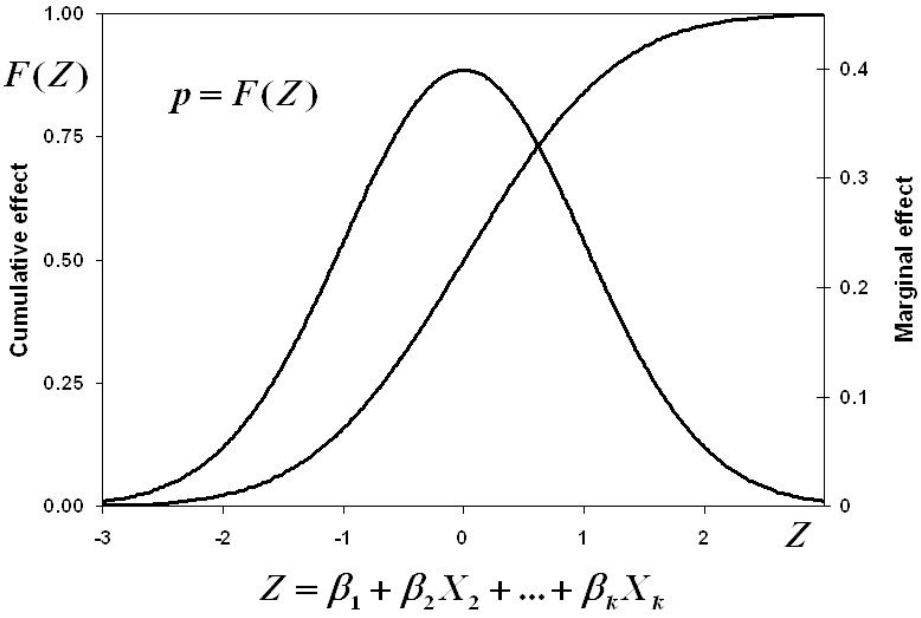
UNIVERSIDAD NACIONAL PEDRO RUIZ GALLO

INSTITUTO DE INVESTIGACIÓN ECONOMÍA Y SOCIEDAD

Guía para la Construcción de Modelos de Regresión Lineal Clásico y Modelos de Elección Binaria con STATA 15.



Lindon Vela Meléndez
Guillermo Eloy Guerrero Carrasco

Lambayeque, Perú, octubre del 2020

TABLA DE CONTENIDO

Guía para la Construcción de Modelos de Regresión Lineal Clásico y Modelos de Elección Binaria con STATA 15

Presentación

La estimación de modelos econométricos se ha vuelto fundamental en la formación profesional de los economistas, debido a que los métodos de la econometría son importantes para la economía aplicada. Su uso va desde la implementación o análisis de políticas públicas hasta la toma de decisiones para las empresas. Esta importancia radica principalmente en la capacidad de relacionar una variable con otra o con un conjunto de variables permitiendo establecer como una variable influye sobre otra. A este estudio de la dependencia de una variable dependiente respecto a variables independientes se denomina análisis de regresión.

Aprender el correcto manejo de datos representa una parte importante para la estimación de cualquier modelo econométrico y en la actualidad, los economistas se apoyan de la informática para crear y/o usar *softwar*es estadísticos siendo los más importantes: Eviews, SPSS, STATA, Gretl, R, Excel, etc. La estimación de modelos econométricos conlleva a usar un software estadístico que permita realizar procesos estadísticos, econométricos y matemáticos que serían imposibles realizar de manera manual y además de una base de datos obtenida mediante un instrumento de recolección de datos, en el caso peruano se usa la Encuesta Nacional de Hogares (ENAHO) dirigida por el Instituto Nacional de Estadística e Informática (INEI), la cual muestra variables económicas y sociales de la población como el ingreso familiar, el gasto familiar, el índice de pobreza, la tasa de informalidad, etc. Además, que a partir de las variables recogidas por la ENAHO es que se pueden calcular otras variables, como el gasto catastrófico, el índice de Gini, etc. es por eso que esta guía servirá para conocer el manejo de datos de la ENAHO con STATA. Con la intención que esta guía de estudios sea lo más orientativa posible se hará uso de la ilustración del trabajo de investigación **medición del impacto de la infraestructura relacionada con acceso a los servicios básicos sobre la pobreza mediante un modelo de regresión logística** publicada por Carlos Aparicio, Miguel Jaramillo y Cristina San Román para mostrar la sintaxis de los comandos que permiten el manejo de datos y el análisis de regresión logística para un modelo con variable dependiente binaria.

Previamente a detallar los comandos en STATA sobre el manejo de base de datos, para realizar una eficiente explicación de cuáles son los pasos para la formulación de modelos econométricos y demostrar que la elaboración de estos modelos están al alcance de cualquier persona en los siguientes capítulos se alcanzara una breve teoría econométrica de cuáles son los aspectos que se deberán tomar en cuenta para la elaboración de los econométricos, **esto ya que aprender econometría no es solo memorizar los comandos de un programa estadístico sino entender la teoría econométrica y su base que es la estadística porque solo de esta manera, el lector no solo será capaz de usar comandos sino de especificar modelos econométricos cada vez más y más complejos.**

De esta manera, lo que se busca lograr con esta guía de estudios es servir como un resumen de lo incomprensible que puede resultar la complicada teoría econométrica y más allá de esto revelar los detalles que se tienen que seguir para la correcta especificación, estimación, evaluación e interpretación de los modelos econométricos para que el lector sea capaz de realizar sus propios modelos econométricos acorde a la investigación que realice requiera un estudio correlacional e ir más allá del estudio descriptivo. Por último, el lector debe recordar que esta no es más que una guía de estudios y que los conocimientos que se pretenden explicar estarán detallados con conceptos simples, por ello es que se le exhorta a complementar lo aprendido con libros especializados de econometría de autores reconocidos.

En el primer capítulo trata sobre una introducción generalizada a conceptos sobre econometría, desde su definición, pasando por su importancia, hasta los pasos que plantea la teoría econométrica para elaborar modelos econométricos. En el segundo capítulo, se detalla algunas especificaciones sobre la población y muestra, las técnicas de muestreo y algunos errores al momento de aplicar muestreo. En el tercer capítulo, se aborda directamente temas en relación al Modelo de Regresión Lineal Clásico y el método de estimación de Mínimos Cuadrados Ordinarios, desde su concepción hasta la detección y tratamiento de violaciones a los supuestos de Gauss-Márkov, y finaliza con una presentación de un ejemplo en STATA sobre los pasos para estimar un modelo econométrico mediante MCO en datos de corte transversal y datos de series temporales. En el cuarto capítulo, se conceptualizan temas referentes al análisis de modelos de elección binaria y se muestra un ejemplo en STATA utilizando un modelo Logit, el cual consiste en una réplica de un trabajo de investigación.

1. Introducción a la Econometría

1.1. ¿Qué es la Econometría y por qué es Importante Aprenderlo?

Desde el origen de la econometría ha existido un debate en cuanto a la definición correcta para la econometría, pero de forma sencilla podemos usar la definición que le otorga Econometric Society, la cual según (Portillo, 2006) esta sociedad plasma su objetivo en el primer artículo y es: promover estudios que se dirijan a una unificación de la aproximación teórico-cuantitativa y empírico-cuantitativa a los problemas económicos y que constituyan reflexiones constructivas y rigurosas similares a las que han llegado a dominar las Ciencias naturales. Según (Portillo, 2006) esta sociedad no define a la econometría como estadística económica ni mucho menos como teoría económica tampoco como matemática aplicada, sino como la unión de estos tres aspectos para la concepción de una herramienta al economista, esta herramienta se llama **econometría**. En síntesis, siguiendo esta definición la econometría es una disciplina compuesta por tres ciencias: la economía, la matemática y la estadística y además está apoyada por la informática. En palabras de (Gujarati & Porter, 2010) La economía, a través de la teoría económica, formula hipótesis sobre las relaciones entre las variables y de su naturaleza cualitativa, la matemática aplicada a la economía es capaz de expresar la teoría económica en forma de ecuaciones y la estadística económica procesa y recopila información en forma de datos estadísticos y cifras económicas que pueden ser visibles en gráficos y cuadros, en consecuencia, **la labor del econometrista es darle contenido empírico a la teoría económica expresadas en el empleo de ecuaciones matemáticas.**

Tal como contempla (Spanos, 1999) El econometrista al elaborar un modelo se enfrenta a datos que provienen de la observación más que la experimentación, por ello la creación de los modelos econométricos requiere dominar habilidades de análisis de datos y familiarizarse con la naturaleza de los datos en cuestión.

Sin embargo, aún queda pendiente responder a la pregunta: ¿Cuál es el fin de la econometría? Para ello se expresa la siguiente cita:

> *"El objetivo básico de la econometría consiste en especificar y estimar un modelo de relación entre las variables económicas relativas a una determinada cuestión conceptual"* (Novales, 1998)

La cita anterior ofrece un alcance sobre la importancia de al econometría, la econometría sirve de herramienta para elaborar un modelo que relacione variables

económicas que describa su comportamiento en un contexto. En otras palabras, es una herramienta que le sirve al economista para que logre determinar las relaciones entre las variables. ¿Cómo se consigue determinar esas relaciones? Se logra determinar una relación entre variables mediante la cuantificación, es decir, la econometría logra cuantificar la influencia de una variable sobre otra. (Hernández A. & Zúñiga R., 2013) Amplían esta idea explicando que el fin más importante y fundamental de la cuantificación de las relaciones entre variables es servir en la previsión de magnitudes económicas, es decir, la econometría sirve para verificar las teóricas económicas.

> *"Se trata de comprobar mediante los resultados del modelo estimado, la validez de la teoría económica que expresa dicho modelo."* (Hernández A. & Zúñiga R., 2013)

Tomando en cuenta ambas citas llegamos a la conclusión que la econometría sirve de herramienta para señalar qué relación existe entre las variables económicas y además señala esa relación cuantificándola cuan influyente es una sobre la otra; con el objeto de determinar la validez de un modelo económico, es decir si verdaderamente se cumple lo que el modelo económico explica. Sin embargo, más adelante caeremos en cuenta que no necesariamente un modelo contradice la realidad debido a que esté equivocado, si esto sucediese también puede deberse a la naturaleza de los datos o a la metodología usada para cuantificar las relaciones o a otro factor, serán temas abordados más adelante.

Pero ¿para qué sirve comprobar la validez de la teoría económica? (Hernández A. & Zúñiga R., 2013) Continúa explicando que la comprobación de la teoría econométrica sirve para la previsión y la evaluación de políticas; la primera se entiende como la predicción que la econometría suministra para conocer el comportamiento futuro de las variables económicas, mientras que la evaluación de políticas se entiende que la econometría permite la valoración de las consecuencias que una acción de un gobierno ejerce sobre la variable explicada.

Para lograr comprobar la validez de la teoría económica se debe comparar los signos de los parámetros estimados con la hipótesis sobre ellos la cual está indicada por la teoría económica, es decir se compara los signos de los estimadores con lo que la teoría económica señala que deberían ser, cuando ambas coinciden entonces el modelo econométrico confirma empíricamente la teoría económica, de no ser así entonces el modelo posibilita la revisión o sustitución de la teoría.

1.2. La Modelización Econométrica

La definición de un modelo económico es una simplificación de la realidad que muestra hipótesis sobre conductas de las variables económicas y sus relaciones. Por ello para (Ouliaris, 2011) los modelos económicos pueden ser teóricos o empíricos, los primeros tratan de buscar implicaciones verificables sobre el comportamiento económico siguiendo el supuesto que los agentes maximizan sus objetivos, mientras que los segundos tratan de verificar las predicciones cualitativas de los modelos teóricos y transformarlas en resultados numéricos. Los modelos económicos suelen constar de ecuaciones que buscan explicar la conducta de los agentes racionales o el funcionamiento de una economía, por lo tanto las ecuaciones buscan simplificar una realidad. En palabras más simples, los economistas usan modelos para explicar la realidad basada en las conductas de las variables económicas y para lograr esa explicación, los economistas usan modelos en forma de ecuaciones.

Sin embargo, al ser la realidad totalmente compleja ningún modelo puede explicar perfectamente la realidad. (Greene, 2012) Postuló que un modelo no puede tener en cuenta todas las influencias (relaciones) pero a pesar de la existencia de esa carencia de relaciones entre la variable dependiente y los aspectos no tomados en cuenta, esta carencia no supone ser importante para nuestro modelo. En otras palabras, ningún modelo podría englobar todos los aspectos aleatorios de las variables económicas, por lo tanto, es necesario tomar en cuenta los aspectos estocásticos en nuestros modelos empíricos. Para (Greene, 2012) La introducción de un aspecto estocástico a un modelo empírico hace que la explicación de la conducta de la variable dependiente, es decir las variaciones de la variable dependiente, no solo sean atribuidas al comportamiento de las variables independientes identificadas en el modelo empírico sino también a la aleatoriedad del comportamiento humano. Entonces al tomar en cuenta el aspecto estocástico, se convierte una afirmación exacta en una descripción probabilística y esta condición de ser probabilístico hace que el modelo sea menos preciso. Es por ello que el uso de modelos como herramientas para explicar un determinado fenómeno económico hace a la economía una ciencia probabilística.

Podemos llegar a la conclusión, que la teoría económica explica la realidad mediante la simplificación que ofrece construir modelos, un modelo económico explica el comportamiento y las relaciones de las variables económicas y puede ser descrito en forma de ecuaciones, pero al no ser capaz de recoger todos los aspectos de una realidad

debido a la conducta humana, es que se debe agregar un elemento estocástico. **Cuando se le agrega ese elemento estocástico entonces deja de ser un modelo económico y pasa a ser un modelo econométrico**, el cual permite cuantificar y contrastar las relaciones entre las variables económicas que señala la teoría económica a través del modelo económico. A continuación, se denotan dos formas funcionales de un modelo económico y econométrico, respectivamente.

$$Y = f(X_1, \dots, X_k) \quad (1.2.1.)$$

$$Y = f(X_1, \dots, X_k) + \varepsilon \quad (1.2.2.)$$

En la segunda forma funcional, se observa el símbolo **ε**, este símbolo representa el aspecto estocástico. No olvidar que el aspecto estocástico hace referencia a lo que el modelo no contempla expresándolo en una variable económica en concreto, pero que existe y representa factores no precisados, es decir factores no observados. Según (Wooldrige, 2009) en el análisis econométrico, como tratar este aspecto estocástico, o mejor dicho en palabras más técnicas, como tratar el **término de error o perturbación** es quizá el componente más importante. Pero ese análisis será detallado posteriormente.

1.3. El Efecto Causal y la Noción de Ceteris Paribus

Observemos el siguiente ejemplo de lo que es el efecto causal, realizado por (Alonso, 2012):

> *"Efecto causal de la educación en el salario.*
>
> *Es el incremento salarial que conseguiría un individuo de la población objeto de estudio si, manteniéndose constantes sus demás características, tuviera un nivel mayor de educación (por ejemplo, un año adicional, tener o no un título universitario, etc.)"* (Alonso, 2012)

A través del ejemplo, podemos identificar la variable dependiente e independiente y el efecto que ejerce la variable independiente sobre la variable dependiente, en este ejemplo la variable dependiente es el nivel de salario, la variable independiente es el grado de estudio y el efecto que hace el grado de estudio sobre el nivel de salario, pues al incrementar la primera variable hace incrementar también la segunda. Sin embargo, el ejemplo menciona un enunciado importante, pues se logra ese efecto causal si se mantienen las demás características constantes, que podrían ser número de hijos, número

de años laborando en la empresa, etc. Es decir, manteniendo constante las demás variables que influyen sobre el nivel de salario.

Para (Kendall & Stuart , 1961) por más que exista una relación estadística fuerte nunca será suficiente para suponer que existe causalidad, esta debe venir de la teoría o de estadísticas externas. Pero, para (Gujarati & Porter, 2010) la causalidad también puede provenir del sentido común y ejemplifica que el rendimiento de un cultivo también depende de la temporada de lluvias, no se necesita de ninguna teoría ni de cuestiones estadísticas sino de sentido común, y concluye afirmando que **una relación estadística no implica la existencia de causalidad y para encontrarla se debe revisar las consideraciones a priori o teóricas**. Sin embargo, la existencia de la condición de mantener las demás variables constantes para medir la influencia de una variable sobre otra, supone la importancia de la condición ceteris paribus, esta condición es importante porque mediante él se pretende aislar el efecto del aspecto estocástico para estimar el efecto de la variable explicativa sobre la variable a explicar. (Wooldrige, 2009) Ejemplifica esta condición, cuando se analiza la demanda del consumidor se quiere determinar cómo el precio explica la cantidad demandada de un bien, por lo tanto en condiciones de ceteris paribus logramos aislar los efectos que también ejercen otras variables como gustos y preferencias, precios de bienes sustitutos complementarios y/o sustitutos, etc. Sin embargo, (Wooldrige, 2009) continuando explicando que este supuesto a pesar que es fundamental para los estudios econométricos, hace resaltar la pregunta: ¿se han mantenido constantes suficientes factores para que se justifique la causalidad? Debido a que explicar con exactitud un fenómeno económico resulta ser complejo e imposible, es que el supuesto de ceteris paribus en la econometría es difícil de seguir, por ello es que haciendo uso de las técnicas econométricas correctas se puede simular esta condición.

1.4. Enfoque de la Econometría Tradicional

Antes de presentar de forma detallada la metodología que sigue la econometría, es necesario presentar el enfoque de la econometría tradicional o clásica. La mayoría de autores concuerdan en que la econometría necesita de tres componentes, y estos son: la teoría económica, los datos cuantitativos o cualitativos y las técnicas estadísticas o econométricas. (Núñez Z., 2007) Explica estos tres componentes y tal como se ha señalado anteriormente, la teoría económica es el primer elemento fundamental para la elaboración de cualquier modelo econométrico, y la simplificación de la teoría económica

se logra a través de un modelo matemático usando ecuaciones. Los datos se obtienen a partir de las observaciones y son mediciones de hechos, esto quiere decir que son provocados por fenómenos que pueden estar expresados en términos de espacio y tiempo. Esta última aseveración es mejor explicado por (Gujarati & Porter, 2010) quienes detallan con precisión los conceptos de **datos de series de tiempo**, **datos de corte transversal** y **datos panel**. Explica que una serie de tiempo es un conjunto de observaciones sobre los valores de una variable en diferentes momentos expresados en forma diaria, semanal, mensual, trimestral, anual, quinquenal, decenal, etc. Por otro lado, los datos transversales o de corte transversal son datos que consisten registrados en el mismo momento del tiempo y finalmente los datos de panel son los datos que estudia a través del tiempo la misma unidad transversal, es decir una combinación de los dos tipos de datos anteriores. Para efecto de este trabajo se hará un ejemplo de regresión con el uso de datos de corte transversal, sin embargo, el análisis de series de tiempo es fundamental para el análisis de la conducta de las variables económicas en un periodo y su predicción. Existen múltiples fuentes de datos, y en el caso peruano, la fuente que pretende representar la población de manera más significativa es sin lugar a dudas la Encuesta Nacional de Hogares (ENAHO), el manejo de STATA para esta encuesta se describirá de manera más detallada en los siguientes apartados. Por último, las técnicas estadísticas o econométricas son el tercer componente del enfoque tradicional y sirven tanto para analizar la base de datos como para estimar los parámetros del modelo especificado según la teoría económica. Sirve además para realizar pruebas para diagnosticar incumplimientos de supuestos, el cual será detallados en los siguientes capítulos, de igual forma se hablará de cada aspecto de forma más detallada en los siguientes capítulos. A continuación, se muestra una figura recogida de (Núñez Z., 2007) que explica el enfoque tradicional de la econometría.

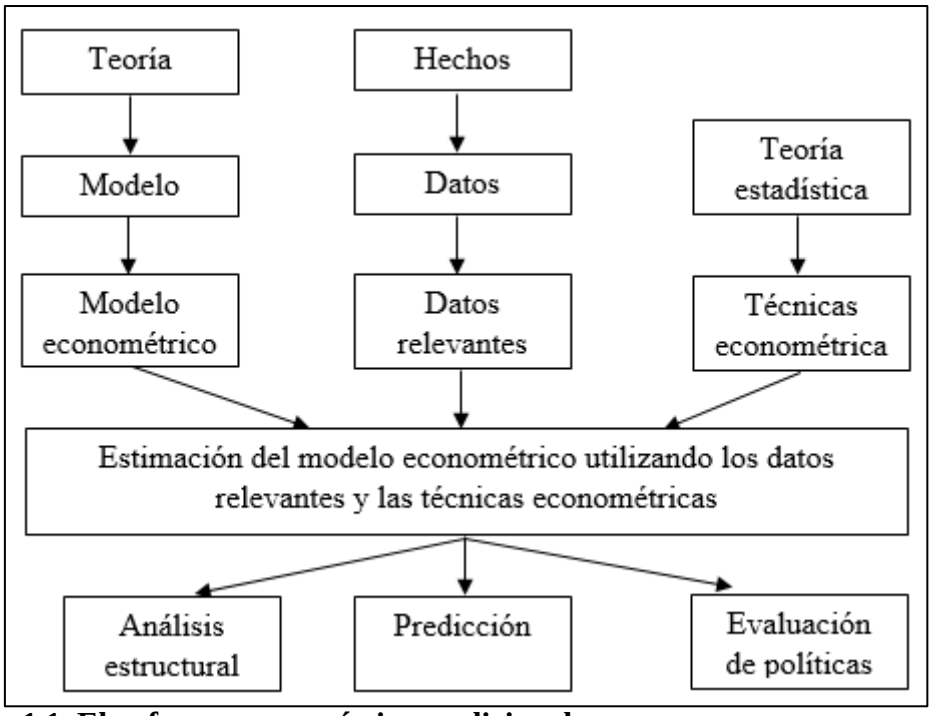


Figura 1.1. El enfoque econométrico tradicional
Elaboracion: Núñez, 2007
Fuente: Intriligator, Bodkin y Hsiao, 1996

Un breve comentario sobre el enfoque tradicional y contemporáneo, el enfoque tradicional combina estos tres componentes anteriormente señalados y le da importancia a la combinación de estos tres componentes para realizar una adecuada especificación y posterior estimación de un modelo econométrico, en otras palabras no basta solo con tener un buen manejo en las técnicas econométricas sino también es importante contar con una teoría económica que explica la realidad y una base de datos que sea representativa de la población, sin embargo, existe un sesgo en este enfoque y es que se busca cada vez más que la teoría y la construcción de base de datos sean lo más representativo posible, condición que difícilmente podrá ser cumplida. Detalla (Núñez Z., 2007) Los problemas que surgen a través del enfoque tradicional que pueden ser problemas metodológicos o relacionados con los modelos econométricos, por ello es que la econometría contemporánea pretende resolver estos problemas, sin embargo, su estudio y análisis merecería otro trabajo dedicado exclusivamente al enfoque contemporáneo.

1.5. Metodología de la Econometría tradicional

A partir del enfoque tradicional, los economistas optan por seguir la metodología tradicional de la econometría para hacer investigaciones, políticas públicas y predicciones. (Gujarati & Porter, 2010) Explican que la metodología se ajusta a los siguientes pasos:

1. Planteamiento de la teoría o de la hipótesis.
2. Especificación del modelo matemático de la teoría.
3. Especificación del modelo econométrico o estadístico de la teoría.
4. Obtención de datos.
5. Estimación de los parámetros del modelo econométrico.
6. Pruebas de hipótesis.
7. Pronóstico o predicción.
8. Utilización del modelo para fines de control o de políticas.

Sin embargo, en algunos trabajos que sirven de guía para el estudio de la econometría tradicional hace mención de cuatro pasos para el uso de modelos econométricos (Aguarto P., 2010) señala esos cuatro pasos:

1. Especificación del modelo.
2. Estimación del modelo.
3. Evaluación de los estimadores.
4. Evaluación de la capacidad predictiva del modelo o interpretación.

Se hablará de estos cuatro para el uso de modelos econométricos pasos con mayor detalle, y posteriormente se hablará de forma más completa de la metodología tradicional en los siguientes capítulos.

1.5.1. Especificación del modelo.

Este es el paso más importante de todos, de este paso depende que el modelo tenga la forma adecuada y que el uso que se le dará al modelo, ya sea para explicar la conducta de las variables en un fenómeno económico determinado o para realizar una política pública o la predicción de una variable, será significativo.

Para (Aguarto P., 2010) especificar un modelo conlleva a determinar la variable dependiente y cuáles serán las variables independientes, el tamaño del modelo y el signo esperado de los parámetros, la forma matemática, es decir si seguirá una forma lineal o no lineal y si el modelo será uniecuacional o multiecuacional.

Ya previamente se había indicado que un buen modelo econométrico se basa en la teoría económica, sin embargo, se suele recurrir también a la evidencia empírica como complemento en la especificación de un modelo econométrico, es importante entonces considerar las variables significativas sin caer en sesgos de omisión de variables

relevantes o la inclusión de variables irrelevantes. (Acosta G., Andrada F. Julián, & Fernández M., 2009) Explican en qué consisten estos sesgos. El principal problema de un modelo de regresión múltiple es la **selección de los regresores** o variables explicativas para el modelo que se trata de especificar, este problema se origina debido a que al ser tantas variables que pueden influir de una manera u otra a la variable dependiente es difícil y en algunos casos imposible tener todas las variables independientes, por ello es que se asume el concepto de la perturbación aleatoria o mejor dicho el aspecto aleatorio de un modelo econométrico. Sin embargo, la sobreparametrización de los modelos econométricos genera una buena predicción intramuestral pero una mala predicción extramuestral. Esto quiere decir que explica muy bien la muestra pero no tiene la capacidad de poder ser generalizado a toda la población. Advierten que a pesar que lo recomendable es que el modelo econométrico se apoye en una teoría economía consolidada, en algunos casos la teoría económica no podrá ayudar a la hora de decidir cuáles serán las variables independientes que serán tomadas en cuenta para elaborar un modelo econométrico capaz de simplificar la realidad.

> *"Los modelos económicos suelen ser menos precisos que los econométricos, de esta manera se corre el riesgo de especificar modelos con variables explicativas irrelevantes, o por el contrario con la omisión de variables explicativas relevantes. Estas circunstancias tendrán determinadas repercusiones en el modelo."* (Acosta G., Andrada F. Julián, & Fernández M., 2009)

Sin embargo, existen algunas soluciones que ayudan al economista a considerar cuales son las variables a tomar en cuenta, por ejemplo, los criterios de información pueden ser una herramienta útil en este tipo de situaciones, no obstante, los criterios de información no "curan" al modelo de este problema, en consecuencia, siempre se debe tener en cuenta la existencia de una duda estadística y ante esta duda, se debe proceder con cuidado y precaución.

Casi de igual forma al problema de los regresores, anteriormente explicado, existe otro problema descrito por (Aguarto P., 2010) quien plantea que es labor del economista determinar la forma matemática del modelo. Esto implica detallar el número de ecuaciones que se usarán y la forma de las ecuaciones. Sin embargo, el enfoque tradicional hace uso del concepto **linealidad** el cual será abordado de forma más detallada más adelante, pero de momento podemos afirmar que según (Wooldrige, 2009) La linealidad permite que todo cambio en una variable independiente en una unidad tiene

siempre el mismo efecto sobre la variable dependiente. El aspecto tradicional hace tomar en cuenta la linealidad, por ello se puede expresar la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i + u_i (1.5.1.)$$

Donde:

Y_i: variable dependiente

X_i: variable independiente

β_1 y β_2: parámetros a estimar

u_i: término de perturbación

La anterior ecuación puede determinarse como modelo econométrico, debido a la consideración del término de perturbación en la especificación del modelo econométrico, el cual distingue de un modelo económico. El subíndice *i* indica que esta ecuación es con datos de corte transversal, si el subíndice sería denotado con *t* entonces sería una ecuación con datos de series temporales y si el subíndice fuese *it* entonces estamos ante un modelo con datos panel. Los parámetros β_1 y β_2 son los parámetros que se estimarán, se explicará la estimación de un modelo en el siguiente paso pero desde ahora ya se asume que se les conoce como **término de intercepto** y **pendiente** respectivamente y ambos son conocidos como **coeficientes de regresión** y el nombre más apropiado para la ecuación anterior es **modelo de regresión lineal uniecuacional simple**, se denomina simple porque solo hace uso de dos variables, una dependiente y una independiente, si hiciera uso de más variables independientes entonces se llamaría **modelo de regresión lineal uniecuacional múltiple**, el cual es la forma de especificación más usada por los economistas. La ecuación 1.5.2. Demuestra ser un modelo de regresión múltiple donde k es el número de variables explicativas que tiene el modelo.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i (1.5.2)$$

Ambas ecuaciones, tanto (1.5.1.) como (1.5.2.) corresponden a función de regresión población, estaremos ante una función de regresión muestral cuando los estimadores tengan encima un *"gorrito"*. La ecuación (1.5.3.) muestra una función de regresión muestral.

$$Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + \widehat{\beta_3} X_{3i} + \cdots + \widehat{\beta_k} X_{ki} + u_i (1.5.3.)$$

Más adelante se abordará con detalle la diferencia de estas dos funciones de regresión. Sin embargo, desde este momento se puede inferir que los coeficientes de

regresión, son **estimadores** también conocidos como **estadísticos** y que la técnica de estimar un parámetro población (sin gorrito) a partir de información muestral corresponde a la inferencia estadística. Finalmente, se presenta un cuadro donde se muestran los sinónimos de las variables independientes y dependientes.

Y	X
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de estímulo
Variable predicha	Variable predictora
Variable Regresada	Variable regresora
Variable Endógena	Variable exógena

Tabla 1.1. Terminología de las variables independiente y la variable dependiente
Elaboración propia
Fuente: Gujarati & Porter, 2010

1.5.2. Estimación del modelo.

En este paso se tratará de la cuantificación de los parámetros del modelo, usando un conjunto de datos que servirán como la muestra del modelo y se logrará a través del método econométrico más indicado.

(Aguarto P., 2010) Explica que realizar para realizar la estimación del modelo se requiere realizar las siguientes tareas:

- La recolección de observaciones estadísticas para cada una de las variables del modelo.
- El examen de ciertos problemas de agregación que están implícitas en algunas variables de naturaleza macroeconómica.
- El examen del grado de asociación que podría existir entre algunas variables que hacen el papel de explicativas en el modelo.
- El examen de las condiciones de identificación de la relación que pretende estimarse.
- La elección del método econométrico más apropiado para la estimación.

1.5.2.1. Recolección de datos.

Ya anteriormente se ha descrito los conceptos de **datos de series de tiempo**, **datos de corte transversal** y **datos de panel**, es importante tomar en cuenta el tipo de datos con el que se trabajara debido a que cada tipo de dato requiere un procedimiento de estimación específico. Se suele encontrar en la literatura econométrica la existencia de problemas que se les relaciona a los tipos de datos con los que se usan para estimar el modelo. (Hernández A. & Zúñiga R., 2013) Explica cuáles son los problemas en los datos de las variables. Entre ellos se detalla la **muestra insuficiente** que hace referencia a que se debe buscar siempre un gran número de observaciones, lo más amplio posible, debido a que la estimación será más precisa cuando se tenga más información consiguiendo que el número de observaciones supera una cuota si se especifica una cuota, por ejemplo, es bien sabido que en los modelos que siguen la metodología Box-Jenkins el cual es una metodología que modela series temporales para su posterior predicción, recomienda y exhorta al economista a usar más de 50 observaciones para que el modelo sea lo más preciso posible. En otras palabras, se debe buscar siempre tener una amplia base de datos para cada variable para tener certeza en que el modelo será significativo y así se evitará que se generen otros problemas en la estimación como fallos a los supuestos.

Otro problema son los **errores en los datos,** esto sucede cuando existen errores en la medición numérica de la información, si se detecta este problema entonces no se puede confiar en los resultados obtenidos de la estimación. Existen dos tipos de error de medición: el error puntual y el error de sesgo; el primero se genera por un valor atípico, el cual es una observación distinta al resto de los datos, por ejemplo, un conjunto de datos sobre las exportaciones de un conjunto de países en un determinado periodo, la mayoría de estos países tienen exportaciones entre 4 millones de dólares y 7 millones de dólares sin embargo, existen países que tienen 30 millones de dólares en el mismo conjunto, estos países que se encuentran alejados del centro se les conoce como **valores atípicos o outliers**. La solución es detectarlos con un gráfico de residuos y aplicar las técnicas apropiados como las técnicas robustas. Por otro lado, el error de sesgo es más difícil de detectar y pueden generar problemas de multicolinealidad o cambio estructural. Sin

embargo, esos problemas requieren un estudio aparte, pero desde ahora podemos asumir que su origen está en un error de medición de los datos o en una muestra insuficiente.

1.5.2.2. Problemas de agregación.

(Aguarto P., 2010) Explica que la estimación de los modelos macroeconómicos obliga a usar variables que estudien el comportamiento agregado de unidades individuales y esta agregación es siempre generadora de sesgos y errores. Sin embargo, este trabajo tomará énfasis en el trabajo microeconométrico y no macroeconométrico. Por lo que se recomienda la lectura de textos especializados en el tema.

1.5.2.3. Multicolinealidad.

Este problema será abordado con mayor detalle en los posteriores apartados, debido a que concierne más a ser tratado como un fallo en los supuestos del modelo clásico de regresión lineal (MCRL), sin embargo, a menudo se le define como la existencia de alto grado de correlación entre las variables explicativas del modelo.

1.5.2.4. Examen de las condiciones de identificación de la relación.

(Aguarto P., 2010) Detalla que el economista debe conocer cuáles son las relaciones entre las variables seleccionadas.

1.5.2.5. Elección del método econométrico más apropiado para la estimación.

Este es el paso que nos permite cuantificar los coeficientes de regresión, es decir, nos permite medir la relación económica entre la variable explicada y las variables explicativas. Siendo la estimación por el método de los mínimos cuadrados ordinarios (MCO) el método más usado para estimar modelos de regresión lineal. Sin embargo, existen otros métodos para estimar modelos econométricos. Otros conocidos son:

- MCG: Método de cuadrados generalizados
- MV: Método de máxima verosimilitud
- MC2E: Método de mínimos cuadrados en dos etapas
- MCR: Mínimos cuadrados robustos

Como todos los pasos anteriores, escoger el método de estimación puede resultar ser difícil para el economista, sin embargo, (Aguarto P., 2010) sugiere que debes tomar en cuenta los siguientes factores:

- La naturaleza de la relación y sus condiciones de identificación.
- El propósito de la identificación. Debemos siempre alinear nuestro método de estimación a lo que se busca en el marco de la investigación, por ejemplo, para estimar un modelo con variable explicada dicotómica, es decir una variable cualitativa que señala la cualidad de una observación, es sugerido que se siga el método de estimación por máxima verosimilitud que el método de mínimos cuadrados ordinarios.
- La simplicidad del método y los requerimientos de tiempo y costo. Se suele recomendar seguir la estimación más simple y que no requiera demasiado esfuerzo, siguiendo el principio de la parsimonia.

Sin embargo, estas no resultan ser más que guías, depende libremente del economista elegir el correcto método para estimar actuando siempre con criterio, buscando los estimadores o coeficientes de regresión que sean MELI (Mejores Estimadores Lineales Insesgados) para ello deben cumplir ciertas propiedades, pero se abordarán en los apartados posteriores.

1.5.3. Evaluación de los estimadores.

Una vez estimado los coeficientes de la regresión, se debe probar que estos valores numéricos tengan utilidad para su posterior interpretación. En palabras de (Aguarto P., 2010) Es determinar cuán significativos y correctos son los estimadores que hemos conseguido en la etapa de estimación. Para ello, se consideran los siguientes criterios:

1.5.3.1. Criterio económico.

Se espera a que los coeficientes de regresión cumplan con el signo y el tamaño según los lineamientos de la teoría económica, por ello el criterio económico contrasta que los coeficientes de regresión, que según el modelo pueden ser propensiones, valores marginales, multiplicadores, etc., cumplan con lo que ya se especificó siguiendo la teoría económica y la evidencia empírica.

Ante el contraste solo quedan dos opciones: que los coeficientes de regresión cumplan los requerimientos especificados anteriormente o que no la cumplan. Es deseable

que ocurra la primera opción, pero cuando no es así es deber del economista demostrar el motivo de que porque esto no sucede. Podríamos pensar en simplemente rechazar la teoría económica, más aún cuando tenemos la especificación del modelo, la técnica y los datos adecuados, pero se debe probar con una investigación consistente y meticulosa. Por lo tanto, frente a la segunda opción es mejor replantear el modelo, lo que implica tomar en cuenta nuevas variables, otra forma funcional o elegir otra metodología de estimación. La afirmación anterior debe quedar muy claro en el lector.

Si continúa sin cumplir con los requerimientos que la teoría económica ha establecido, entonces se tendría los elementos necesarios para concluir que la teoría económica debe ser rechazada porque no puede ser demostrada econométricamente.

1.5.3.2. Criterio estadístico.

Anteriormente se detalló, que la estadística tendría un papel importante en la formulación de la teoría econométrica, y ahora se explicara porque la econometría se apoya en la teoría estadística. El criterio estadístico consiste en someter a los coeficientes de regresión a pruebas para medir su certeza, estas pruebas también llamadas como test o exámenes se apoyan en la **prueba de hipótesis**, un concepto recogido por la **estadística inferencial**.

Por lo tanto, la estadística inferencia, tendrá un rol sumamente fundamental para lograr probar mediante prueba de hipótesis la certeza o grado de confiabilidad de los coeficientes de regresión. Para comprender la econometría, es necesario comprender los procedimientos de la estadística inferencial, por ello se expone un breve concepto de la estadística inferencial.

> *"La estadística ha desarrollado técnicas y procedimientos para generalizar datos relacionados con los parámetros de una población, con base en la información contenida en una muestra representativa de dicha población."* (Pérez-Tejada, 2007)

El enunciado anterior muestra la definición de lo que es la estadística inferencial, concluimos que cuando intentamos inferir los aspectos de una población a partir de una muestra estamos haciendo estadística inferencial, sin embargo, siempre se ejecuta tomando cierto grado de fiabilidad o mejor dicho **nivel de confianza**. Para la literatura estadística, existen dos formas de estadística inferencial: **la estimación** y **la prueba de**

hipótesis y dentro del método de la estimación se encuentra **la estimación puntual** y **la estimación intervalar.** En la etapa anterior se señaló la existencia de propiedades que deben tener los estimadores, en esta etapa se menciona cuáles son esas propiedades. (Pérez-Tejada, 2007) Señala que los estimadores conseguidos por estimación puntual tienen las siguientes propiedades:

- **Insesgado:** Se dice que un estimador o coeficiente de regresión es insesgado cuando el **valor esperado** del estimador muestral coincide con el verdadero valor del parámetro poblacional. Matemáticamente se expresa de la siguiente manera:
- $E(\hat{\beta}) = \beta$ (1.5.4.)
- Al asumir que el valor esperado, conocido también como **esperanza**, **media** o **promedio**, del estimador muestral es igual verdadero valor del parámetro poblacional entonces el estimador muestral es insesgado.
- **Eficiente:** La eficiencia de un estimador muestral compara las varianzas de dos estimadores muestrales y elige al que tenga varianza mínima. Matemáticamente se expresa de la siguiente manera:
- $V(\widehat{\beta_1}) < V(\widehat{\beta_2})$ (1.5.5.)
  - En algunos textos puede encontrarse la siguiente forma matemática:
- $\sigma^2_{\hat{\beta}_1} < \sigma^2_{\hat{\beta}_2}$ (1.5.6.)
- Tanto en (1.5.5.) cómo (1.5.6.) se puede interpretar que el estimador $\hat{\beta}_1$ es más eficiente que $\hat{\beta}_2$. En cuyo caso se prefiere el estimador $\hat{\beta}_1$ debido a que tiene una varianza mínima con respecto a $\hat{\beta}_2$. Para que un estimador sea eficiente debe cumplir la propiedad de insesgamiento.
- **Consistente:** Un estimador muestral es consistente cuando al ir aumentando el tamaño de la muestra, el estimador muestral se acerca al verdadero valor del parámetro poblacional.
- (Ponce A. & Nolberto S., 2008) Explican que esta propiedad se cumple debido a que al aumentar el tamaño de la muestra podemos estar más seguros que el error entre el estimador muestral y el parámetro población será menor y lo expresan matemáticamente:
- $\lim_{n\to\infty} P(|\hat{\beta} - \beta|) < c = 1$ (1.5.7.)

- Interpretan de la siguiente manera: la ecuación (1.5.7.) el estimador muestral ($\hat{\beta}$) es consistente del parámetro poblacional (β) si y solo si para cada c>0. En palabras sencillas, cuanto menor es la diferencia entre el estimador muestral y el parámetro poblacional con probabilidad igual a 1, el estimador muestral se aproxima lo más posible al parámetro poblacional.
- **Suficiente:** Un estimador muestral es suficiente cuando se utiliza toda la información muestral para su estimación.

Es importante conocer las propiedades de los estimadores muestrales debido a que en muchas ocasiones durante el proceso de validación se puede detectar violaciones de los supuestos de regresión lineal, como la **heterocedasticidad** o la **autocorrelación** que provocan que los estimadores pierdan las propiedades que los hacen estimadores. Si un estimador muestral pierde sus propiedades entonces no se puede concluir que está representando al parámetro poblacional, en otras palabras, la estimación de los coeficientes de regresión mediante un método de estimación debe asegurar el cumplimiento de estas propiedades. Un último detalle es que la evaluación del cumplimiento de los supuestos lineales corresponde al criterio econométrico, pero tienen una base muy sólida al criterio estadístico.

La segunda forma de estimación es a través de la estimación intervalar, también llamado estimación por **intervalos de confianza o región de confianza**. Determinado por dos valores, uno superior y otro inferior, donde se puede afirmar que con un determinado **nivel de confianza**, el verdadero valor del estimador se encuentra entre esos dos valores.

> *"La construcción de ese intervalo a partir de la información observada y recopilada de una muestra provee una banda alrededor del parámetro estimado, asegurando con una probabilidad determinada que dicho parámetro está ubicado dentro del intervalo."* (Pérez-Tejada, 2007)

Matemáticamente, se expresa de la siguiente manera:

$$P[LI \leq \beta \leq LS] = 1 - \alpha \quad (1.5.8.)$$

El símbolo α, se define como el **nivel de significancia o nivel de significación**, y representa la probabilidad de fallar la estimación. Mientras la expresión 1-α, se define como el **nivel de confianza** y es la probabilidad que el verdadero valor del estimador se

encuentre en el intervalo de confianza construido. El valor más usado en la mayoría de investigaciones económicas y programas estadísticos es un nivel de confianza de 95%, por lo tanto el valor del nivel de significancia es 0.05. Entonces se interpreta de la siguiente manera: el 95% de las veces de los intervalos de confianza construidos incluirán dicho parámetro y el 5% no lo incluirán. La construcción de los intervalos se mostrará más adelante con un amplio detalle, sin embargo, para que la aseveración anterior pueda ser comprendida con facilidad, se presenta el siguiente ejemplo, expuesto por (Ponce A. & Nolberto S., 2008)

$$P[106.08 \leq \beta \leq 113.92] = 1 - 0.05 \quad (1.5.9.)$$

$$P[106.08 \leq \beta \leq 113.92] = 0.95 \quad (1.5.10.)$$

Obviando el proceso de construcción de la expresión (1.5.10.) se interpreta de la siguiente manera: Con un 95% de nivel de confianza podemos concluir que el verdadero valor del estimador β se encuentre entre 106.08 y 113.92.

Finalmente, la última forma de la estadística inferencial, **test de hipótesis**, al principio de este apartado se mencionó que la prueba de hipótesis tendría un importante papel en el criterio estadístico para realizar la evaluación de los coeficientes de regresión estimados mediante algún método de estimación. (Aguarto P., 2010) Reconoce dos test de pruebas de **significancia** y **test de bondad de ajuste**. Sin embargo, las pruebas de hipótesis tienen mayores aplicaciones, son usadas también para evaluar si los supuestos de la regresión lineal se cumplen, como el supuesto de homocedasticidad o no autocorrelación. En los siguientes apartados, se entrará en gran detalle de cómo los economistas usan pruebas de hipótesis para medir la fiabilidad de un modelo. Finalmente, (Aguarto P., 2010) Realiza un comentario sobre el criterio estadístico y económico; señala que las pruebas deben hacerse con criterio básico de económico, esto quiere decir que se puede obtener estimadores significativos y un modelo igualmente significativo, pero no sirve de nada si contradice con los criterios económicos, es decir si el signo esperado no se consigue en la estimación. **En otras palabras, el criterio económico tiene un mayor peso que el criterio estadístico**.

1.5.3.3. Criterio econométrico.

En la etapa de estimación, se usaba una técnica econométrica para estimar los estimadores muestrales, ahora se harán uso de técnicas econométricas para validar la técnica de estimación usada en el anterior paso.

Tal como señala (Aguarto P., 2010) el criterio econométrico consiste en determinar que los supuestos se hayan cumplido en el proceso de estimar los coeficientes de regresión y en el caso que no cumplan estos supuestos el economista deberá proceder a ejecutar una técnica econométrica para corregir esos fallos a los supuestos.

(Pérez L., 2012) Clasifica a estos supuestos o hipótesis del modelo de regresión lineal en cuatro grupos según los componentes del modelo, son:

- Supuestos sobre la perturbación aleatoria
- Supuestos sobre los regresores
- Supuestos sobre los parámetros
- Supuestos sobre la forma funcional

Antes de pasar a explicar cada grupo, es necesario mencionar a modo de recordatorio que todo modelo econométrico tiene un aspecto aleatorio y ese aspecto aleatorio está representado por las perturbaciones aleatorias, que son todos los factores que no se especifican en la ecuación econométrica pero de igual forma explican el comportamiento de la variable dependiente en el modelo econométrico, el **término de error o llamado también termino de perturbación** representa de manera simbólica todas las perturbaciones aleatorias. Se suele usar al símbolo μ como término de error. Debido a que el término de error proviene de las perturbaciones aleatorias, llegamos a la conclusión que **el término de error es una variable aleatoria**. Antes de continuar con la explicación de los supuestos de la regresión lineal, es necesario explicar que es una variable aleatoria debido a que la terminología empleada en la teoría estadística y teoría econométrica pueden resultar definiciones confusas para el lector.

(Véliz C., 2011) Definen como **variable aleatoria** como una función que asigna valores reales a cada resultado de un experimento aleatorio. Se denotan con letras mayúsculas: *X, Y, Z*, etc. y sus valores con letras minúsculas: *x, y, z*, etc. Una variable aleatoria puede ser discreta o continua, se dice que una variable aleatoria es discreta cuando el conjunto de sus valores se puede contar, por lo general describe el número de veces de ocurrencia de un evento. Por otro lado, una variable aleatoria es continua cuando sus valores pueden encontrarse en un determinado intervalo.

Por ejemplo: se tiene una comunidad donde las personas leen periódicos entonces la variable aleatoria discreta *X* muestra el número de veces que una persona lee un periódico durante el día. Por otra parte, una empresa registra sus ventas durante un mes, *Y* es una variable aleatoria continua que muestra sus valores que pueden encontrarse en el intervalo]0,+∞ [.

Toda variable aleatoria sigue una **función de distribución de probabilidad**, la cual describe matemáticamente la probabilidad que sigue la variable aleatoria al momento de designar valores aleatorios. A continuación, se expone un ejemplo planteado por (Freund & Walpole, 1990).

Se tienen dos dados, un dado rojo y otro verde, donde se realizará el siguiente experimento: se tirarán los dos dados y el resultado de sus caras se sumarán, habiendo asignado la probabilidad 1/36 a cada elemento del espacio de la muestra, sin embargo los valores de cada resultado del experimento aleatorio tiene su propia probabilidad. La tabla (1.2.) muestra los valores de la variable aleatoria y sus probabilidades.

x (valor de la variable aleatoria)	P(*X*=*x*)	*x* (valor de la variable aleatoria)	P(*X*=*x*)
2	1/36	10	3/36
3	2/36	11	2/26
4	3/36	12	1/36
5	4/36		
6	5/36		
7	6/36		
8	5/36		
9	4/36		

Tabla 1.2. Valores de la variable aleatoria
Elaboración propia
Fuente: (Freund & Walpole, 1990)

Tomemos el caso de *x* cuando vale 9, es decir *x*=9, este valor tiene una probabilidad de 4/36, debido a que sigue la siguiente función:

$$f(x) = \frac{6-|x-7|}{36} \text{ (1.5.11.)}$$

Si reemplazamos los valores de x en la función (1.5.11.) obtendremos las probabilidades de la variable aleatoria:

$$f(2) = \frac{6 - |2 - 7|}{36} = \frac{1}{36}$$

$$f(3) = \frac{6 - |3 - 7|}{36} = \frac{2}{36}$$

… … … … … … … … … … … … …

$$f(12) = \frac{6 - |12 - 7|}{36} = \frac{1}{36}$$

Finalmente, del ejemplo anterior podemos concluir que **toda variable aleatoria sigue una función de distribución de probabilidades, que designa la probabilidad con la que sus valores aparecen en el espacio muestral**. El ejemplo anterior corresponde a una **función de distribución de probabilidades para una variable discreta.** Y (Freund & Walpole, 1990) Señala que la función de distribución cumple su propósito en una variable aleatoria discreta *X* si y sólo si:

- $f(x) \geq 0$ para cada valor en su dominio.
- $\sum_x f(x) = 1$ donde la sumatoria se extiende sobre todos los valores contenidos en su dominio.

Ambos teoremas señalan que la función arroja probabilidades positivas y que además la suma de esas probabilidades siempre será igual a 1.

Ahora se procederá a explicar la **función de distribución de probabilidad para una variable aleatoria**. (Véliz C., 2011) Explica que usando la **función de densidad** es posible calcular la probabilidad que un valor de la variable aleatoria esté en el determinado intervalo. Esta definición se complementa con la que da (Freund & Walpole, 1990) Quienes señalan que las áreas situadas debajo de la curva darán las probabilidades relacionados con los intervalos correspondientes situadas en el eje horizontal. En palabras más sencillas, se usa una función de densidad la cual calcula la probabilidad que un valor de X se encuentre en el intervalo. La función de densidad se matematiza de la siguiente forma:

$$P(a \leq x \leq b) = \int_a^b f(x)dx \text{ (1.5.12.)}$$

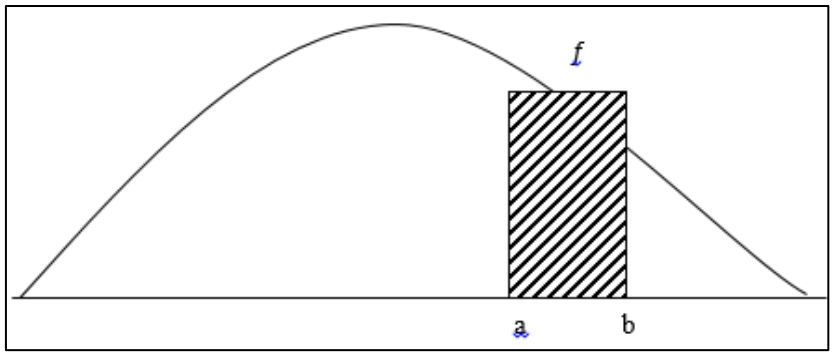


Figura 1.2. Área de la región sombreada =P [a≤X≤b]
Elaboración: Propia
Fuente: (Véliz C., 2011)

Usando la función de densidad se calcula la probabilidad que el valor de una variable aleatoria caiga en el intervalo [a,b]. La función de densidad presenta las siguientes propiedades, explicadas por (Véliz C., 2011):

- Los valores de *f* son mayores o iguales a 0, es decir, *f*(x) ≥0.
- El área por debajo de la gráfica de la función, es decir debajo de la curva y por encima del eje horizontal es 1, lo que quiere decir que la suma de las probabilidades es 1. Lo que equivale a decir: $\int_{-\infty}^{\infty} f(x)dx = 1$.
- La probabilidad que la variable aleatoria *X* tome valores entre a y b se denota como P [a≤*X*≤b] y es igual al área comprendida entre la gráfica (la curva) de *f*, el eje horizontal y las rectas paralelas que pasan por el intervalo [a,b].

Se expone el siguiente ejemplo planteado por (Freund & Walpole, 1990). La función de densidad de probabilidad de la variable aleatoria X está dada por:

$$f(x) = \begin{cases} k.e^{-3x}, & x > 0 \\ 0, & en\ cualquier\ otra\ parte \end{cases}$$

Pide hallar *k* y $P\ (0.5 \leq X \leq 1)$. El ejemplo pide determinar el valor de *k* dada la función de densidad de probabilidad para los valores que se encuentren en el intervalo [0.5, 1]. Además los únicos valores que admite son aquellos mayores de 0 debido a que la probabilidad será 0 para cualquier valor que no sea mayor a 0 indicado así por la expresión *f*(x)=0. También pide calcular la probabilidad de que un valor mayor a 0 se encuentre en el intervalo [0.5, 1].

Solución:

$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} k.e^{-3x}dx = k.\lim_{n\to\infty} \frac{e^{-3x}}{-3}\Big|_0^n = \frac{k}{3} = 1$, para que k/3=1 entonces asumimos que k es 3. Por lo tanto para calcular la probabilidad se efectúa la siguiente integral:

$$P\ (0.5 \leq X \leq 1) = \int_{0.5}^{1} 3e^{-3x}dx = -e^{-3x}\Big|_{0.5}^{1} = -e^{-3} + e^{-1.5} = 0.173$$

Entonces dado la función de densidad *f*(x)=$3e^{-3x}$ podemos calcular que existe la probabilidad de 0.173 de que un valor de la variable aleatoria *X* se encuentre en el intervalo [0.5, 1].

Se ha demostrado entonces que en la teoría estadística se usa una función de distribución de probabilidad para explicar matemáticamente el comportamiento que siguen los valores de una variable aleatoria, pero **existen momentos o también llamadas resúmenes numéricos** que pretenden precisar una descripción más completa y las más usadas son la **esperanza** y la **varianza**. Es importante conocer estos resúmenes debido a que los supuestos de regresión lineal se basan fuertemente en estos resúmenes numéricos.

La esperanza de una variable aleatoria continua *X* se denota con *E(X),* también se le conoce como valor esperado. La esperanza es aquel valor central que sirve de eje para los demás valores ya que los demás valores de la variable están alrededor de él. El concepto se asemeja mucho al promedio aritmético, no obstante, se diferencia en que el valor esperado está ligado a la teoría de probabilidades. La esperanza se matematiza con la siguiente manera:

$E(X) = \sum_i x_i P[X = x_i]$ (1.5.13.)

La esperanza de una variable aleatoria sea continua o discreta se le denomina también como media o mediana, y en palabras de (Véliz C., 2011) es la suma ponderada de los valores de la variable aleatoria, es decir suma los productos de los valores de la variable aleatoria con sus respectivas probabilidades que cada valor tiene de aparecer, y describe la tendencia central que tiene una variable aleatoria sin embargo no muestra la dispersión de sus valores con respecto a ese valor que representa ser la media de la variable aleatoria, el valor esperado. Por lo tanto es necesario calcular la varianza de la variable aleatoria para poder medir cuánto están dispersos los valores de la variable aleatoria.

La varianza es por tanto una medida de dispersión, la cual es el cuadrado de la **desviación estándar** o **desviación típica** y este es el nombre que reciben todas aquellas dispersiones del conjunto de valores de la variable aleatoria con respecto a la esperanza o media. Cuando la desviación estándar es baja entonces, los valores de la variable aleatoria se encuentran cerca a la media, caso contrario sucede cuando la desviación estándar es muy alta. La varianza y la desviación estándar se denota simbólicamente como: σ^2 y σ respectivamente. Debido a que la varianza es el cuadrado de la desviación estándar entonces se le puede definir como el valor esperado del cuadrado de la desviación estándar respecto a su media. Matemáticamente se representa de la siguiente manera:

$$V(X) = \sum_j (x_j - E(X))^2 . P[X = x_j] \quad (1.5.14.)$$

Tanto las ecuaciones (1.5.13.) y (1.5.14.) describen la esperanza y varianza respectivamente para una variable aleatoria discreta. Antes de pasar al caso de la variable aleatoria continua se presenta un ejemplo expuesto por (Véliz C., 2011):

Cuando se invierte en un negocio se gana 2000 dólares con probabilidad de 0.2, se gana 1500 dólares con probabilidad 0.7 y se pierde 3000 dólares con probabilidad 0.1. Se pide calcular el valor esperado y la varianza:

Para calcular el valor esperado denotamos:

$$E(X) = 2000(0.2) + 1500(0.7) - 3000(0.1) = 1150$$

Interpretación: En promedio se gana 1150 dólares si se realiza muchas veces la inversión en ese negocio. Por lo tanto el valor 1150 muestra la tendencia central que seguirán los valores de la variable aleatoria X que muestra la ganancia o pérdida. Para calcular la varianza de X, se visualiza la siguiente tabla:

Ganancia	Probabilidad	$(\text{Ganancia-1150})^2$ x Probabilidad
2000	0.2	144500
1500	0.7	85750
-3000	0.1	1722250
	Varianza	1952500

Tabla 1.3. Varianza y desviación estándar del ejemplo
Elaboración propia
Fuente: (Véliz C., 2011)

Solamente se puede interpretar la desviación estándar no la varianza, y se interpreta de la siguiente manera, en promedio cada valor de la variable aleatoria X se aleja de la media en 1397.31 dólares.

Ahora, tanto para una variable aleatoria discreta o continua, las definiciones de valor esperado y varianza son las mismas, por ello solo expondrán las fórmulas matemáticas del valor esperado y la varianza de una variable aleatoria continua:

$$E(X) = \int_{-\infty}^{\infty} x.f(x)dx \text{ (1.5.15.)}$$

$$V(X) = \int_{-\infty}^{\infty} x^2.f(x)dx - E(X)^2 \text{ (1.5.16.)}$$

Note como ahora para las fórmulas (1.5.15.) y (1.5.16.), las cuales son la esperanza y varianza de una variable aleatoria continua respectivamente, toma en cuenta la función de densidad de probabilidad.

Una vez explicado lo que es una variable aleatoria a continuación se muestra una tabla donde se menciona de manera general los supuestos, en algunos textos puede encontrarse como hipótesis.

Supuestos o hipótesis del modelo de regresión lineal

Supuestos sobre la perturbación aleatoria	El término de error, μ, es una variable aleatoria con esperanza nula, una matriz de covarianzas constantes y diagonal. Y además $Cov(\mu_i, \mu_j) = 0$ cuando $i \neq j$ este es el supuesto de la **no autocorrelación** esto quiere decir que el término de error no tiene relación consigo misma debido a que es una variable aleatoria. Y al ser la varianza constante significa que no cambia y es independiente para cada valor del término de error, este es el supuesto de la **homocedasticidad**.
	El término de error, μ, es una variable aleatoria no observable, implica que la variable endógena sea aleatoria, ya que depende de una variable aleatoria, μ.
	El término de error es una variable aleatoria que sigue una distribución normal, es decir, que el valor esperado del término de error es 0, $E(\mu)=0$, y además tiene una varianza constante. Se le denota de la siguiente manera: $\mu \sim N(0, \sigma^2)$. Este es el supuesto de la **normalidad** de los residuos.
Supuestos sobre los regresores	Las variables explicativas son linealmente independientes, es decir no existe relación lineal exacta entre ellas. Este es el supuesto de **independencia** y cuando no se cumple, el modelo presenta problema de **multicolinealidad**.

	Las variables explicativas son deterministas, es decir se pueden medir y no son inobservables. Sucede así porque su valor es constante y proviene de una muestra tomada en el tiempo y no tienen correlación con el término de error. Este supuesto se le conoce como la **exogeneidad**.
	Las variables no tienen error de medida y además el número de observaciones, *n*, debe ser igual o mayor al número de regresores, *k*.
Supuestos sobre los parámetros	Los parámetros son fijos y además cumplen sus propiedades anteriormente explicadas. Este supuesto quiere decir que los parámetros tienen estabilidad en el tiempo de las estimaciones, de este supuesto surge la teoría de la cointegración. Una teoría muy usada en la estimación de series temporales.
Supuestos sobre la forma funcional	La relación entre la variable dependiente y las variables independientes es lineal. Es el supuesto de la **linealidad**.
	Se asume que el modelo especificado tiene ausencia de error de especificación, significa que se han incluido solamente las variables independientes relevantes para la explicación de la variable dependiente.

Tabla 1.4. Supuestos del modelo de regresión lineal
Elaboración propia
Fuente: (Pérez L., 2012)

En realidad, el cumplimiento supuestos del modelo de regresión lineal son importantes para obtener un buen modelo econométrico capaz de explicar el comportamiento de la variable endógena, sobre todo el supuesto de la **normalidad** de los residuos, debido a que permite la estimación de los intervalos de confianza, las test de hipótesis sobre los parámetros del modelo, cuando no se tiene una distribución normal se suele tener pruebas de hipótesis inválidas, podríamos cometer error tipo 1 o 2. En los siguientes apartados se explicará con un detalle meticuloso más sobre los supuestos, que sucede cuando no se cumplen y porque es importante siempre guardar el cumplimiento de estos supuestos.

1.5.4. Evaluación de la capacidad predictiva o interpretación.

Una vez demostrado el cumplimiento de los supuestos de MCO, se procede a darle un sentido económico a los coeficientes de regresión a través del principio de ceteris paribus y en el mejor de los casos a predecir o pronosticar el comportamiento futuro de la variable endógena mediante el modelo estimado. En realidad en el último paso podemos usar el modelo de regresión especificado como una buena forma de explicar los fenómenos económicos investigados y plantear políticas públicas o cual fuese el objetivo del tema investigado. En lo que concierne a esta guía de estudios solo se pretende enseñar los pasos para la elaboración de modelos econométricos

desde la base de datos hasta la interpretación del modelo. Dependerá del lector construir sus propios modelos econométricos, especificarlos, estimarlos y darles el uso que requiera para su trabajo de investigación.

2. La base de datos y la Encuesta Nacional de Hogares.

Ya se ha explicado que la construcción de la base de datos, resulta ser el primer paso en la elaboración de modelos econométricos debido a que es el correcto manejo de los datos de las variables económicas lo que permite que la estimación de los parámetros muestrales sea consistente y puedan ser usados para la interpretación. Saber construir una base de datos es necesario en la formación no solo de modelos econométricos sino también cuando se pretende elaborar estadísticos descriptivos. Por ello el siguiente capítulo procurará señalar los aspectos importantes cuando se requiere construir una sólida base de datos y también se detallará algunos temas importantes sobre la ENAHO.

2.1. Los datos y las variables

Como detalla (Gil F., 1994)

> *"La mayoría de los autores asumen que el investigador desempeña un papel activo respecto de los datos: el dato es el resultado de un proceso de elaboración, es decir, el dato hay que construirlo."* (Gil F., 1994)

La cita anterior quiere decir que los datos son la información extraída de la realidad mediante una técnica de recolección de datos, por tanto, se puede decir que son hechos que describen sucesos y estos deben ser convertidos en información para ofrecer un significado, es decir, por si mismos no significan nada, pero cuando se les asocia en un contexto adquieren sentido, entonces se habrán convertido en información.

Tal como (Novales, 1998) explica que cuando los economistas se enfrentan a una base de datos el principal problema es que deben organizarla y precisamente una forma de organización es seguir un proceso denominado **muestreo estadístico**. El muestreo se refiere a toda técnica que recoge datos para construir bases de datos que sean capaces de representar a las variables que el modelo requiere. Debido al tipo de naturaleza de la variable es que la técnica de muestreo es distinta.

Por lo general se define a la variable como una característica presente en la realidad que se pretende explicar y al ser cambiante tanto en el tiempo como en el espacio es que se le designa el nombre de "*variable*". Por ejemplo: el número de niños menores

de 5 años en un hogar, el ingreso familiar anual, etc. La clasificación más común para diversos autores señala que una variable puede ser **cualitativa** o **cuantitativa**. Una variable es cuantitativa cuando sus elementos expresan cantidad, se suele emplear el término observación como sinónimo de datos, por ejemplo: el gasto de bolsillo de una familia. Por otro lado, una variable es cualitativa cuando los valores de sus elementos expresan una cualidad, por lo general son **variables dicotómicas**, es decir que toman el valor 1 cuando la observación cumple una cualidad y toma el valor de 0 cuando no la cumple la cualidad estudiada, por ejemplo: cuando se requiere la creación de una variable que permita representar si una persona tiene o no afiliación al SIS, entonces cada persona será una observación y tomarán el valor de 1 cuando cumplan la condición de estar afiliados al SIS y 0 cuando no cumplan la condición.

Las variables cuantitativas pueden dividirse en **variables discretas** o **variables continuas**, la primera es aquella que admite un número contable de observaciones mientras que la segunda no admiten un número contable y en su mayoría se usan intervalos para agrupar sus datos.

2.2. Población y muestra

En las investigaciones empíricas, se trata de elaborar un análisis de datos que no solo recopile datos, sino que además los organice para realizar conclusiones. Estas conclusiones se elaboran a partir de un conjunto de datos que representan a un conjunto más grande de datos, es decir, se utiliza una **muestra** que representa a la **población** para elaborar conclusiones.

Se puede decir que lo que se busca es que la muestra sea representativa ya que esta característica es lo que permite concluir desde unos cuantos datos particulares hacia toda la población. (Pardo, Ruiz, & San Martín, 2009) Explican que es necesario contar con una buena técnica de recojo de datos, debido a que esta técnica es lo que asegurará que la muestra seleccionada representa a la población. A este proceso de utilizar observaciones de una muestra para concluir, describir e inferir a una población, se le conoce como **estadística inferencial**, una definición ya explicada anteriormente, es por tanto que el objetivo de las técnicas de recojo de información es asegurar que la muestra represente a la población para realizar una buena inferencia estadística.

(Moya C., 2007) Alcanza la definición de población.

> *"Población. Es la colección de todos los individuos, objetos u observaciones que poseen al menos una característica común."* (Moya C., 2007)

Algunos ejemplos de población podrían ser:

- Los pesos de sandías que se comercializan en un mercado.
- Los pacientes de un hospital que padecen de TBC.
- Las personas que viven en casas construidas de adobe en una ciudad.

Tal como (Moya C., 2007) Menciona, en estos ejemplos todos tienen una característica en común que los convierte en elementos de la población. (Moya C., 2007) También señala la importancia de determinar la población acorde a su naturaleza y a la extensión del problema bajo estudio. (Moya C., 2007) Se refiere al término "naturaleza" como la característica o materia del estudio, por ejemplo: si quisiéramos estudiar pesos de un conjunto de personas, la naturaleza o característica de la población serían todos los pesos de ese conjunto de personas, y al término "extensión del problema" como la característica que la población debe ser tan extensa y cuantiosa como la investigación lo requiera, por ejemplo: se pide describir las características de los solicitantes de créditos en una ciudad, entonces la población serían todas las personas que solicitan créditos en toda la ciudad.

(Moya C., 2007) Explica las definiciones de **población objeto** y **población objetivo** textualmente:

> *"... entendemos por población objeto, el conjunto de elementos materia de estudio y por población objetivo las diferentes medidas de la característica que nos interesa de la población objeto."* (Moya C., 2007)

Se explica el siguiente ejemplo: Cuando un estudio pretende describir a los infantes menores de 5 años, podemos identificar a la población objeto como el conjunto de todos los infantes menores de 5 años y a la población objetivo como lo que nos interesa medir, por ejemplo: sus pesos al nacer, etc.

Debido a que se necesita realizar inferencias para conocer los parámetros poblaciones, es necesario emplear una muestra. Para (Pardo, Ruiz, & San Martín, 2009) la **muestra** es un subconjunto de elementos de la población, y la señala como las fuente de información para describir las propiedades de la población, por ello es que la muestra

tiene que ser representativa. Esta definición se ve apoyada con una cita textual que se recoge de (Moya C., 2007).

> *"…en otras palabras, nuestro propósito es conocer la población, para lo cual se extrae una muestra de esta."* (Moya C., 2007)

(Lind, Marchal, & Wathen, 2015)Detalla algunas razones para muestrear:

- Establecer contacto con toda la población requiere mucho tiempo.
- El costo de estudiar todos los elementos de una población resulta prohibitivo.
- Es imposible verificar de manera física todos los elementos de la población. Los autores se refieren a que existen poblaciones que son infinitas y que por lo tanto es imposible identificar sus elementos.
- Algunas pruebas son de naturaleza destructiva. Por lo general, las empresas durante sus pruebas de calidad de sus productos eligen una muestra para controlar si cumplen los estándares requeridos, por ejemplo, alguna empresa dedicada a producir pisco, eligen a una parte de la producción para poder examinar y degustar su aroma y sabor; si eligieran a toda la producción entonces se beberían todo el licor producido y no habría producción para la comercialización.
- Los resultados de la muestra son adecuados. En algunos casos la utilización de toda la población no sea necesaria para algún estudio requerido. (Lind, Marchal, & Wathen, 2015) ejemplifican: El gobierno de Estados Unidos decide usar una muestra de tiendas de alimentos en vez de toda la población de tiendas de alimentos para calcular el índice mensual de precios de los alimentos, debido a que es poco probable que la inclusión de toda la población haga cambiar significativamente al índice de precios porque el precio de cada una de las cadenas de alimentos varía en centavos con respecto a la otra.

2.3. Técnicas de muestreo

Previamente a mostrar las técnicas de muestreo se debe tener en cuenta los conceptos de reposición y probabilidad de selección. (Pérez L., 2005) Explica que estos son criterios de selección de muestras y se clasifican de la siguiente forma:

- Criterio de probabilidades de selección:
  - Con probabilidades iguales: Todos los elementos de la población tienen la misma probabilidad de ser seleccionadas para pertenecer a la muestra.
  - Con probabilidades desiguales: Al menos dos elementos tienen diferentes probabilidades de pertenecer a la muestra.
- Criterio a la mecánica de selección:
  - Muestreo con reposición: Cada unidad que es seleccionada para pertenecer a la muestra se repone a la población antes de volver a extraer una muestra, la estructura poblacional permanece invariante. Ejemplo: Suponga que se realiza un sorteo entre 45 alumnos donde se sortean 3 libros entonces habrá 3 sorteos, cuando el primer sorteo otorgue un ganador, este alumno ganador del primer sorteo volverá a participar en el segundo y tercer sorteo. De esta forma siempre habrá una población que no cambiará.
  - Muestreo sin reposición: Cada elemento que es extraído para pertenecer a la muestra no vuelve a la población antes de volver a extraer una muestra, es decir el número de la población cada vez va disminuyendo. Ejemplo: Siguiendo el ejemplo anterior, el primer ganador ya no participará del segundo sorteo ni del tercer sorteo, por lo que habrá 44 y 43 alumnos sorteándose los libros respectivamente de cada sorteo.

(Pérez L., 2005) Sigue explicando que podemos combinar estos criterios y podemos obtener 4 tipos de muestreo:

- Muestreo con reposición y probabilidades iguales.
- Muestreo sin reposición y probabilidades iguales.
- Muestreo con reposición y probabilidades desiguales.
- Muestreo sin reposición y probabilidades desiguales.

(Otzen & Manterola, 2017)Detalla que existen dos tipos de técnicas de muestreo y son: las técnicas de muestreo probabilístico y no probabilístico. En esta guía de estudios, se detalla las técnicas de muestreo probabilístico, sin embargo, se alcanza una definición de las técnicas de muestreo no probabilística, (Moya C., 2007) Define que estas técnicas también llamadas conveniencia o de juicio, tiene base en el conocimiento y la opinión para identificar los elementos que deben incluirse, por lo general dadas por un experto en

la materia. Además, existen 3 tipos de muestreo no probabilística y (Otzen & Manterola, 2017) las mencionan:

- **Intencional**. Permite seleccionar casos que cumplan las características limitando la muestra a esos casos, se suele usar cuando la población es muy variante y la muestra pequeña.
- **Por conveniencia**. Se seleccionan los casos que aceptan ser incluidos en la muestra, por lo general por conveniente accesibilidad y proximidad.
- **Accidental o consecutivo**. Selecciona los casos hasta completar el número de muestra deseado, esto se elige casualmente.

El muestreo probabilístico consiste en el uso de las probabilidades para obtener la muestra, en palabras de (Moya C., 2007) cuando los elementos tienen probabilidad alguna de pertenecer a la muestra, ya sea una probabilidad igual o desigual, estamos ante el muestreo probabilístico. A continuación, se presenta una tabla que contiene las diferentes técnicas de muestreo haciendo énfasis en las técnicas de muestreo probabilísticas.

Técnicas de muestreo no probabilística	• Intencional • Por conveniencia • Accidental o consecutivo
Técnicas de muestreo probabilística	• **Muestreo aleatorio simple** Este es el diseño más básico de todos y consiste en seleccionar *n* elementos de muestreo de tal manera que cada elemento tiene la misma oportunidad de ser seleccionada, la probabilidad para cada elemento puede ser igual o desigual. • **Muestreo estratificado** Cuando en la población existen estratos o clases con características únicas que de cierta forma constituyen una población dentro de la población, es decir una subpoblación, entonces se debe hacer uso del muestreo estratificado. Esto sucede porque la población total es demasiado heterogénea, por ello es que se debe dividir a la población en grupos homogéneos. Nótese que para lograr tal división se debe tener en cuenta una variable, como los ingresos de cada grupo, las edades de cada grupo, el nivel socioeconómico de cada grupo, etc. En otras palabras, el muestreo estratificado divide la población que es heterogénea en grupos lo más homogéneos posibles denominados estratos.

- Muestreo por conglomerados

El muestreo por conglomerados se usa en las poblaciones particularmente grandes donde los elementos están dispersos desde un punto de vista geográfico.

En el muestreo por conglomerados se divide a la población con el fin de estudiar varios elementos, no se debe confundir con el muestreo estratificado donde cada división tiene características propias que la hace un estrato. En el muestreo conglomerado se divide (casi siempre por zona territorial) para estudiar unidades que representan un grupo de elementos, por ejemplo familias, comunidades, etc. Por ejemplo imagine que se quiere investigar el ingreso de las familias de Lambayeque entonces según el muestreo por conglomerados se seleccionan 100 familias (conglomeraciones) sin importar su estrato.

Este muestreo a diferencia del muestreo estratificado, permite grupos lo más heterogéneos posible.

- Muestreo sistemático

Tal como su nombre indica, el muestreo que realiza sigue un orden en el que cada elemento que se selecciona está en el mismo lugar dentro de la zona que ocupa la primera unidad seleccionada en la primera zona. Por ejemplo: se pretende seleccionar una muestra de 40 elementos con una población de 1200 elementos, entonces k=1200/40 = 30. La muestra se obtiene tomando cada 30-ésima unidad de la población.

Tabla 2.1. Técnicas de muestreo
Elaboración propia
Fuente: (Pérez L., 2005), (Moya C., 2007) & (Scheaffer, Mendenhall III, & Lyman O., 2007)

2.4. Determinación del tamaño muestral

(Gallardo & Moreno, 1999) Detalla cómo se logra determinar el tamaño de la muestra, pero previamente señalan que el investigador debe considerar los antecedentes del estudio en cuestión, además de tener en cuenta si se tienen los recursos económicos para lograr tener los resultados de la muestra, pero sobre todo considerar en todo momento a los objetivos de la investigación para determinar el tamaño de la muestra.

(Aguilar-Barojas, 2005) Detalla que para calcular el tamaño de la muestra depende del tipo de investigación, los siguientes son para investigaciones descriptivas de tipo cualitativo.

(Gallardo & Moreno, 1999)Explican la fórmula para determinar el tamaño muestral y es:

$$n = \frac{n_o}{1+\frac{n_o}{N}} \text{ (2.4.1.)}$$

La ecuación anterior determina el tamaño de la muestra *n*, además n_o es la primera aproximación al tamaño de la muestra y se calcula con:

$$n_o = \frac{Z^2 pq}{d^2} \text{ (2.4.2.)}$$

Donde *Z* es el nivel de confianza y se obtiene de las tablas de la distribución normal, ejemplo:

- Para un nivel de confianza del 90% *Z*=1.645
- Para un nivel de confianza del 95% *Z*=1.96
- Para un nivel de confianza del 99% *Z*=2.58

Además, *p* y *q* representan la probabilidad que ocurra el evento y que no ocurra el evento, respectivamente; recordar que *q* se calcula con *(1-p)*. Si no se conocen sus valores, se puede asumir sus valores de 0.5 para ambos. Y finalmente, *d* es el margen de error y *n* es el tamaño de la muestra. **Esta fórmula se utiliza cuando se tiene una población infinita.**

(Aguilar-Barojas, 2005) Detalla la fórmula para calcular el tamaño de la muestra para una investigación de tipo cualitativo con población finita:

$$n = \frac{N\, Z^2 pq}{d^2(N-1)+Z^2 pq} \text{ (2.4.3.)}$$

Agregamos *N*, el cual representa el tamaño de la población. También muestra las fórmulas para calcular la muestra en los trabajos de investigación de tipo cuantitativo para poblaciones infinita y finita, a continuación, se muestran respectivamente:

$$n = \frac{Z^2 S^2}{d^2} \text{ (2.4.4.)}$$

$$n = \frac{NZ^2 S^2}{d^2(N-1)+Z^2 S^2} \text{ (2.4.5.)}$$

La fórmula (2.4.4.) se debe usar para las poblaciones infinitas y la formula (2.4.5.) para las poblaciones finitas.

2.5. Técnicas de recolección de datos

(Hernández S., Fernández C., & Baptista L., 2010) Detallan cuales son las formas de recolectar datos cuantitativos y cualitativos, además explican que recolectar datos requiere ejecutar un plan que nos permita reunir los datos con un objetivo específico. Especifican los pasos del plan en forma de preguntas:

- ¿Cuáles son las fuentes de donde se obtendrán los datos?, se refiere a que se debe tener en cuenta de donde vendrán los datos recopilados.
- ¿En dónde se localizan tales fuentes?, casi siempre la muestra seleccionada tiene la respuesta.
- ¿A través de qué medio o método vamos a recolectar los datos? Esta es la pregunta que implica definir el medio para recoger datos guardando siempre confiabilidad, validez y objetivos.
- ¿Cómo serán preparados para que puedan analizarse?

(Acosta G., Andrada F. Julián, & Fernández M., 2009) Definen al proceso de asignar números, símbolos o valores a las propiedades de los objetos o eventos de acuerdo con reglas, como **medir,** sin embargo, algunos aspectos son tan abstractos que es difícil o ya de por sí, imposible de medir, ejemplo: la disonancia cognitiva, la pareja ideal, el clima organizacional, etc. Un instrumento de medida adecuada acorde a (Hernández S., Fernández C., & Baptista L., 2010) es aquel que registra los datos de las variables que se quiere investigar. Debe cumplir tres requisitos esenciales:

- **Confiabilidad**, se refiere al grado en que su aplicación repetida a los mismos objetos reproduce los mismos resultados.
- **Validez**, se refiere al grado en que un instrumento realmente mide la variable. Podemos validar la medida a través del contenido, relacionada con el criterio y relacionada con el constructo. La suma de estas tres partes resulta la validez total
- **Objetividad**, se refiere al grado en que el instrumento es permeable a los sesgos. La objetividad es el requisito más difícil de lograr.

Cuando se tiene una investigación de tipo cuantitativo existen algunos instrumentos que permiten la recolección de datos, entre ellos: el cuestionario. Es el instrumento más usado para recolectar datos debido a su congruencia con el

planteamiento del problema e hipótesis. El cuestionario tiene dos tipos de preguntas: **preguntas cerradas** y **preguntas abiertas**. Las primeras contienen opciones de respuesta, las cuales han sido previamente establecidas, acortando las respuestas que el encuestado puede dar a solo una de un conjunto de posibles respuestas. Pueden incluir dos o varias opciones de respuesta. Por otro lado, las preguntas abiertas ofrecen una posibilidad ilimitada de opciones de respuesta, ya que no delimitan las alternativas de respuesta.

(Hernández S., Fernández C., & Baptista L., 2010) Explican cuando es recomendable el uso de cada una de ellas. Las preguntas cerradas son fáciles de codificar y preparar para su análisis y requieren menor tiempo de contestar para los encuestados además que ofrece disminuir las ambigüedades y comparar las respuestas, sin embargo este tipo de preguntas pueden representar una desventaja y es que pueden ser muy limitantes ocasionando que algunos encuestados no sienten que su respuesta está expresada en las alternativas, entonces se puede intuir que si bien es fácil de responder, la calidad de respuesta depende de la calidad de redacción de preguntas del encuestador. Por otro lado las preguntas abiertas proporcionan información más amplia y sirven para profundizar sobre motivos y razones de gustos y preferencias. Sin embargo, el hecho que sea tan amplias sus posibles respuestas provoca que se haga difícil que puedan ser codificadas y preparadas para el análisis.

2.6. Errores de la recolección de datos

(Scheaffer, Mendenhall III, & Lyman O., 2007) Explican que las encuestas pueden presentar errores dado que los resultados obtenidos podrían estar incorrectos o incompletos. Clasifican los errores de encuesta principalmente en **errores ajenos al proceso de observación** y **errores del proceso de observación**. El primer tipo de error aparece cuando las observaciones sólo representan una parte de la población objetivo y el segundo tipo de error aparece cuando las observaciones no son representativas, es decir se desvían de la verdad.

2.6.1. Errores del proceso de observación.

Cuando los errores se producen en el proceso de recolección de datos, pueden ocasionados por el entrevistador, el entrevistado, el instrumento de medida o el método para la relación de datos. (Scheaffer, Mendenhall III, & Lyman O., 2007) Explican que los entrevistadores pueden afectar los resultados de forma directa, y ejemplifica: si el

entrevistador emplea un énfasis distinto al que debería usar, entonces podría dar un sentido a la pregunta diferente al que se espera, por lo que el entrevistado puede dar una respuesta equivocada. Continúan explicando que la respuesta también puede verse afectada por la afinidad del entrevistado con el entrevistador ya que muchas veces la mayoría de entrevistados no desean ser descorteses con el entrevistador al momento de dar una respuesta o si quiera aceptar ser entrevistados, por lo general tratarán de agradar con sus respuestas al entrevistador.

Uno de los motivos por el cual los entrevistados también pueden ser la fuente del error es que cada uno de ellos tiene una idea y percepción distinta a cada pregunta que se le tiene enfrente. (Scheaffer, Mendenhall III, & Lyman O., 2007) Recomiendan el uso de tarjetas con las preguntas escritas para que el encuestado no pierda el sentido de la pregunta y tenga una percepción clara. Además clasifican los sesgos que se pueden obtener de los encuestados:

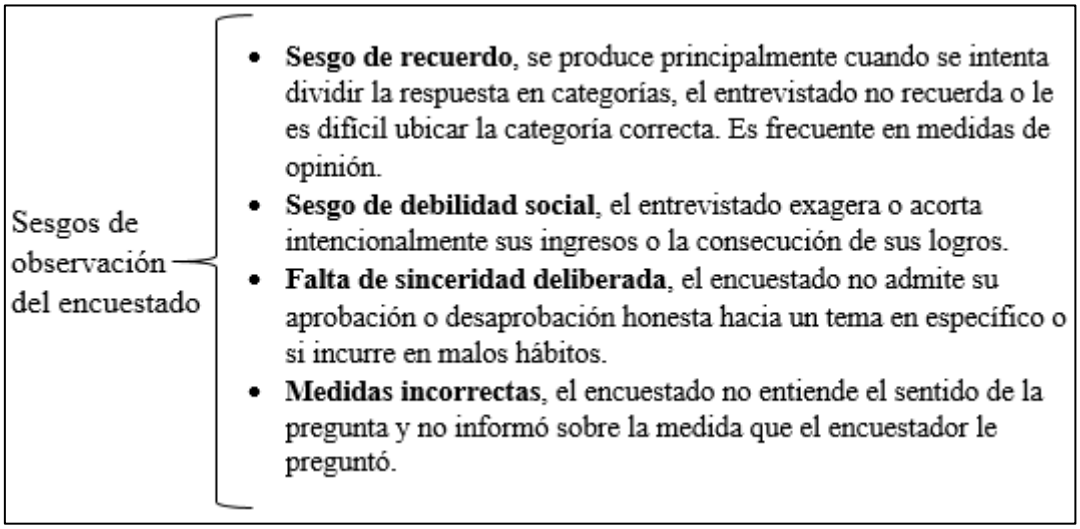


Figura 2.1. Sesgos de observación del encuestado.
Elaboración propia
Fuente: (Scheaffer, Mendenhall III, & Lyman O., 2007)

El instrumento también puede ser un origen de problemas de error por observación, se debe definir bien las medidas que se desea investigar.

> *"Las respuestas poco precisas se deben normalmente a errores de definición en las preguntas de la encuesta."* (Scheaffer, Mendenhall III, & Lyman O., 2007)

(Scheaffer, Mendenhall III, & Lyman O., 2007) Hablan sobre los 4 tipos de recopilación de datos:

2.6.1.1. Entrevistas personales.

Las observaciones que se obtienen de entrevistas, normalmente se requiere de preguntas preparadas y el registro de sus respuestas a menudo son grabadas. Suelen ser muy ventajosas porque el entrevistador puede señalar cuál es el correcto sentido que se quiere de la pregunta, sin embargo, el correcto uso de las entrevistas depende de cuan bien entrenado está el entrevistador, salirse del protocolo, olvidar cual es el objetivo de la entrevista, realizar expresiones faciales y énfasis en preguntas pueden manipular las respuestas y provocar un sesgo en los datos muestreados.

2.6.1.2. Entrevistas telefónicas.

Suelen ser menos costosas que las entrevistas personales, sin embargo, el problema de esta forma de entrevista existe cuando no consigue un marco que representa a la población. En algunas ocasiones se marca los dígitos aleatoriamente con el fin de tener representatividad y porque algunos números ya no son del hogar que la guía indica que sí; sin embargo, esta técnica parece producir muestras insesgadas de hogares en poblaciones objetivo y evita problemas que puedan deberse por el uso de la guía telefónica.

2.6.1.3. Cuestionarios auto administrados.

El principal problema que existe es la “no respuesta”, es decir que el encuestado deje preguntas en blanco y la imposibilidad de poder identificar al encuestado no deja lugar a la oportunidad de completar el cuestionario. Otro problema es confiar en que el encuestado no se equivoque al momento de responder, para solucionar estos problemas se debe tener en cuenta siempre el empleo de preguntas lo más cortas posible, sin caer en redundancias o términos incomprensibles.

2.6.1.4. Observación directa.

Este método se emplea cuando no se quiere estudiar personas y debe colocar a una persona a contar el número de elementos. El problema que ocurra en este método, es la posibilidad de errores en la observación.

Como se puede dar cuenta, estos problemas de error del proceso de observación, existen cuando se intenta medir un objeto o a las personas, tener siempre en cuenta la medida correcta y lo que se quiere conseguir, es la clave para evitar caer en estos errores.

2.7. Encuesta Nacional de Hogares (ENAHO)

Una de las fuentes de información en Perú es la Encuesta Nacional de Hogares, que, desde mayo del 2003 hasta la actualidad, viene recogiendo datos sobre las condiciones de vida y pobreza de los hogares de forma continua. Esta encuesta es llevada a cabo por el Instituto Nacional de Estadística e Informática (INEI), que ya desde 1995 empezaba a recolectar datos mediante ENAHO pero de una forma muy diferente a la que se conoce en la actualidad. En sus inicios, en el año 1995, tenía una frecuencia trimestral y planteaba variables sobre condiciones de vida y pobreza de los hogares como variables educativas, salud, gasto, ingreso, etc.

No fue hasta 1997 que con el auspicio del Banco Interamericano de Desarrollo (BID), Banco Mundial (BM) y la Comisión de Económica para América Latina y el Caribe (CEPAL) quienes fortalecieron el programa Mejoramiento de Encuestas y de la Medición de las Condiciones de Vida (MECOVI) del INEI. En aquel entonces se ejecutaban 4 encuestas trimestrales y cada una de ellas correspondía a un tema en específico: en el primer trimestre se recolectaba información de fecundidad y salud, en el segundo se recolectaba información sobre educación y programas sociales, en el tercer trimestre se recolectaba información sobre el empleo y en el último trimestre la información recolectada era sobre el gasto del hogar. Esta modalidad estuvo en vigencia hasta 2002 y tenía como finalidad servir de fuente de información para el seguimiento de la realidad demográfica. Sin embargo, era necesario mejorar los lineamientos de recolección de datos y con la ayuda de varios expertos del MECOVI, el asesoramiento de la Organización Internacional del Trabajo (OIT) y el Ministerio de Trabajo y Promoción del Empleo (MTPE) en el 2003 se empezó a ejecutar el ENAHO de la forma que se conoce hasta hoy.

Al aplicarse en 2003 una encuesta única y continua es que ahora se podía contar con una nueva medición: la dimensión temporal, permitiendo medir los choques que hacen frente los hogares, tales como económicos, sociales, demográficos, etc. Además, que ahora con la implementación de indicadores de pobreza y empleo, es que se podía monitorear la eficiencia de los programas sociales y el seguimiento a la pobreza.

Es cierto que ha recibido mejoras desde el 2003 hasta la actualidad, pero es sin lugar a dudas la del 2003 la más importante. El ENAHO desde su creación hasta ahora tiene la finalidad de servir como fuente de información sobre las condiciones de vida de

los hogares a fin de realizar mediciones sobre la pobreza y ejecutar estudios sobre la sociedad en un determinado espacio geográfico y temporal.

La aplicación del ENAHO alcanza los 24 departamentos del país y la Provincia Constitucional del Callao, y tiene como población objetivo las viviendas y sus ocupantes residentes en el área urbana y rural excluyendo a la población que vive en cuarteles, campamentos y tampoco toma en cuenta a los residentes de viviendas colectivas como hoteles, hospitales, asilos, cárceles y claustros, etc.

El ENAHO tiene un tipo de muestra probabilística, de áreas, estratificada, multietápica e independiente en cada departamento de estudio. En el 2008 se implementó las muestras panel de viviendas mientras que en las muestras no panel se visitan cada año los mismos conglomerados en el mismo mes de encuesta, pero distintas viviendas. La siguiente tabla muestra los temas investigados y el número de preguntas de cada uno de los temas:

1. Caratula (7 preguntas)
2. Características de la vivienda y el hogar (34 preguntas)

2.1. Vivienda (7 preguntas)
2.2. Hogar (27 preguntas)

3. Características de los miembros del hogar (20 preguntas)
4. Educación – Para personas de 3 años y más de edad (43 preguntas)
5. Salud – Para todas las personas (32 preguntas)
6. Empleo e Ingreso – Para personas de 14 años y más de edad (87 preguntas)

6.1. Condición de actividad – Semana Pasada (4 preguntas)
6.2. Ocupados
6.2.1. Ocupación principal (14 preguntas)
6.2.2. Ocupación secundaria (9 preguntas)
6.2.3. Total horas (6 preguntas)
6.2.4. Búsqueda de otro empleo (2 preguntas)
6.2.5. Desocupados (7 preguntas)
6.2.6. Trabajo anterior (3 preguntas)
6.3. Ingreso por trabajo del Hogar
6.3.1. Ocupación Principal (7 preguntas)
6.3.1.1. Por trabajo dependiente
6.3.1.2. Por trabajo independiente
6.3.2. Ocupación Secundaria (7 preguntas)
6.3.2.1. Ingresos por trabajo dependiente
6.3.2.2. Ingresos por trabajo independiente
6.3.3. Ingresos Extraordinarios por trabajo dependiente (ocupación principal y/o secundaria) (1 pregunta)

6.4. Ingreso por trabajo del productor agropecuario (25 preguntas)
6.5. Ingreso por trabajo del trabajador independiente informal (22 preguntas)
6.5.1. Características básicas del negocio o establecimiento
6.5.2. Producción de bienes
6.5.3. Comercio
6.5.4. Servicios
6.5.5. Otros gastos
6.5.6. Características de la mano de obra y empleo
6.5.7. Hoja de control
6.6. Ingresos por transferencias corrientes (últimas 6 meses) – 1 pregunta
6.7. Ingresos por rentas de la propiedad (últimos 12 meses) – 1 pregunta
6.8. Ingresos extraordinarios (últimos 12 meses) – 1 pregunta

7. Sistema de pensiones (2 preguntas)

8. Etnicidad (3 preguntas)

9. Desplazamiento de la población a otros distritos por trabajo (1 pregunta)

10. Inclusión Financiera (4 preguntas)

11. Gastos del hogar (82 preguntas y 329 ítems)

11.1. Alimentos (26 preguntas)

11.1.1. Gastos en alimentos y bebidas consumidas dentro del hogar (últimos 15 días), (5 preguntas - 203 ítems)
11.1.2. Alimentos para consumir dentro del hogar obtenidos de instituciones benéficas (últimos 15 días) (7 preguntas - 3 ítems)
11.1.3. Alimentos consumidos fuera del hogar obtenidos de instituciones benéficas (Menores de 14 años) (7 preguntas - 3 ítems)
11.1.4. Alimentos consumidos fuera hogar obtenido de restaurantes, ambulante, etc. (7 preguntas - 4 ítems)

11.2. Otros gastos

11.2.1. Mantenimiento de la vivienda (mes anterior) (5 preguntas - 15 ítems)
11.2.2. Gastos en transportes y comunicaciones (mes anterior) (5 preguntas - 13 ítems)
11.2.3. Gastos en transportes y comunicaciones (semana anterior) (7 preguntas - 4 ítems)
11.2.4. Gastos en servicios a la vivienda (mes anterior) (4 preguntas - 8 ítems)
11.2.5. Esparcimiento, diversión y servicios de cultura (mes anterior) (5 preguntas - 8 ítems)
11.2.6. Bienes y servicios de cuidados personales (mes anterior) (5 preguntas - 11 ítems)
11.2.7. Vestido y calzado (últimos 3 meses) (5 preguntas - 7 ítems)
11.2.8. Gastos de transferencia (últimos 3 meses) (2 preguntas - 9 ítems)
11.2.9. Muebles y enseres (últimos 12 meses) (5 preguntas - 6 ítems)
11.2.10. Otros bienes y servicios (últimos 12 meses) (5 preguntas - 11 ítems)
11.2.11. Equipamiento del hogar (7 preguntas - 22 ítems)
11.2.12. Venta de inmuebles, equipos (1 pregunta - 2 ítems)

12. Programas sociales de ayuda alimentaria (7 preguntas)

13. Programas sociales no alimentarios (4 preguntas)

14. Participación ciudadana (6 preguntas)

15. Módulo de opinión (45 preguntas)
15.1. Gobernabilidad (Personas de 18 años y más de edad) (3 personas)
15.2. Corrupción (Personas de 18 años y más de edad) (5 preguntas)
15.3. Democracia (Personas de 18 años y más de edad) (9 preguntas)
15.4. Discriminación (Personas de 18 años y más de edad) (2 preguntas)

15.5. Corrupción (Solo para el jefe/a del hogar y cónyuge) (1 pregunta)
15.6. Acceso a la justicia (Solo para el jefe/a del hogar y cónyuge) (6 preguntas)
15.7. Percepción del hogar (Solo para el jefe/a del hogar y cónyuge) (3 preguntas)
15.8. Percepción de los programas no alimentarios (Jefe/a del hogar y cónyuge) (2 preguntas)
15.9. Percepción de los programas alimentarios (Jefe/a del hogar y cónyuge) (2 preguntas)
15.10. Nivel de vida / Situaciones adversas (Solo para el jefe/a del hogar y cónyuge) (9 preguntas)
15.11. Educación de los padres del jefe/a del hogar (1 pregunta)

Tabla 2.2. Temas del ENAHO
Elaboración propia
Fuente: Ficha Técnica - ENAHO

Es muy común el uso de factores de expansión en la metodología del procesamiento de datos y el diseño muestral del ENAHO no es ajeno a esto. El factor básico de expansión que usa el INEI en la ENAHO es el inverso de la probabilidad final de selección cuyo cálculo toma en cuenta las etapas de selección de la muestra del ENAHO, en el área urbana se realizan 3 etapas de selección, en el centro poblado, el conglomerado y la vivienda. Mientras que para el área rural son dos tipos de selección para centros poblados rurales son: CP de 500 a 2000 hab. También llamado AER SIMPLE y el AER compuesto. En los archivos tanto de SPSS y STATA, el factor de expansión se representa con la variable FACTOR07.

El uso del factor de expansión sirve para hacer proyecciones desde la muestra hacia la población, es decir, a la muestra se le concibe un peso ponderado el cual se puede interpretar como el número de elementos de la población que el elemento de la muestra representa. Las siguientes tablas obtenidas de STATA pueden ayudar a tener una mejor perspectiva del uso de factores de expansión:

	estrato socio-económico						
pobreza	a	b	c	d	e	rural	Total
pobre extremo	0	0	3	27	139	1,160	1,329
pobre no extremo	0	7	80	457	1,433	3,176	5,153
no pobre	552	964	2,164	6,840	6,932	8,254	25,706
Total	552	971	2,247	7,324	8,504	12,590	32,188

Figura 2.2. Tabla del nivel de pobreza por estratos socioeconómicos
Elaboración propia
Fuente: Base de datos del ENAHO

pobreza	estrato socio-económico						
	a	b	c	d	e	rural	Total
pobre extremo	0	0	2,041.88	4,341.1	41,880.4	224,732.7	272,996.1
pobre no extremo	0	1,882.02	26,541.97	171,099.4	490,530.9	553,189.6	1243243.8
no pobre	95,587.69	260,701.8	749,803.5	2309166.7	2318486.6	1,260,551	6994297.3
Total	95,587.69	262,583.8	778,387.3	2484607.2	2850897.9	2038473.3	8510537.2

Figura 2.3. Tabla del nivel de pobreza por estratos socioeconómicos con factor de expansión
Elaboración propia
Fuente: Base de datos del ENAHO

Tanto las figuras 2.2. Y 2.3. Representan el nivel de pobreza por estratos socioeconómicos, la diferencia es que en la primera tabla no se ha usado el factor de expansión, por lo que el total es de 32188 observaciones el cual es el total de la muestra. Sin embargo, en la segunda tabla se muestra como el uso del factor de expansión proyecta el total de la muestra hacia el total de la población y no solo el total sino también el número de pobres y no pobres en cada estrato socioeconómico.

El factor de expansión se aplica teniendo en cuenta el número de miembros en el hogar, de tal forma que $facp = factor07 * mieperho$, donde la variable $mieperho$ representa el número de miembros en el hogar en la base de datos de ENAHO, posteriormente se explicará a profundidad cómo aplicar el factor de expansión.

Otro proceso de análisis de datos que brinda la ENAHO es la deflactación, el cual consiste en transformar los valores monetarios nominales en valores reales mediante el índice de precios. Este proceso permite realizar comparaciones en precios constantes de un determinado periodo. El diccionario del ENAHO indica que una variable está deflactada cuando el nombre de la variable empieza con la letra D.

Finalmente, el proceso de imputación se utiliza cuando no se tiene una observación registrada y se debe asignar un valor mediante observaciones que sí están registradas. El diccionario del ENAHO indica que una variable está imputada y deflactada cuando el nombre de la variable empieza con la letra I.

3. Análisis Clásico de Regresión Lineal

Cuando por fin se han seleccionado las variables que conciernen al estudio, se han recolectado los datos y se han procesado en tablas y gráficos, es cuando estaremos listos para estimar los parámetros del modelo econométrico. Sin embargo, para entender las distintas metodologías de estimación es necesario comprender un tema que puede ocasionar estrés en los estudiantes de economía: **el Análisis Clásico de Regresión Lineal**.

Previamente se ha definido a la econometría, como una mezcla entre tres ciencias: la teoría económica, la estadística inferencial y las matemáticas, también se dijo que el análisis de las ecuaciones como una forma de expresar la conducta de las distintas variables económicas para simbolizar sus relaciones da lugar al empleo de modelos económicos y posteriormente modelos econométricos. Para (Novales, 1998) estas ideas se complementan con el hecho que la econometría tiene por objetivo: especificar y estimar un modelo con el propósito de cuantificar las relaciones entre la variable dependiente y la variable o variables independientes. Para ello, la econometría tiene una fuerte base en la estadística inferencial, pues se parte de una muestra para inferir sobre la población.

El análisis clásico de regresión lineal permite cuantificar las relaciones entre las variables del modelo econométrico, sin embargo debemos recordar que existe una parte de la ecuación que no podremos medirla pero igualmente influye sobre la variable dependiente, a esta parte del modelo se le conoce como término de error o término de perturbación. (Cid S., Mora C., & Valenzuela H., 1990) Explican el término de error:

> *"En rigor, el término ϵ representa nuestra incapacidad para predecir en forma exacta el comportamiento de la variable aleatoria Y. Lo anterior significa que ϵ_i resume toda la imprecisión de estos valores y por tanto la variabilidad de Y es exactamente la de ϵ. A este término lo llamaremos error aleatorio o simplemente error."* (Cid S., Mora C., & Valenzuela H., 1990)

Pero ¿de dónde sale ϵ_i? (Véliz C., 2011) Especifica la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (3.1.)$$

Donde señala a la expresión $\beta_0 + \beta_1 x$ como la parte estructural, mientras el término de perturbación representado con ϵ es una variable aleatoria con distribución normal, tiene media 0 y varianza constante. La expresión anterior hace caso al supuesto

de normalidad de los residuos, un supuesto muy útil e importante pues este es el supuesto que permite la correcta estimación de los parámetros. La ecuación 3.1. Es una ecuación de regresión simple, debido a que solo usa dos variables, pero es más útil la inclusión de más variables explicativas en el modelo, dando lugar a la ecuación de regresión múltiple, expresada en su forma matemática como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \text{ (3.2.)}$$

Donde las betas son los coeficientes de regresión, que miden la relación de cada una de las variables explicativas con la variable independiente. Ambas ecuaciones tanto 3.1. Cómo 3.2. Son funciones de regresión poblacionales, para estimar los parámetros de las ecuaciones se debe explicar los fundamentos de la regresión poblacional. A continuación, se procede a explicar la **función de regresión poblacional** y **la función de regresión muestral**.

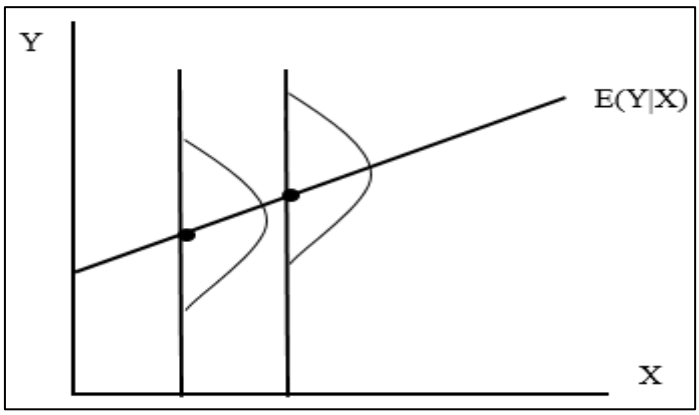
3.1. Análisis de Regresión Simple

3.1.1. Función de regresión poblacional.

Recordemos que la población es el conjunto de todos los elementos, mientras que la muestra es solo una parte que representa el total de los elementos, que mediante una determinada técnica de muestreo se logra extraer una muestra que cumpla condición de representatividad. En econometría, la función de regresión poblacional se expresa de la siguiente manera:

$$E(Y|X_i) = f(X_i) \text{ (3.1.1.)}$$

La ecuación (3.1.1.) denota de manera simbólica que $E(Y|X_i)$ es una media condicional que depende de una la función de X_i y X_i es un valor de X. El enunciado anterior significa que el valor esperado de Y dado X depende de cualquier valor de X. Gráficamente se representa de la siguiente forma:



Gráfica 3.1. Línea de Regresión Poblacional.
Elaboración propia
Fuente: (Gujarati & Porter, 2010)

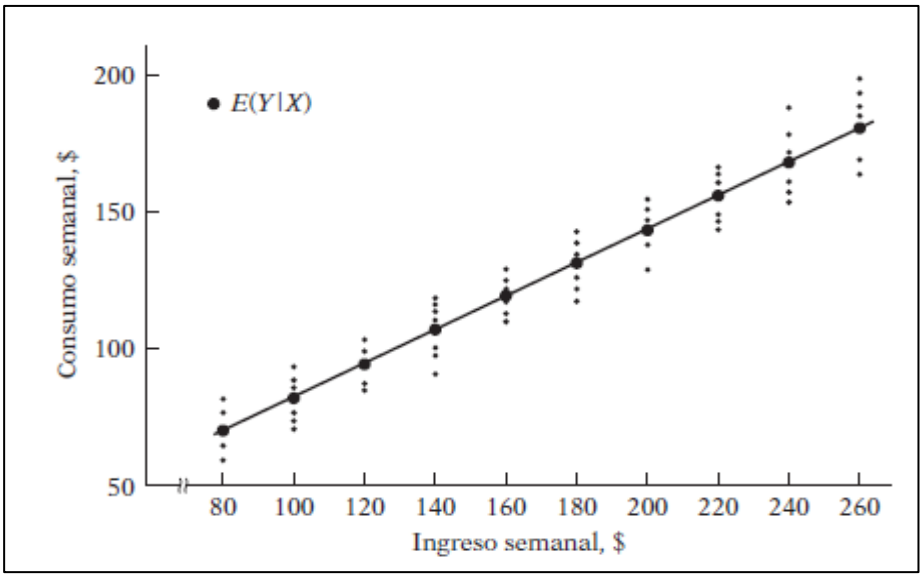
Donde cada punto de la recta en el gráfico 3.1. Representa cada valor esperado condicionado de Y dado cada valor de X. Para explicar mejor lo anteriormente expuesto, se procede a usar el ejemplo de (Gujarati & Porter, 2010). Dada una comunidad cualquiera se tiene la información sobre el ingreso semanal y el consumo semanal de 60 familias **que representan ser toda la población**, la variable dependiente será el consumo semanal mientras la variable explicativa será el ingreso semanal. La siguiente tabla muestra la información:

X → / Y ↓	80	100	120	140	160	180	200	220	240	260
Consumo familiar	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
Total	325	462	445	707	678	750	685	1043	966	1211
Media condicional de Y dado X	65	77	89	101	113	125	137	149	161	173

Tabla 3.1. Ingreso y consumo familiar semanal.
Elaboración (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

En la tabla anteriormente mostrada, se ha agrupado a la población acorde a su nivel de ingreso semanal (*X*) donde cada grupo tiene a un número determinado de

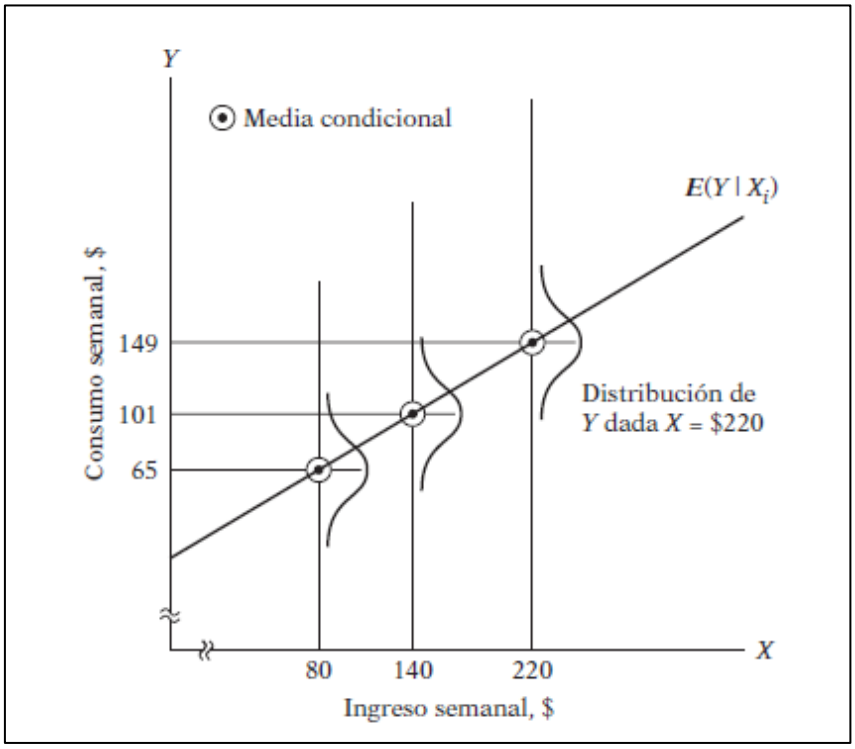
familias, siendo 10 grupos o subpoblaciones, en consecuencia existen 10 valores fijos de *X*. (Gujarati & Porter, 2010) Continúan explicando que al existir 10 valores fijos de *X* es que también encontramos 10 valores esperados de *Y* dado cada valor de *X*. Para que quede más claro, observe el primer grupo donde existen 5 familias que tienen el mismo ingreso semanal siendo 80 dólares, este sería el primer valor fijo de *X*, sin embargo las 5 familias que conforman el primer grupo tienen diferentes niveles de consumo semanal siendo el menor 55 dólares y el mayor 75 dólares, si nosotros calculamos la media del primer grupo obtendremos que cada familia que conforma el grupo consume 65 dólares en promedio. Entonces llegaremos a la conclusión que el valor esperado de *Y* cuando el valor de *X* es 80, es igual a 65 dólares. En términos matemáticos estaría expresado como $E(Y|X=80)=65$. **Recuerde que media, promedio, valor esperado o esperanza** pueden ser empleados como sinónimos, todos estos términos indican el valor que se espera obtener de *Y* dado un valor de *X*. Si calculamos los 9 valores fijos restantes y los alineamos en una gráfica obtenemos la siguiente línea de regresión poblacional:



Gráfica 3.2. Distribución condicional del gasto en los niveles de ingreso.
Elaboración (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

En la gráfica 3.2. Se observa que en el eje horizontal se muestran los 10 valores fijos de *X* y en cada punto que conforma la línea están los valores esperados condicionales y alrededor de esos valores se encuentran los valores observados de *Y*, por ejemplo, en el primer grupo donde el valor esperado de *Y* dado $X=80$ es 65 alrededor del valor esperado se encuentran dispersos el consumo semanal de las familias. Es decir tenemos el valor

esperado y una dispersión alrededor de él. Si recordamos la teoría de distribución de probabilidad, esta definición hace suponer que estamos ante una distribución, donde el valor esperado está en el centro de la curva de distribución y dentro de la curva se encuentra los demás valores. Por ello, en el siguiente gráfico podemos observar el valor esperado de *Y* dado *X* cuando *X* vale 80, 140 y 220.



Gráfica 3.3. Línea de regresión población del ejemplo.
Elaboración (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

Una vez entendido estas aseveraciones, se puede entender la siguiente cita:

> *"Así, desde el punto de vista geométrico, una curva de regresión poblacional es tan solo el lugar geométrico de las medias condicionales de la variable dependiente para los valores fijos de la variable explicativa"* (Gujarati & Porter, 2010)

Ahora cobra más sentido la ecuación $E(Y|X_i) = f(X_i)$ pues hemos concluido que el valor promedio de *Y* varía con cada valor de *X*, sin embargo queda responder a la pregunta: ¿Cuál debería ser la función correcta que adopta *f(X)*?. (Gujarati & Porter, 2010) Explican que esta pregunta empírica tiene una solución en la que cada economista podría darle, es decir, depende el investigador que función utilizar. La más usada es sin lugar a

dudas la forma funcional lineal. Por lo tanto podemos asumir la siguiente expresión matemática:

$$E(Y|X_i) = \beta_0 + \beta_1 X_i \text{ (3.1.2.)}$$

La anterior expresión matemática es la primera aproximación, donde los parámetros β_0 y β_1 se les conoce como el **coeficiente de intercepción** y la **pendiente** respectivamente, estos términos son propios del concepto de linealidad. (Wooldrige, 2009) Detalla que al ser una función lineal, el aumento en una unidad de *X* el valor esperado de *Y* se modifica en la cantidad β_1. Es entonces el objetivo de la econometría, cuantificar el valor de los parámetros poblaciones mediante la estimación de la información muestral representado por los estimadores muestrales. (Gujarati & Porter, 2010) Expande el concepto de linealidad y denota que el concepto de linealidad tiene dos enfoques: **linealidad en las variables** y **linealidad en los parámetros**.

Sin embargo, aún hace falta expresar las desviaciones o dispersiones de la variable *Y* con respecto a su valor promedio. Por ello es que a la ecuación (3.1.2.) agregamos una parte no sistemática, matemáticamente se expresa de la siguiente manera:

$$\mu = Y_i - E(Y|X_i) \text{ (3.1.3.)}$$

O

$$Y_i = E(Y|X_i) + \mu_i \text{ (3.1.4.)}$$

Según (Gujarati & Porter, 2010), la ecuación 3.1.4 Indica que el consumo de las familias depende de una parte sistemática o determinada, compuesta por $E(Y|X_i)$ el cual señala que el consumo de las familias depende del ingreso semanal que es la media del consumo de las familias de un mismo grupo y además el término μ_i es una variable que representa a todas las variables que no están especificadas en el modelo pero que de igual manera tienen influencia sobre la variable dependiente, y se muestra en las desviaciones con respecto a su valor medio de la variable dependiente. Por ello es que si la ecuación (3.1.4.) Sigue una función lineal entonces se expresa en la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i \text{ (3.1.5.)}$$

La variable μ_i es conocida como **término de perturbación y es una variable aleatoria que sigue ciertos supuestos** necesarios que se tomarán en cuenta para la estimación de los estimadores muestrales que representen a los parámetros poblacionales.

La definición, demostración y las implicaciones de esos supuestos se presentarán más adelante con el fin de no generar cansancio para el lector.

Podemos concluir que la función de regresión poblacional se estima mediante la función muestral de regresión y para lograr cierta estimación se tiene que tener en cuenta supuestos y una fuerte base en la estadística inferencial. Por ello, se muestra a continuación la función de regresión muestral.

3.1.2. Función de regresión muestral.

Al igual que un trabajo descriptivo, cuando se requiere construir una base de datos que represente a la población, es imposible tomar en cuenta todos los datos de la población, por ello es que en el modelo clásico de regresión lineal al tener la limitante de no poder estimar los parámetros poblacionales entonces se estima estimadores muestrales, que como su propio nombre indica estos estimadores tendrán su base en una muestra que representa a la población. La siguiente cita lo explica:

> *"La FRP es un concepto idealizado, pues en la práctica pocas veces se tiene acceso al total de la población de interés. Por lo general se cuenta solo con una muestra de observaciones de la población. En consecuencia, se utiliza la función de regresión muestral estocástica (FRM) para estimar FRP."* (Gujarati & Porter, 2010)

La función de regresión muestral se expresa en la siguiente ecuación:

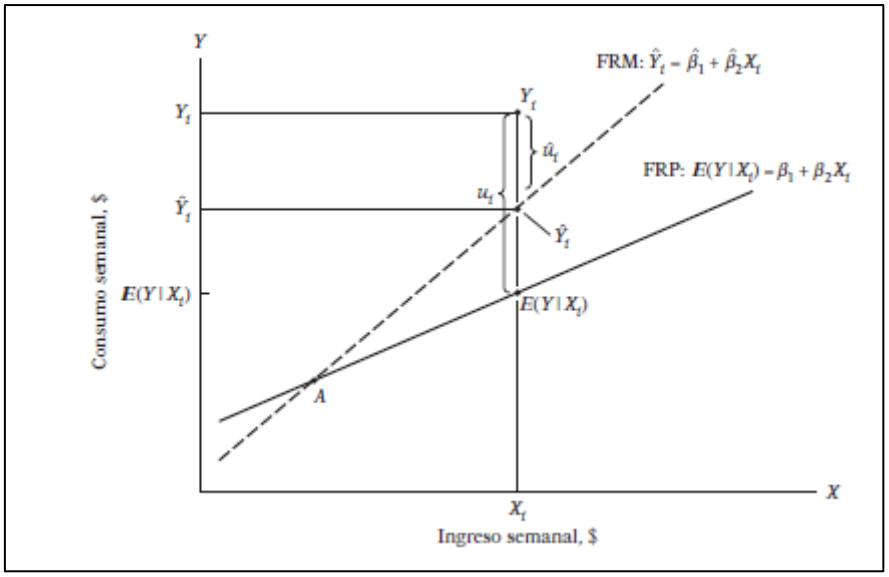
$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} X_i \quad (3.1.6.)$$

La principal diferencia entre la ecuación (3.1.5.) y la ecuación (3.1.6.) Es la presencia de un "*gorrito"* sobre los coeficientes de regresión, donde $\widehat{Y_i}$ se lee como "Y sombrero" y **estima** el valor esperado *Y* dado cada valor de *X*, es decir estima $E(Y|X_i)$, en otras palabras **es el valor estimado o ajustado de Y**, mientras $\widehat{\beta_1}$ y $\widehat{\beta_2}$ son los estimadores de β_1 y β_2. Sin embargo, a la ecuación 3.1.6. Falta el componente no sistemático, es decir el componente estocástico, por lo tanto, ajustemos su forma estocástica.

$$Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + \widehat{\mu_i} \quad (3.1.7.)$$

Donde $\widehat{\mu_i}$ es el estimador de μ_i, conocido como **término residual** o simplemente residuo. Es bien sabido que el uso de una muestra supone que represente a la población,

por lo tanto la línea de regresión muestral debe ser ajustada, de tal manera que sea igual o lo más parecido posible con la línea de regresión poblacional. (Gujarati & Porter, 2010) Explica esta definición en la siguiente gráfica:



Gráfica 3.4. Línea de regresión población y muestral.
Elaboración (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

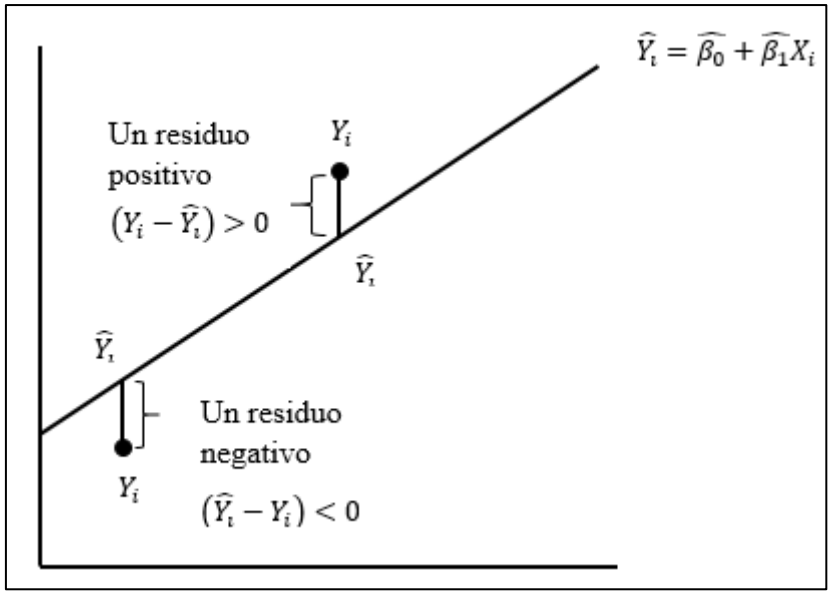
El gráfico 3.4. Supone que los residuos representados con $\widehat{\mu_i}$ es la diferencia entre el valor observado de *Y* con el valor ajustado o estimado de *Y*, entonces el valor de *Y* observado es la suma del valor estimado de *Y* más los residuos, expresado en forma ecuacional:

$$Y_i = \widehat{Y_i} + \widehat{\mu_i} \text{ (3.1.8.)}$$

Si despejamos el término residual de la ecuación (3.1.8.) obtendremos

$$\widehat{\mu_i} = Y_i - \widehat{Y_i} \text{ (3.1.9)}$$

(L. Webster, 2005) Explica lo que implica la diferencia entre el valor observado de *Y* y el valor estimado de *Y*. Cada punto que conforma la función de regresión muestral (FRM) representa cada valor estimado de *Y*, además que depende del valor observado *Y* si se determina si es positivo o negativo el residuo. En la gráfica 3.4. Se aprecia una sobrestimación, es decir debido a que el valor observado *Y* es mayor a valor estimado es que el residuo es positivo, pero si fuese al revés, es decir si el valor estimado de Y es mayor al valor observado de Y, entonces sería una subestimación. La siguiente gráfica representa lo explicado anteriormente:



Gráfica 3.5. Línea de regresión muestral.
Elaboración propia
Fuente: (L. Webster, 2005)

En la gráfica 3.5. Podemos observar de forma más clara lo expuesto en el párrafo anterior, sólo falta precisar:¿A qué se refiere con valor observado?, el valor observado hace referencia a los valores que conforman la muestra representativa, debido a que se tiene que trabajar con una muestra ya que es imposible obtener la información de toda la población.

Podemos concluir entonces que la función de regresión muestral estima a la función de regresión poblacional y además se usa con la ayuda de una muestra que represente a la población. Al ser la FRM una estimación de la FRP es que los coeficientes de regresión de la FRM son los estimadores muestrales con los que lograremos acercarnos a los parámetros poblacionales, en varios textos de econometría se suele encontrar conceptos que indican que nunca se podrá conocer el verdadero parámetro poblacional, esta afirmación requiere entonces que se asume que los estimadores muestrales son iguales a los parámetros poblacionales, es decir que los coeficientes de la FRM cumplen con la condición de estimador insesgado, una condición muy importante ya que se busca que los modelos econométricos cuantifiquen estimadores insesgados. Al usar una muestra que estime a la población ocasiona que el valor observado o muestreado de la variable dependiente sea distinto de su valor estimado, por lo tanto se generan residuos que no es más que la diferencia entre lo que se observa con lo que se estima y dependiendo del lugar del valor de la variable dependiente con respecto a su valor ajustado concluimos si cada

residuo tiene signo positivo o negativo. Tomando en cuenta a (L. Webster, 2005) quien explica que debido a que algunos residuos serán positivos y negativos es que la **Suma de los errores** o **Suma residual** sea igual a 0. Asumir que la Suma residual es igual a 0 significa que los errores pueden ser omitidos a pesar que sean incluidos en la especificación del modelo. (Orellana, 2008) Argumenta que al tener una población estimada a partir de una muestra aleatoria, podemos usar varias muestras para estimar la población por ello es que se pueden obtener varias funciones de regresión muestrales con diferentes estimadores muestrales, además señala que una forma de elegir un modelo sobre otro es darse cuenta cual es el modelo que minimiza las distancias de los residuos. El método de estimación que hace referencia es el método de estimación mediante **MÍNIMOS CUADRADOS ORDINARIOS**, sin embargo, abordaremos más adelante este tema con mayor detalle.

3.2. Análisis de Regresión Múltiple.

Hasta este momento se ha utilizado solamente una variable explicativa para determinar cuánto influye sobre la variable explicada, sin embargo, el comportamiento de una variable puede ser afectada por el comportamiento de múltiples variables, por lo que es necesario realizar un análisis de regresión múltiple, una vez más se replica el modelo de regresión múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu \quad (3.3.)$$

Donde existen *k* variables explicativas para explicar a la variable endógena *Y*, esta es la forma más general de representar una regresión lineal múltiple. A continuación, se explica este tema. Para empezar, al igual que el modelo clásico de regresión simple, cada β es el coeficiente de regresión, donde β_0 es el término independiente debido a que no le acompaña ninguna variable explicativa, el resto de coeficientes desde β_1 hasta β_k son las pendientes del modelo para cada variable explicativa. (Novales, 1998) Explica que las pendientes del modelo de regresión múltiple también se le pueden conocer como **coeficientes de regresión parcial**, debido a que cada uno de los coeficientes podrá medir el efecto que tiene cada variable explicativa sobre la variable explicada cuando sean estimados y manteniendo el principio de ceteris paribus, es decir suponiendo que las demás permanecen inalterables se lograra medir el efecto de cada variable explicativa sobre la explicada.

Sin embargo, el supuesto de ceteris paribus es difícil de mantener en la realidad, debido a que algunas variables explicativas podrían tener cierto grado de relación entre ellas, por lo tanto al introducir más variables en el modelo de regresión lineal podríamos romper con el supuesto de independencia entre las variables explicativas por lo tanto caeríamos en el problema de **multicolinealidad**, un problema que será detallado más adelante pero una definición breve sobre el problema de multicolinealidad se presenta en la siguiente cita textual.

> *"La relación lineal entre dos o más variables independientes se llama multicolinealidad."* (Hanke & Wichern, 2006)

Al igual que el modelo de regresión simple, el modelo de regresión múltiple también mide el valor medio de la variable dependiente. (Gujarati & Porter, 2010) Detallan que los coeficientes de regresión parcial miden el efecto directo de cada variable explicativa sobre la variable dependiente, así β_1 mide el efecto directo que tiene la variable explicativa X_1 sobre el valor medio de la variable dependiente $E(Y)$ manteniendo el supuesto de ceteris paribus en las demás variables explicativas. Se puede expresar mediante una forma ecuacional, siendo:

$$E(Y|X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (3.4.)$$

Al igual que en el modelo de regresión simple también posee la función de regresión poblacional expresado en la ecuación (3.3.) y la función de regresión muestral, con los sombreros encima de los parámetros y también se debe hacer uso de la estadística inferencial para lograr estimar los estimadores muestrales que permitan acercarse al parámetro poblacional. También cuenta con un término de perturbación que sigue una distribución normal con media 0 y varianza constante.

3.2.1. Matriz de correlación

Entender la correlación es importante para la construcción de modelos de regresión múltiple, debido a que las variables explicativas pueden estar altamente correlacionadas entre sí, es aconsejable medir la correlación existente para poder decidir cuáles variables o no incluir en el modelo especificado.

La correlación en términos simples quiere decir la dependencia que existe entre una variable con otra y se puede medir mediante el **coeficiente de correlación**. Y esta es la fórmula usada para lograr calcularla.

$$r = \frac{cov(X,Y)}{\sqrt{[var(X).var(Y)}} \text{ (3.2.1.)}$$

Donde *r* es el coeficiente de correlación y puede tomar valores comprendidos entre -1 y 1. Si disponemos de un conjunto de variables y queremos mostrar sus coeficientes de correlación, podemos usar una matriz de correlación. Los programas estadísticos pueden calcular las correlaciones de las variables y mostrarlas mediante una tabla conocida como matriz de correlación. (Hanke & Wichern, 2006) Muestra la siguiente tabla que es la matriz de correlación.

Variables	Variables		
	X_1	X_2	X_3
X_1	r_{11}	r_{12}	r_{13}
X_2	r_{21}	r_{22}	r_{23}
X_3	r_{31}	r_{32}	r_{33}

Tabla 3.2. Matriz de correlación.
Elaboración propia
Fuente: (Hanke & Wichern, 2006)

Donde cada r_{ij} son los elementos de la matriz que representan ser los coeficientes de correlación, para comprender un poco mejor. Se muestra el siguiente ejemplo.

Suponga que se quiere modelar un modelo econométrico que pretende explicar la **cantidad demandada** de cierto bien, con las variables **ingreso del consumidor** y el **precio del bien**. Pero previamente calculamos la correlación para observar el grado de relación que existe entre las variables. Y estos son los resultados.

Variables	Variables		
	Cantidad demandada	Ingresos	Precio
Cantidad demandada	1	0.86	0.65
Ingresos	0.86	1	0.54
Precio	0.65	0.54	1

Tabla 3.3. Ejemplo de matriz de correlación.
Elaboración propia

El ejemplo anterior sugiere que la **cantidad demandada** y los **ingresos** tienen una relación más fuerte que con el precio, por lo que se podría asumir que la variable **ingresos** podría ser más significativa al momento de explicar la cantidad demandada, la matriz de correlación ayuda al momento de determinar la existencia o no de multicolinealidad.

3.3. Supuestos del Modelo de Regresión Lineal de Mínimos Cuadrados Ordinarios

Ya sea para el modelo de regresión simple o el modelo de regresión múltiple, los modelos siguen supuestos para lograr la estimación mediante **MÍNIMOS CUADRADOS ORDINARIOS**.

Debido a que los supuestos en el modelo de regresión simple y múltiple tienen supuestos similares que deben cumplirse para ambos, es que al momento de representar la(s) variable(s) explicativa(s) en el modelo simple o múltiple se utilizara **X** para representarla(s) tanto en el modelo simple o múltiple.

Previamente ya se ha mostrado un cuadro que representa los supuestos del modelo clásico de regresión lineal con MCO. Se vuelve a mostrar el cuadro anterior, pues el siguiente apartado intentara explicar cuáles son los supuestos, el motivo por el cual son utilizados para estimar los coeficientes de regresión y cuáles son las consecuencias de no cumplirse los supuestos. A continuación, se muestra una réplica de la tabla sobre los supuestos del modelo de regresión lineal.

Supuestos o hipótesis del modelo de regresión lineal

Supuestos sobre la perturbación aleatoria	El término de error, μ, es una variable aleatoria con esperanza nula, una matriz de covarianzas constantes y diagonal. Y además $Cov(\mu_i, \mu_j)= 0$ cuando $i \neq j$ este es el supuesto de la **no autocorrelación** esto quiere decir que el término de error no tiene relación consigo misma debido a que es una variable aleatoria. Y al ser la varianza constante significa que no cambia y es independiente para cada valor del término de error, este es el supuesto de la **homocedasticidad**. El término de error, μ, es una variable aleatoria no observable, lo cual implica que la variable endógena sea aleatoria, ya que depende de una variable aleatoria, μ.

Supuestos sobre los regresores	El término de error es una variable aleatoria que sigue una distribución normal, es decir, que el valor esperado del término de error es 0, $E(\mu)=0$, y además tiene una varianza constante. Se le denota de la siguiente manera: $\mu \sim N(0, \sigma^2)$. Este es el supuesto de la **normalidad** de los errores. Las variables explicativas son linealmente independientes, es decir no existe relación lineal exacta entre ellas. Este es el supuesto de **independencia** y cuando no se cumple, el modelo presenta problema de **multicolinealidad**. Las variables explicativas son deterministas, es decir se pueden medir y no son inobservables. Sucede así porque su valor es constante y proviene de una muestra tomada en el tiempo y no tienen correlación con el término de error. Este supuesto se le conoce como la **exogeneidad**. Las variables no tienen error de medida y además el número de observaciones, *n*, debe ser igual o mayor al número de regresores, *k*.
Supuestos sobre los parámetros	Los parámetros son fijos y además cumplen sus propiedades anteriormente explicadas. Este supuesto quiere decir que los parámetros tienen estabilidad en el tiempo de las estimaciones, de este supuesto surge la teoría de la cointegración. Una teoría muy usada en la estimación de series temporales.
Supuestos sobre la forma funcional	La relación entre la variable dependiente y las variables independientes es lineal. Este es el supuesto de la **linealidad**. Se asume que el modelo especificado tiene ausencia de error de especificación, significa que se han incluido solamente las variables independientes relevantes para la explicación de la variable dependiente.

Tabla 3.4. Supuestos del modelo de regresión lineal.
Elaboración propia
Fuente: (Pérez L., 2012)

La tabla 3.4. Brinda un resumen de los supuestos del modelo clásico de regresión lineal explicados por la teoría que presenta (Pérez L., 2012) El cual clasifica los supuestos en cuatro grupos acorde a una parte del modelo econométrico, el primero de ellos: supuestos sobre la perturbación aleatoria.

3.3.1. Supuestos sobre la perturbación aleatoria.

3.3.1.1. *La normalidad de los residuos.*

Los errores representados en el modelo econométrico por el término de error tienen una distribución normal, este es el primer supuesto: **la normalidad** de los errores, el cual se representa matemáticamente como:

$$\mu \sim N(0, \sigma^2) \quad (3.3.1.)$$

La expresión (3.3.1.) se lee: "**el término de perturbación sigue una distribución normal con media 0 y varianza constante.**" Para (Cid S., Mora C., & Valenzuela H., 1990) La expresión anterior indica que el término de error es una variable aleatoria con distribución de probabilidad normal, además su transcendencia como supuesto del modelo clásico de regresión se centra en que el cumplimiento de la normalidad del término error garantiza que los estimadores cumplan la condición de **Mejores Estimadores Lineales Insesgados**, es decir permite obtener estimadores MELI. Esto será explicado cuando se hablen sobre los supuestos de los estimadores, pero este es el punto de partida.

Para (Cid S., Mora C., & Valenzuela H., 1990) El hecho que tenga una media 0, como consecuencia de la distribución normal, hace suponer que la esperanza o el valor esperado de la variable aleatoria μ es igual a 0. Matemáticamente se representa como.

$$E(\mu) = 0 \quad (3.3.2.)$$

Pero ¿Qué implica que la esperanza sea nula? (Gujarati & Porter, 2010) Explica este supuesto.

> *"(…) los factores no incluidos explícitamente en el modelo y, por consiguiente, incorporados en μ_i, no afectan sistemáticamente el valor de la media de Y; es decir, los valores positivos μ_i se cancelan con los valores negativos de μ_i, de manera que el efecto medio o promedio sobre Y es cero."* (Gujarati & Porter, 2010)

La cita anterior explica que los valores no incluidos en el modelo econométrico no afectan el valor de la media de Y, es decir los errores representados por el término de error no explican ni tampoco afectan a la variable dependiente debido a que los valores positivos se cancelan con los valores negativos. (Wooldrige, 2009)Explica que cuando el

intercepto aparece en la ecuación se supone que la media del término de error es cero, esta implicancia también es explicada por la expresión (3.3.2.).

Este supuesto ocurre para todo valor observado de X, por lo tanto la expresión (3.3.12.) se amplía en:

$E(\mu|x) = E(\mu) = 0$ (3.3.3.)

(Wooldrige, 2009)Explica que si el término de error y la(s) variable(s) explicativa(s) no están **correlacionadas** y además son aleatorias entonces no están relacionadas linealmente. Sin embargo, advierte al mismo tiempo que es posible que el término de error podría estar correlacionada con alguna función de la(s) variable(s) explicativa(s), por ejemplo con X^2. La expresión (3.3.3.) significa entonces que dado cualquier valor de X, la media del término de perturbación, es decir de μ, es igual a 0, si este supuesto se cumple entonces ambas variables no dependen entre sí. Recuerde que al emplear la expresión "término de error" o "término de perturbación" indica que estos supuestos se deben cumplir para la Función de Regresión Poblacional, por lo tanto al estimar la Función de Regresión Muestral también debe cumplir los supuestos.

Al suponer que la media condicional es 0, es decir que $E(\mu|X)$=0, entonces podemos suponer que:

$E(Y_i|X_i) = \beta_1 + \beta_2 X_i$ (3.3.4.)

(Wooldrige, 2009)Explica que la expresión (3.3.4.) es la FRP, donde la linealidad muestra que el aumento en una unidad del valor de X hace que el valor esperado de Y aumente en β_2. (Gujarati & Porter, 2010) Señalan que el supuesto de la normalidad del término de error implica que el modelo no tiene sesgo de especificación. Aunque no es la temática de esta guía de estudios, se alcanza una breve descripción de lo que es un **sesgo de especificación**, (Galán F., y otros, 2016) Alcanzan una definición de lo que es un sesgo de especificación también llamado error de especificación.

> *"La especificación incorrecta del modelo puede deberse a una formulación no adecuada de la forma funcional o bien, a que se violan los supuestos del error aleatorio o incluso a la información empírica que se incorpora al modelo para su estimación."* (Galán F., y otros, 2016)

Es importante evitar los sesgos de especificación ya que podrían ocasionar que los estimadores muestrales no sean los idóneos para estimar a los parámetros poblacionales, lo que nos puede conducir a errores en los resultados y en la inferencia sobre la población.

(Hernández A. & Zúñiga R., 2013) Identifican los principales errores de especificación.

- Variable irrelevante
- Variable omitida
- Error en la función

3.3.1.2. Homocedasticidad.

El supuesto de homocedasticidad parte de la distribución normal del término de error. Al tomar en cuenta que el término de error sigue una distribución normal se asume que la varianza del término de error es constante. La siguiente expresión representa el supuesto de homocedasticidad.

$$var(\mu_i) = \sigma^2 \text{ (3.3.5.)}$$

Podría surgir la pregunta ¿Exactamente, que significa que el término de error sea constante? (Gujarati & Porter, 2010) Formulan el supuesto de homocedasticidad extiendo la expresión (3.3.5.)

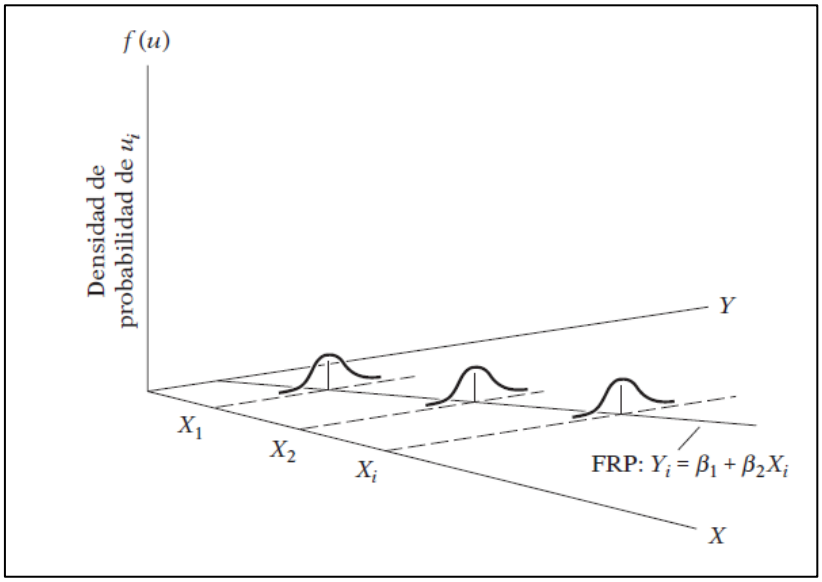
$$var(\mu_i|X_i) = \sigma^2 \text{ (3.3.6)}$$

La ecuación (3.3.6.) es una expresión más detallada del supuesto de **homocedasticidad**, (Gujarati & Porter, 2010) Mencionan que la varianza del término de error μ_i para cada valor de la(s) variable(s) explicativa(s) es constante. Explicación: anteriormente se expuso que el término de error puede ser definido como la diferencia que existe entre los valores poblacionales de *Y* con respecto a su media. Matemáticamente se expresa de la siguiente forma:

$$\mu_i = Y_i - E(Y|X_i) \text{ (3.1.3.)}$$

Entonces, dado cada valor de *X* sobre *Y*, genera una media condicional expresada con *E(Y|X)* de forma muy general, por definición de la teoría de probabilidades, alrededor de la media están dispersos los valores poblacionales de *Y* para cada valor *X*. Entonces el supuesto de homocedasticidad manifiesta que la varianza del término de error será igual

para cada valor de *X* que explica *Y*. La idea anterior puede verse resumida en la siguiente gráfica que han sido tomada de (Gujarati & Porter, 2010).



Gráfica 3.6. Varianza constante.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

La gráfica 3.6. Resume lo anteriormente enunciado, la línea que representa a la ecuación $Y_i = \beta_1 + \beta_2 X_i$ es la función de regresión poblacional (FRP) donde cada valor de la línea es la esperanza condicional, también llamada media condicional de *Y* dado cada valor de *X* representado con *E(Y|X)* y alrededor de cada media condicional están dispersos los valores poblacionales de *Y* dado cada valor de *X*, a esa dispersión se le conoce como **término de error** representado con μ. Al ser las curvas de distribución iguales para cada valor de *X* concluimos que la varianza del término de error no varía, por lo tanto, la varianza del término de error dado cada valor de *X* es constante.

(Wooldrige, 2009) Expande esta concepción del supuesto de homocedasticidad. La independencia que existe entre el término de error y la(s) variable(s) explicativa(s) es la causante por la que se supone que la varianza del termino de error dado la(s) variable(s) explicativa(s) sea constante. Para que quede más claro, veamos sus expresiones matemáticas:

$$var(\mu_i) = E[\mu_i - E(\mu_i)]^2 \quad (3.3.7.)$$

Si recordamos que $E(\mu_i)$=0 obtenemos lo siguiente:

$$var(\mu_i) = E\{\mu_i^2 - 2\mu_i E(\mu_i) + [E(\mu_i)^2]\}$$

$var(\mu_i) = E(\mu_i^2) = \sigma^2$(3.3.8.)

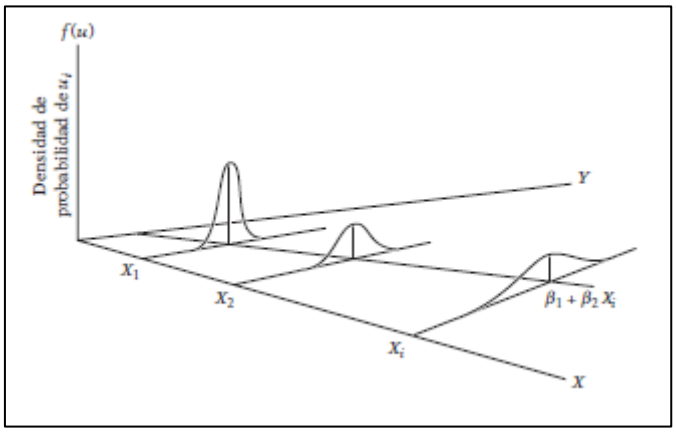
La varianza σ^2 que aparece en (3.3.8.) es la **varianza incondicional de μ_i**, σ^2 también es conocido como varianza del término de terror o varianza del término de perturbación según la teoría explicada por (Wooldrige, 2009). Y a la raíz cuadrática de la varianza del error se le conoce como **error estándar de la regresión**. De esta manera, al asumir que la varianza del término de error es constante también debe quedar claro que la varianza condicional de la variable dependiente dado la(s) variable(s) explicativa(s) es constante (Gujarati & Porter, 2010) Lo representan matemáticamente como:

$var(Y_i|X_i) = \sigma^2$ (3.3.9)

Debe quedar claro que el valor esperado de Y dado X, $E(Y|X)$, es lineal y la varianza de Y dado X $var(Y|X)$ es constante. (Núñez Z., 2007) Expresa que este supuesto garantiza que las distribuciones de probabilidades del término de error son iguales, sin embargo, mantener este supuesto es difícil y es más que probable que al momento de construir modelos econométricos nos encontremos con un problema con respecto a una varianza que no es constante, es decir, la presencia de **heterocedasticidad** en el modelo significa que el modelo especificado no está cumpliendo con este supuesto. (L. Webster, 2005) Menciona que la presencia de heterocedasticidad en el modelo indica que las varianzas del término de error para cada valor de X son diferentes, entonces se asume que los valores Y se dispersan ampliamente a medida que incrementan los valores de X. (Gujarati & Porter, 2010) Expresan lo anterior en la siguiente ecuación:

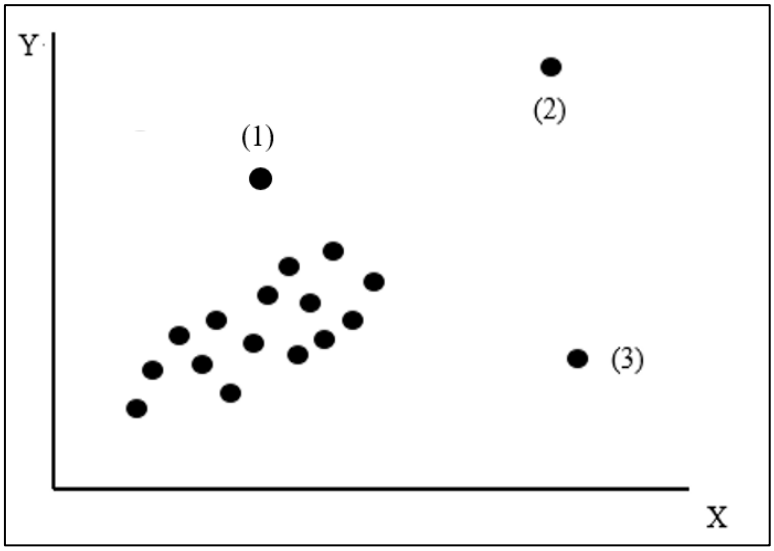
$var(\mu_i|X_i) = \sigma_i^2$(3.3.10.)

La ecuación (3.3.10.) se puede representar gráficamente. (Gujarati & Porter, 2010) Detallan la gráfica.



Gráfica 3.7. Varianza no constante.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

La detección, tratamiento y explicación de sus causas serán resumidas más adelante. Pero una causa muy frecuente del problema de heterocedasticidad son los **datos atípicos**. A través del siguiente gráfico elaborado a partir del trabajo de (Cid S., Mora C., & Valenzuela H., 1990) se define el concepto de dato atípico.



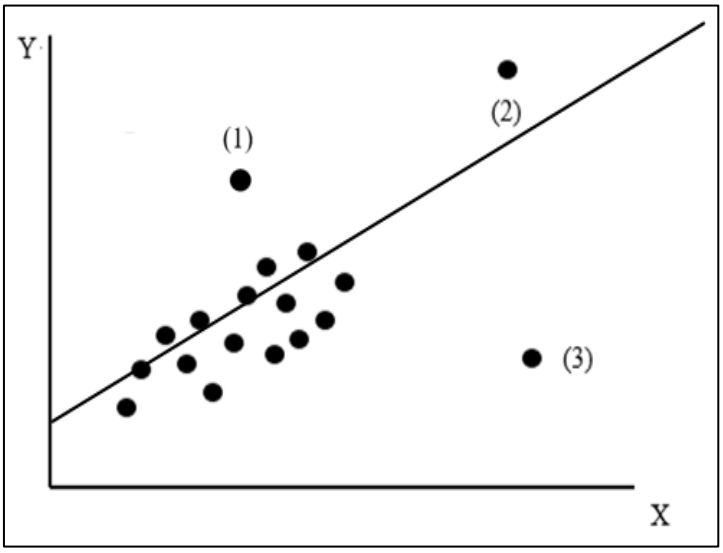
Gráfica 3.8. Datos atípicos (outliers).
Elaboración propia
Fuente: (Cid S., Mora C., & Valenzuela H., 1990)

El gráfico 3.8. Muestra una idea sobre el concepto de dato atípico. Para (Cid S., Mora C., & Valenzuela H., 1990) La inclusión de datos atípicos producidos por un error de registro o muestreo puede inferir a conclusiones erróneas, debido a que un dato atípico es un tipo de dato que se aleja del resto de las observaciones. En muestreo, el dato atípico pertenece a una población muy distinta de la que está incluido en la mayoría de casos. En el gráfico 3.8. Los tres puntos son ejemplos de datos atípicos, y la inclusión de estos datos en la base de datos extraída podría llevarnos a un problema de heterocedasticidad. **Esta es la causa más común de heterocedasticidad, prevenida solamente por el correcto registro de las observaciones**. Para (Cid S., Mora C., & Valenzuela H., 1990) los tres puntos en la gráfica 3.8. Son tres distintos tipos de puntos atípicos. Explica textualmente:

> *"El punto (1) es atípico con respecto del comportamiento global, pero, no es anómalo ni respecto de las variables X ni de la variable dependiente Y. (...). El punto (2), es atípico respecto a ambas distribuciones, pero su ubicación en el plano hace que este no afecte el resultado de la función de regresión resultante. El punto (3) sin embargo, es también atípico respecto de ambas distribuciones y*

su posición afectará significativamente los parámetros de la recta de regresión." (Cid S., Mora C., & Valenzuela H., 1990)

La cita textual anterior menciona que los dos puntos (1) y (2) de la gráfica 3.8. No afectan directamente a pesar de ser atípicos, esta definición no es una contradicción hacia la teoría econométrica, está claro que la inclusión de datos atípicos afecta al modelo que se quiere especificar, pero debido a su ubicación cercana a la línea de regresión es que podría darse el caso de no afectar directamente la varianza del término de error. Siguiendo la lógica anterior podemos caer en cuenta porque el punto (3) si afecta directamente la varianza del término de error, debido no solo a la distancia alejada que tiene con respecto a los demás datos sino también con respecto a la línea de regresión. La inclusión de la línea de regresión en el gráfico puede ayudar a concebir la idea expuesta en la cita textual.



Gráfica 3.9. Datos atípicos (outliers) con la línea de regresión.
Elaboración propia
Fuente: (Cid S., Mora C., & Valenzuela H., 1990)

Cabe mencionar, que la gráfica 3.9. Resume lo anterior dicho. Al momento de construir un modelo econométrico es recomendable prevenir los datos atípicos debido a que afectan a la varianza del término de perturbación y por ende al error estándar de la regresión del modelo, sin importar la ubicación con respecto a la línea de regresión los datos atípicos afectarán al modelo, pero según la explicación de la cita anterior, puntos (1) y (2) no afectarán directamente, sin embargo el punto (3) si afecta directamente ocasionando resultados equivocados.

3.3.1.3. No autocorrelación.

Si tomamos en cuenta que el término de perturbación es una variable aleatoria entonces estamos aceptando que sus valores tienen que haber sido generados mediante un proceso aleatorio obteniendo una muestra aleatoria, si consideramos la aleatoriedad presente en el término de error concluimos que sus elementos no deben depender entre ellos. Es decir, el supuesto de la ausencia de autocorrelación manifiesta que existe independencia en los valores del término de perturbación. Lo anterior se puede describir de la siguiente forma ecuacional:

$$cov(\mu_i, \mu_i) = 0 \text{ (3.3.11)}$$

La expresión (3.3.11.) es la forma matemática de expresar el supuesto de no autocorrelación, donde *cov* significa covarianza, (Pérez-Tejada, 2007) Brinda el concepto de covarianza, la covarianza es una medida que describe la forma en que dos variables se relacionan, siendo más específicos, como ambas varían. Además establece que para poder calcular la covarianza previamente debe realizarse una relación lineal entre ellas.

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \text{ (3.3.12.)}$$

Al asumir la ausencia de autocorrelación entre los valores del término de error, también se concluye la ausencia de correlación entre el término de error y la(s) variable(s) explicativa(s). Lo anterior se expresa como:

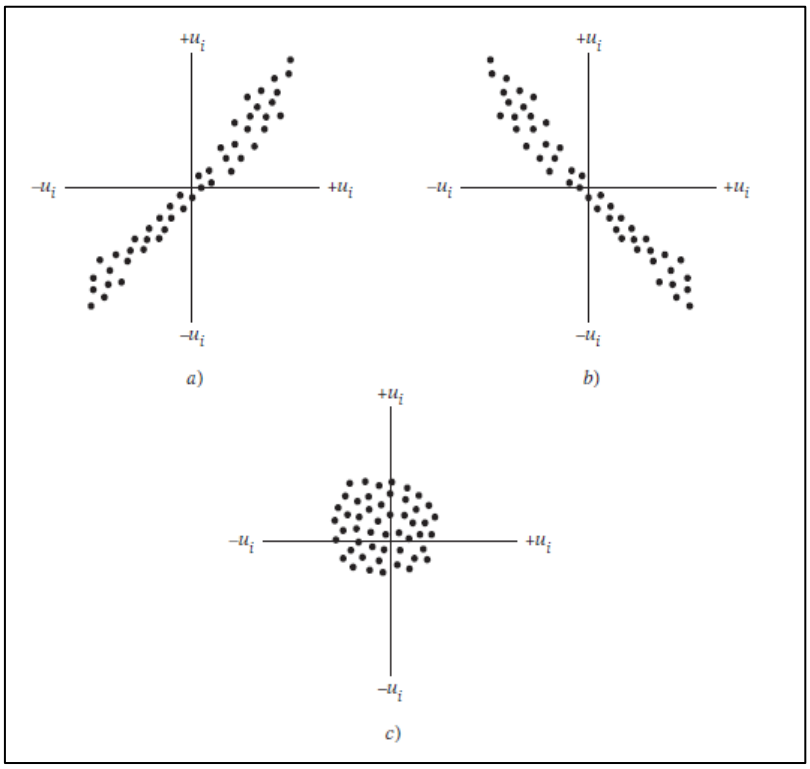
$$cov\left(\mu_i, \mu_j \middle| X_i X_j\right) = 0 \text{ (3.3.13.)}$$

La expresión (3.3.13.) significa la ausencia de correlación entre los términos de error dado los valores de *X* si las observaciones *i* y *j* son distintas de 0 y entre sí, por lo tanto al asumir que en diferentes valores de los términos de error no existe una dependencia dada en cada valor de la(s) variable(s) explicativa(s) estamos asumiendo que el término de error es independiente de sus valores y para cada valor de *X*.

El supuesto de independencia entre los términos de error es más difícil de mantener en las series temporales que en los cortes transversales. (Hanke & Wichern, 2006) Explican, una serie de tiempo donde a medida que avanza en el número de observaciones se estará avanzando también en las fechas que registran los datos observados (muestreados), no puede considerarse inicialmente como una muestra aleatoria, debido a que los valores actuales de una serie de tiempo en la gran mayoría de

casos, por no decir que se cumple como una regla general, depende fuertemente de los valores pasados de la serie temporales, provocando que el modelo econométrico estimado con datos de series temporales puedan tener un problema de autocorrelación. La autocorrelación está presente con frecuencia en modelos con datos de series temporales mientras que la heterocedasticidad es un problema frecuente con datos de corte transversal, sin embargo esto no quita la probabilidad de encontrar heterocedasticidad en una serie temporal ni autocorrelación en datos de corte transversal.

Al caer en cuenta que existe autocorrelación en el modelo, podemos notarlo en un gráfico que representa a los valores del término de error siguiendo un patrón o tendencia. (Gujarati & Porter, 2010) Muestran lo anterior en el siguiente gráfico.



Gráfica 3.10. Autocorrelación.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

El gráfico anterior que ha sido tomado de (Gujarati & Porter, 2010), muestra los tres posibles casos con respecto a la ausencia o no de autocorrelación, en los gráficos (a) y (b), los patrones de los valores del termino de error son positivos y negativos, por lo que al seguir un patrón bien definido podemos intuir que el modelo viola el supuesto de no autocorrelación, además dependiendo de la forma del patrón podemos decir que el gráfico (a) tiene autocorrelación positiva mientras que el gráfico (b) tiene autocorrelación negativa. Lo ideal es que el modelo tenga un gráfico parecido al grafico (3).

Finalmente es necesario señalar la diferencia entre el término "**autocorrelación**" y "**correlación serial**", (Gujarati & Porter, 2010) Explican que se le define como

autocorrelación a la dependencia que existe entre los elementos de una variables, mientras que correlación serial es la correlación existente entre dos variables. En algunos textos se utiliza a ambos términos como sinónimos.

Antes de pasar al siguiente punto, podemos adelantarnos que los residuos representados en los gráficos de dispersión pueden dar una idea si se viola o no algún supuesto sobre el término de perturbación. A esta forma de diagnosticar se le conoce como **método informal** y tiene como criterio observar y determinar la existencia o no de algún patrón de los residuos en el modelo estimado.

3.3.2. Violaciones a los supuestos sobre el término de perturbación.

Los siguientes cuadros ayudarán a entender que sucede cuando no se cumplen estos supuestos sobre el término de perturbación y cuáles son las causas de las violaciones de los supuestos.

Violación a los supuestos	Causas de la violación a los supuestos
No normalidad	• Existencia de datos atípicos. • Distribuciones no normales, ya sea porque no están centradas en la media o por una masa considerablemente grande en los extremos de la curva de probabilidades. • En las series de tiempo, si la(s) variables(s) tienden a incrementar o disminuir de forma no constante entonces su varianza es heterocedástica. En otras palabras, una causa de heterocedasticidad en las series temporales es su misma naturaleza de estar en crecimiento o decrecimiento.
Heterocedasticidad	• Incorrecta especificación del modelo ocasionada por la omisión de variables relevantes o agregando variables irrelevantes o una forma funcional incorrecta. • Existencia de datos atípicos. • La incorrecta transformación de los datos. • La mejora en la recolección de datos permite disminuir la variabilidad, por lo tanto, deja de ser constante.

	• En las series de tiempo, la autocorrelación por lo general es inherente a este tipo de datos, debido a que en las series de tiempo, las variables económicas tienen dependencia con sus valores pasados.
	• Al igual que la heterocedasticidad, las transformaciones en los datos y el sesgo de especificación podría ocasionar autocorrelación al momento de omitir variables relevantes o agregando variables irrelevantes, el sesgo de especificación por una forma funcional incorrecta también puede ser la causa de este problema. El sesgo de especificación por la forma funcional sucede por lo general cuando no se ejecuta un modelo lineal.
Autocorrelación	
	• En ocasiones la teoría económica hace relacionar una variable dependiente con el rezago de la variable independiente, un ejemplo de esto sería el **fenómeno de la telaraña**, que en palabras sencillas manifiesta que la cantidad ofertada es explicada por el precio en el periodo anterior. Una situación muy común en los mercados de libre competencia.
	• El mal manejo de los datos producto de una mala recopilación de datos también es una causa de autocorrelación.

Tabla 3.5. Causas de la violación a los supuestos del modelo de regresión lineal.
Elaboración propia
Fuente: (Pérez L., 2012) (Hanke & Wichern, 2006)

Cabe recalcar, que cuando elaboramos un modelo econométrico es demasiado probable cometer una violación a los supuestos, más aún cuando no se cuenta con la experiencia requerida, por ello al momento de estimar modelos econométricos mediante MCO y encontrar que los supuestos sobre el término de perturbación no se están cumpliendo podemos enfrentarnos a problemas tanto en los estimadores muestrales como también en la varianza del error del modelo estimado:

Supuesto no cumplido	Consecuencias
	• Los estimadores son insesgados, pero dejan de ser eficientes, este concepto se explica en las líneas siguientes.
No normalidad	
	• La varianza deja de ser insesgada por lo que existen problemas al momento de inferir sobre la población a partir de la muestra, en el siguiente cuadro se explica.

- En los estimadores del modelo conservan su insesgamiento sin embargo dejan de ser eficientes, por lo tanto, el estimador por MCO ya no tiene varianza mínima haciendo que los estimadores ya no sean MELI. Al perder la eficiencia de los estimadores ya no es posible estimar mediante MCO.

Heterocedasticidad

- Muy diferente a los estimadores, la **varianza del error** estimada del modelo se vuelve una varianza sesgada, esto quiere decir que la varianza del error estimada del modelo es diferente a la varianza poblacional, por lo tanto, ya no es posible hacer inferencias sobre la población desde la muestra debido a que sólo arrojaría conclusiones equivocadas. Este es el principal problema de la heterocedasticidad, ya que al ser la varianza del error sesgada generaría un **error estándar de la regresión** ineficiente por lo que el error estándar de la regresión estaría subestimado o sobreestimado, es decir el error estándar de la regresión estaría equivocado, derivado de ello, probar las hipótesis de **significancia individual** y **global** estarían erradas.

- Debido a que el error estándar de la regresión es ineficiente, el **coeficiente de determinación**, que mide cuánto explican la(s) variable(s) explicativa(s) a la endógena también estaría equivocado.

- Una vez más, debido al error estándar de la regresión estimado del modelo ineficiente, la **matriz de varianza y covarianza** de los estimadores mostraría valores incorrectos.

- Los pronósticos y predicciones que se quieran realizar a partir del modelo ajustado pueden estar equivocados.

Autocorrelación

En realidad, las consecuencias que implica tener un modelo con autocorrelación o correlación serial, son muy similares a las consecuencias de tener un modelo con heterocedasticidad.

- Los estimadores muestrales continúan siendo insesgados, pero pierden la condición de ser MELI debido a que no son eficientes.

- La varianza del error deja de ser insesgada por lo que el error estándar de regresión también es ineficiente y derivado de esto las pruebas de significancia global e individual pueden estar equivocadas.

- El coeficiente de determinación puede ser incorrecto.
- Y la matriz de varianza y covarianza de los estimadores puede realizar valores equivocados.

- Puede mostrar falsas predicciones y pronósticos.

- El problema de la autocorrelación es que puede mostrar una **relación espuria**, en términos simples, mostrar la existencia de alguna relación de dos variables cuando en realidad no existe. Aunque este problema también existe en la heterocedasticidad es más probable encontrarlo cuando existe autocorrelación en el modelo.

Tabla 3.6. Consecuencias de la violación a los supuestos del modelo de regresión lineal con estimación por MCO.
Elaboración propia
Fuente: (Pérez L., 2012) (Hanke & Wichern, 2006) (Novales, 1998)

La principal consecuencia de la presencia de heterocedasticidad, autocorrelación o no normalidad es que el método de estimación mediante MCO deja de ser el apropiado para estimar estimadores MELI. En los siguientes apartados se explicará el diagnóstico y el tratamiento a estos problemas debido a que se debe hondar en los fundamentos y principios del MCO.

3.3.3. Supuestos sobre sobre los regresores.

3.3.3.1. Independencia o no multicolinealidad.

Este supuesto, es exclusivo del modelo de regresión múltiple, y tal como ya se indicó anteriormente, las variables explicativas no tienen una relación lineal exacta entre ellas. Sin embargo, este supuesto en algunas ocasiones puede carecer de sentido incluirlo cuando queremos especificar un modelo econométrico, debido a que las variables económicas suelen estar relacionadas linealmente con otras variables, por ello para (Uriel & Aldás, 2005) La multicolinealidad, que es el nombre que se le otorga a la violación de este supuesto, puede ser exacta o aproximada, siendo **perfecta** o **imperfecta** respectivamente.

(Wooldrige, 2009) Detalla, si alguna variable explicativa es una combinación lineal perfecta o exacta de otras regresoras, entonces existe multicolinealidad exacta en el modelo estimado y de ser cierto sería imposible estimar los estimadores muestrales, por el contrario, cuando el modelo estimado tiene algún grado de relación lineal en sus regresoras, implica que el modelo estimado contenga multicolinealidad imperfecta o aproximada y aunque la multicolinealidad imperfecta no impide la estimación de los estimadores, si genera problemas en la estimación. En conclusión, el principal problema es cuando la combinación lineal es perfecta puesto que la multicolinealidad perfecta entre variables regresoras no permite la estimación.

(Gujarati & Porter, 2010) Detallan que si tenemos el siguiente modelo econométrico especificado: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu_i$ .Asumimos que las regresoras tienen multicolinealidad perfecta cuando la dependencia lineal se escribe como:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} = 0 \text{ (3.3.14.)}$$

Donde la condición de que $\lambda_1, \ldots, \lambda_k$ son constantes que simultáneamente no todas son iguales a 0, se está cumpliendo por lo que una de las regresoras tiene una dependencia lineal sobre las demás regresoras. Imagine que la constante $\lambda_2 \neq 0$ entonces acorde a (Gujarati & Porter, 2010) la ecuación (3.3.14.) se reescribe como:

$$X_2 = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - -\frac{\lambda_k}{\lambda_2} X_{ki} \text{ (3.3.15.)}$$

En (3.3.15.) se observa que **la regresora X_2 es una combinación lineal exacta de las demás variables regresoras**, de ser así entonces el modelo econométrico tiene multicolinealidad perfecta.

A diferencia de la multicolinealidad perfecta, la multicolinealidad imperfecta no es una combinación lineal exacta, sino aproximada la cual se plantea como:

$$\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \cdots + \lambda_k X_{ki} + v_i = 0 \text{ (3.3.16.)}$$

Donde v_i se le conoce como un error estocástico el cual, al no ser determinado admite que la dependencia es aproximada y por lo tanto la multicolinealidad no es perfecta. Si suponemos que la constante $\lambda_2 \neq 0$, entonces (3.3.16.) se transforma en:

$$X_2 = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - -\frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i \text{ (3.3.17.)}$$

Por lo tanto, se concluye que el modelo econométrico tiene multicolinealidad generada por la regresora X_2, sin embargo esta no es perfecta porque el error estocástico no permite conocer como realmente es el grado de correlación entre las regresoras.

El problema de la multicolinealidad es cuando existe una variable que es la combinación lineal exacta sobre las demás regresoras, es decir cuando la multicolinealidad es perfecta, porque esta genera una **influencia combinada**, un término que (Gujarati & Porter, 2010) Explican con el siguiente ejemplo, si tenemos el modelo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i$ donde se comprueba que $X_{2i} = 2X_{1i}$ la ecuación puede resultar como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 (2X_{1i}) + \mu_i$$

$$Y_i = \beta_0 + (\beta_1 + 2\beta_2) X_{1i} + \mu_i$$

$Y_i = \beta_0 + \alpha X_{1i} + \mu_i$ (3.3.18)

Donde $\alpha = (\beta_1 + 2\beta_2)$ es la influencia combinada y no existe forma de estimar por separado los parámetros β_1, β_2. Por ello, es que cuando la multicolinealidad perfecta está presente en un modelo econométrico no es posible obtener una estimación exacta de los estimadores.

> *"Si la multicolinealidad es perfecta (...), los coeficientes de regresión de las variables X son indeterminados, y sus errores estándar, infinitos. Si la multicolinealidad es menos que perfecta (...), los coeficientes de regresión, aunque determinados, poseen grandes errores estándar (...), lo cual significa que los coeficientes no pueden ser estimados con gran precisión o exactitud."* (Gujarati & Porter, 2010)

La cita anterior, indica que los estimadores son indeterminados producto de la influencia combinada que no permite capturar el efecto de una variable debido a su dependencia lineal es exacta. La estimación cuando la relación lineal no es exacta también genera problemas pues esta produce estimadores con poca precisión, tal como menciona la cita, es decir los **errores estándar** no son pequeños.

La presencia de multicolinealidad perfecta o imperfecta en el modelo hace que existan varios problemas además de la incorrecta estimación de los estimadores. A continuación se explican cuáles son las consecuencias de presentar multicolinealidad en el modelo:

- **No significancia de los estimadores.**

Esta es una consecuencia directa de la presencia de multicolinealidad, ya que al existir errores estándares tan grandes, los estimadores pueden ser no significativos, lo que quiere decir que se podría tener indicios de descartar una variable cuando en realidad debería estar presente en el modelo especificado. Posteriormente se explicará el concepto de significancia individual.

- **Estimación por intervalos incorrecta.**

Esta es otra consecuencia directa de la presencia de multicolinealidad en el modelo. Debido a que los errores estándares de los estimadores son más grandes de lo que deberían ser y estos son usados para la estimación por intervalos, entonces podrían hacer una estimación incorrecta del modelo especificado. Y al igual que en el anterior punto, el hecho de tener intervalos tan grandes puede hacer que una regresora aparentemente sea no significativa cuando en realidad podría serlo.

- **Un coeficiente de determinación demasiado alto.**

Esta consecuencia se puede usar también como un indicio sobre la existencia de multicolinealidad en el modelo. El coeficiente de determinación que mide en un porcentaje de cuánto explican las regresoras a la variable explicada puede ser muy alto, lo cual sería bueno, pero en presencia de multicolinealidad el coeficiente de determinación es demasiado alto y no solo eso sino que las variables explicativas no son significativas.

Es decir, aparentemente el modelo tiene una buena bondad de ajuste, sin embargo sus variables explicativas no explican individualmente a la variable explicada. Y como ya se dijo anteriormente, esto puede ser considerado como un indicio que existe multicolinealidad, tal como indican (Gujarati & Porter, 2010).

Una vez vistas las consecuencias de la multicolinealidad surge una pregunta: ¿Cuáles son las causas del problema? En realidad es difícil de determinar ya que no existe un consenso claro sobre el problema de multicolinealidad, sin embargo (De Grange C., 2005) Detalla algunas causas:

- La más obvia de todas, la existencia de una relación causal entre dos o más variables explicativas.
- La naturaleza de las variables económicas, esta es la causa más importante de todas, de hecho esta causa es la que origina a la primera causa, y es que para (De Grange C., 2005) Las variables económicas están correlacionadas entre ellas y se hace más evidente cuando se trabaja con datos de series temporales ya que basta que exista una tendencia creciente entre dos variables explicativas para que su correlación aumenta. Es por ello, que es casi seguro que la multicolinealidad estará presente en los modelos

econométricos. Esto es algo irónico, porque se podría pensar que quitando del modelo las variables explicativas que estén correlacionadas entonces la multicolinealidad desaparecerá, sin embargo, conforme a (Wooldrige, 2009) Esto no solo no puede ocurrir sino que además podríamos caer en un sesgo de especificación por omisión de variable relevante.

Posteriormente se explicará cómo detectar y solucionar el problema de multicolinealidad, pero desde ahora se tiene que entender que en ocasiones la solución de la multicolinealidad es no hacer nada al menos cuando la multicolinealidad que presenta el modelo no es perfecta.

3.3.3.2. Exogeneidad.

(Ahumada, 2014) Explica de manera sencilla a través de una cita de Wooldrige, que la exogeneidad es el supuesto que indica que las variables explicativas no están correlacionadas con el término de error. Cuando no se cumple este supuesto, el modelo tiene **endogeneidad**.

Debido a que los conceptos de exogeneidad, endogeneidad y causalidad son tan complejos y extensos, se debería redactar otra guía con el fin de explicar a profundidad este supuesto. Sin embargo, a manera de introducción podemos detallar un poco más sobre el supuesto de exogeneidad y los problemas de endogeneidad. Tal como (Bravo & Vásquez Javiera, 2008) Detallan que para estimar estimadores insesgados por MCO no debe existir correlación entre la(s) regresora(s) y el término de error. Sin embargo, al omitirse variables relevantes, o debido a la simultaneidad o al error de medición es que se incumple este supuesto.

La simultaneidad en econometría se debe a que las variables econométricas suelen estar demasiado relacionadas entre sí, tal como (Alonso, 2010) ejemplifica con el siguiente sistema de ecuaciones:

$$Y_1 = \alpha_1 Y_2 + \alpha_2 X_1 + \mu_1 \quad (3.3.19)$$

$$Y_2 = \alpha_3 Y_1 + \alpha_4 X_2 + \alpha_5 X_3 + \mu_1 \quad (3.3.20.)$$

Con las ecuaciones (3.3.19.) y (3.3.20.) se construye el sistema de ecuaciones, donde la variable Y_1 aparece tanto como variable explicada y explicativa, esta condición de aparecer tanto en la izquierda de una ecuación y a la derecha en otra se le conoce como **ecuaciones simultáneas**. Estas ecuaciones conforman un modelo econométrico

multiecuacional y son frecuentemente usados en los modelos macroeconométricos. El problema con la simultaneidad es que no siempre se cumple con la causalidad unidireccional, (Novales, 1998) Define que la causalidad unidireccional como el supuesto que indica la relación unidireccional entre las regresoras y la variable explicada, es decir la(s) variables(s) explicativa(s) ejercen una influencia sobre la variable explicada pero nunca al revés, continúa explicando que esta cuestión debe tratarse con cuidado y justificar cuáles son las relaciones causales entre variables con ayuda de la teoría, sin embargo en la simultaneidad la condición de causalidad unidireccional no se cumple o al menos no del todo. Para (Bravo & Vásquez Javiera, 2008) la causalidad en las ecuaciones simultáneas obliga a tomarse en cuenta en ambos sentidos. Algunos ejemplos pueden ser la ecuación de Mincer, que toma al nivel de escolaridad y al nivel de ingresos como variable explicada y explicativa, pero que a su vez se puede tomar a la escolaridad como una variable que explica al nivel de ingresos. Por lo tanto podemos encontrar una variable explicativa endógena siendo el nivel de escolaridad la variable que presenta endogeneidad.

En realidad, el tema es demasiado amplio pero la breve descripción anterior debería servir como ilustración para comprender mejor el supuesto de exogeneidad.

3.3.3.3. No existen errores de observación.

Al momento de la construcción de un modelo econométrico, debemos tomar en cuenta la no existencia de errores de especificación, en algunos textos, se pueden encontrar como sesgo de especificación, este es un supuesto que puede resultar difícil de cumplir más aún cuando no se cuenta con suficiente experiencia en la especificación de modelos econométricos. Al asumir la ausencia de errores en la especificación estamos asumiendo que los supuestos se cumplan debido a que los supuestos anteriores tienen una base que se apoya sobre la presunción de especificar un correcto modelo econométrico.

Es absolutamente complicado, por no decir que imposible, construir un modelo econométrico que no tengo un sesgo de especificación, (Gujarati & Porter, 2010) Asemejan la búsqueda de ese modelo econométrico *perfecto* a la búsqueda del Santo Grial y no es para menos, puesto que seleccionar las variables correctas, con los mecanismos correctos para el recojo de datos y revisar la teoría que mejor explique las relaciones entre variables es altamente improbable. (Gujarati & Porter, 2010) Muestran una clasificación sobre los tipos de sesgos de especificación.

- Errores de especificación del modelo.
  - Omisión de una variable relevante.
  - Inclusión de una variable innecesaria.
  - Adopción de una forma funcional incorrecta.
  - Errores de medición u observación.
- Errores de especificación incorrecta del modelo.
  - Especificación incorrecta del término de error estocástico.
  - Suposición de que el término de error esta normalmente distribuido.

Cuando consideramos que el modelo especificado es el verdadero pero no somos capaces de estimarlo debido a los errores entonces estamos ante el grupo de errores de especificación, por otro lado, cuando no tenemos ni idea cual es el verdadero modelo entonces estamos ante el segundo grupo conocido como errores de especificación incorrecta del modelo. Así lo han definido (Gujarati & Porter, 2010).

A continuación, se muestra un cuadro que pretende ser un resumen sobre las consecuencias de la omisión de una variable relevante y la inclusión de una variable irrelevante en un modelo econométrico. Pero previamente mostraremos ejemplos de ecuaciones con subajuste y sobreajuste tomados de (Gujarati & Porter, 2010).

Modelo subajustado

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$ (3.3.21.) (Modelo verdadero)

$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$ (3.3.22.) (Modelo subajustado)

Modelo sobreajustado

$Y_i = \beta_1 + \beta_2 X_{2i} + \mu_i$ (3.3.21.) (Modelo verdadero)

$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$ (3.3.22.) (Modelo sobreajustado)

Sesgo de especificación	Consecuencias
Subajuste (omisión de una variable relevante)	• Los estimadores por MCO son sesgados e inconsistentes cuando la regresora omitida está correlacionada con alguna regresora incluida o cuando la regresora omitida explica a la variable dependiente. Debido a que la regresora al no estar incluida explícitamente en el modelo, forma parte del

	término de error y al estar la regresora omitida correlacionada con la regresora incluida, entonces se concluye que el modelo especificado tiene correlación entre la regresora incluida y el término de error. • Debido a que la regresora omitida está en el término de error, la varianza de perturbación no está correctamente estimada, y al igual que la heterocedasticidad y la no autocorrelación, una varianza sesgada produce conclusiones equivocadas al momento de realizar pruebas de hipótesis tanto de significancia global e individual. De hecho esta es la razón por la cual los sesgos de especificación suelen ser la causa de violaciones a los supuestos de MCO. • Debido a que los estimadores no son consistentes, al aumentar la muestra no se obtendrán estimadores insesgados.
Sobreajuste (inclusión de una variable irrelevante)	• A diferencia del sesgo por subajuste, en un modelo sobreajustado los estimadores por MCO son insesgados y consistentes, sin embargo estos estimadores son ineficientes, es decir no tienen varianza mínima debido a que hay menos **grados de libertad**. • La varianza del error está correctamente estimado y por ello las pruebas de significancia global e individual conservan su validez.

Tabla 3.7. Consecuencias de la estimación por MCO con modelos que tienen sobreajuste y subajuste.
Elaboración propia
Fuente: (Gujarati & Porter, 2010) (Bravo & Vásquez Javiera, 2008) (De Grange C., 2005)

En los modelos econométricos se asume la ausencia de errores de medición u observación en las variables regresoras, sin embargo algunos autores consideran que este supuesto también debe tomarse en cuenta para las variables explicadas, es decir este supuesto sostiene que no existen errores de observación tanto para las variables explicativas como explicadas. (Wooldrige, 2009) Pone en claro que los errores de observación solo es un problema cuando las variables tienen datos que difieren de las variables que influyen en las decisiones de los sujetos.

Cuando se intentan estimar modelos econométricos con variables dependientes que son variables monetarias es frecuente la existencia de errores de medición en la variable explicada. Debido a que las familias suelen no revelar sus verdaderos ingresos a

los encuestadores, por lo general se recomienda precauciones al tratar con estas variables. Cuando un modelo econométrico tiene un error de medición en la variable dependiente se originan problemas para estimar mediante MCO. (Wooldrige, 2009) Explica el siguiente modelo econométrico.

$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$ (3.3.23.)

Donde asumimos que y^* es el ahorro familiar anual y el modelo (3.3.23.) es la función de regresión poblacional, por lo tanto al medir y^* utilizamos la variable y, la cual es la medición de la variable y^*. Sin embargo, al asumir que es posible que las variables y^* y y difieran, es decir que el valor real y el valor observado serán distintos por un error de medición, admitimos la existencia del error en la población, y este error es la diferencia entre el valor observado y el valor real. Lo denota de la siguiente manera.

$e_0 = y - y^*$ (3.3.24.)

La ecuación puede ser familiar con lo explicado anteriormente en los temas de la FRP y FRM, sin embargo no debe ser confundido con el término de perturbación ni el término residual respectivamente para la FRP y la FRM. Recordemos que el término de error en la FRP, se denomina a la diferencia entre los valores de *Y* con *E(Y|X)*, es decir la diferencia de cada valor de *Y* con la media condicional, mientras que el termino residual en la FRM es la diferencia entre el valor medido u observado con el valor ajustado o también llamado valor estimado. Lo que la ecuación (3.3.24.) manifiesta es la diferencia entre el valor observado, y, con el valor real, y^*. Reemplacemos (3.3.24.) en (3.3.23.)

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu + e$ (3.3.25.)

(Gujarati & Porter, 2010) Nombran a la expresión $(\mu + e)$ como **término de error compuesto**, que contiene al error de medición, e y al término de error μ. En realidad, se podría omitir este término de error compuesto y proseguir con la estimación de MCO. Suponiendo que cumplen los supuestos de MCO no habría ningún problema en la estimación. El problema de este tipo de errores de observación se expresa en la varianza de los estimadores, pues dejan de ser varianzas mínimas por lo que los estimadores no son eficientes.

Por otro lado, cuando los errores de observación están presentes en las variables regresoras los estimadores ya no son MELI. (Wooldrige, 2009) Plantea las siguientes ecuaciones y una explicación de estas.

$y = \beta_0 + \beta_1 x_1^* + \mu$ (3.3.26.)

Donde se asume que la variable y cumple los supuestos de MCO, mientras que la regresora al no estar correctamente medida o por ser inobservable, se usa la variable x_1, donde al igual que los errores de medición en la variable dependiente, el error se expresa de la siguiente forma:

$e = x_1 - x_1^*$ (3.3.27.)

Al despegar x_1^* de (3.3.27.) y reemplazarla en (3.3.26.) Se obtiene lo siguiente:

$$y = \beta_0 + \beta_1(x_1 - e) + \mu$$

$y = \beta_0 + \beta_1 x_1 + (\mu - \beta_1 e)$ (3.3.28.)

(Gujarati & Porter, 2010) Agregan la siguiente ecuación a partir de (3.3.28.) $y = \beta_0 + \beta_1 x_1 + z$, donde z es la composición de los errores ecuacional y de medición. El problema de (3.3.28.) Es que no se puede suponer que z es independiente de x_1, es decir no se puede asumir la ausencia de correlación entre el error combinado con la variable regresora, por ello es que se puede sospechar que estén relacionados con lo cual se generan estimadores sesgados e inconsistentes. A modo de conclusión, el error de medición en las regresoras afecta de peor manera la correcta estimación mediante MCO.

Finalmente, una condición para que obtengamos una correcta estimación es que el número de observaciones, *n*, debe ser igual o mayor al número de regresores, *k*.

3.3.4. Supuestos sobre los estimadores.

La correcta estimación mediante MCO, brinda estimadores que son MELI, y se les llama MELI porque cumplen ciertas propiedades, a continuación se vuelve a explicar las propiedades de los estimadores MELI.

- **Insesgado:** Se dice que un estimador o coeficiente de regresión es insesgado cuando el **valor esperado** del estimador muestral coincide con el verdadero valor del parámetro población. Matemáticamente se expresa de la siguiente manera:
- $E(\hat{\beta}) = \beta$ (1.5.4.)

- Al asumir que el valor esperado, conocido también como **esperanza**, **media** o **promedio**, del estimador muestral es igual verdadero valor del parámetro poblacional entonces el estimador muestral es insesgado.
- **Eficiente:** La eficiencia de un estimador muestral compara las varianzas de dos estimadores muestrales y elige al que tenga varianza mínima. Matemáticamente se expresa de la siguiente manera:
- $V(\widehat{\beta_1}) < V(\widehat{\beta_2})$ (1.5.5.)
- En algunos textos puede encontrarse la siguiente forma matemática:
- $\sigma^2_{\hat{\beta}_1} < \sigma^2_{\hat{\beta}_2}$ (1.5.6.)
- Tanto como (1.5.5.) cómo (1.5.6.) se puede interpretar que el estimador $\hat{\beta}_1$ es más eficiente que $\hat{\beta}_2$. En cuyo caso se prefiere el estimador $\hat{\beta}_1$ debido a que tiene una varianza mínima con respecto a $\hat{\beta}_2$. Para que un estimador sea eficiente debe cumplir la propiedad de insesgamiento.
- **Consistente:** Un estimador muestral es consistente cuando al ir aumentando el tamaño de la muestra, el estimador muestral se acerca al verdadero valor del parámetro poblacional.
- (Ponce A. & Nolberto S., 2008) Explican que esta propiedad se cumple debido a que al aumentar el tamaño de la muestra podemos estar más seguros que el error entre el estimador muestral y el parámetro población será menor y lo expresan matemáticamente:
- $\lim_{n\to\infty} P(|\hat{\beta} - \beta|) < c = 1$ (1.5.7.)
- Interpretan de la siguiente manera: en la ecuación (1.5.7.) el estimador muestral es consistente del parámetro poblacional si y solo si para cada c>0. En palabras sencillas, cuanto menor es la diferencia entre el estimador muestral y el parámetro poblacional con probabilidad uno, el estimador muestral se aproxima lo más posible al parámetro poblacional.
- **Suficiente:** Un estimador muestral es suficiente cuando se utiliza toda la información muestral para su estimación.

En los siguientes apartados se explicará porque se asume que estas propiedades se cumplen para los estimadores obtenidos mediante MCO. De momento es importante saber qué es lo que significa que el estimador cumpla con las propiedades.

3.3.5. Supuestos sobre la forma funcional.

3.3.5.1. Linealidad.

Anteriormente se ha explicado, que el supuesto de linealidad en los modelos econométricos permite medir el efecto de la variable exógena sobre la variable endógena, cuando esta primera aumenta su valor en una unidad. Sin embargo, existen otras formas funcionales que, aunque no son el tema principal de este trabajo es interesante tomarlas en cuenta para el desarrollo de otros modelos que explican mejor en algunos puntos que el modelo lineal. (Gujarati & Porter, 2010) Mencionan que las siguientes transformaciones cumplen el supuesto de linealidad en los parámetros más no en las variables.

3.3.5.1.1. Modelo log-lineal.

En los trabajos de economía se suele investigar la medición de la elasticidad de cierta variable, por ello la teoría econométrica ofrece este modelo para la correcta estimación de aquellos modelos.

(Gujarati & Porter, 2010) Expresan el término **modelo de regresión exponencial** para referirse a este tipo de modelos. Lo explica con las siguientes ecuaciones:

$$Y_i = \beta_1 X_i^{\beta_2} e^{\mu_i} \text{ (3.3.29.)}$$

La ecuación (3.3.29.) puede **transformarse** en la siguiente forma:

$$\ln Y_i = ln\beta_1 + \beta_2 lnX_i + \mu_i \text{ (3.3.30.)}$$

Donde, $\alpha = ln\beta_1$, entonces al estimar mediante MCO los parámetros α y β_2 tendrán estimadores que serán MELI ya que este modelo es lineal en los parámetros. Otros nombres que se le dan a este tipo de modelos es **log-log**, **doble-log** ya que se introducen los logaritmos en ambas partes de la ecuación.

Lo importante a destacar es que este modelo permite medir la elasticidad en el coeficiente β_2 de *Y* respecto a *X*, en palabras de (Gujarati & Porter, 2010) Miden el cambio porcentual de *Y* ante un pequeño cambio porcentual de *X*. Este modelo supone que el coeficiente de elasticidad, β_2, permanece constante sin importar cuanto el cambio de ln*X* haga cambiar a ln*Y*. Finalmente, el intercepto en estos modelos suelen estar sesgados pero su importancia es mínima, por lo que no debería generar preocupación en conseguir su insesgamiento.

3.3.5.1.2. Modelos semilogarítmicos.

La principal diferencia con el anterior modelo logarítmico, es que ahora solo se aplicara logaritmos en un solo lado de la ecuación, tanto en la variable dependiente o independiente.

- Modelo log-lin

Con este tipo de modelos **log-lin**, se permite medir la tasa de crecimiento, una variable muy importante. Para lograrlo, es común utilizar dos variables, la variable dependiente, *Y*, y el tiempo expresado en *t*. Se expresa de la siguiente manera:

$$lnY_t = \beta_1 + \beta_2 t + \mu_t \text{ (3.3.31.)}$$

Observe que el subíndice *t* indica que el modelo (3.3.31.) es un modelo con datos de serie de tiempo, en este caso solamente la variable regresada está expresada en su logaritmo mientras que la regresora es el tiempo que ocupa los valores de 1,2,3,…, t.

Lo importante de este modelo es que se busca medir el cambio porcentual o también llamado la tasa de crecimiento. En algunos textos es llamado **semielasticidad** de *Y* con respecto a *X*.

- Modelo lin-log

Por otro lado, cuando se especifica un modelo lin-log, se tiene el modelo:

$$Y_i = \beta_1 + \beta_2 \ln X_i + \mu_i \text{ (3.3.32.)}$$

Ahora se aplica un logaritmo a la(s) variable(s) explicativa(s) con el fin de medir el cambio absoluto de *Y* con el cambio porcentual de *X*.

(De Grange C., 2005) Explica otras transformaciones de variables.

- Transformación Box-Tidwell
- Transformación Box-Cox

3.3.5.2. Ausencia de errores de especificación en la función.

Este supuesto asume que la forma funcional es el correcto para especificar el modelo econométrico, sin embargo, al usarse el modelo lineal este supuesto podría obviarse.

Luego de haber leído todo sobre los supuestos de MCO tanto en el modelo simple como en el modelo múltiple, concluimos que la correcta especificación del modelo

garantiza que los estimadores sean MELI, el cual es el objetivo de la estimación. Cuando no se cumple este supuesto, el modelo puede contener algún problema causado por la violación de los supuestos de MCO, de ser así entonces tendría que aplicarse medidas correctivas a la estimación. En el siguiente apartado se explicara entonces el proceso de estimar mediante Mínimos Cuadrados Ordinarios y como se usan estos supuestos para la estimación de los estimadores muestrales.

3.4. Estimación del Modelo de Regresión Múltiple mediante Mínimos Cuadrados Ordinarios

Previamente a pasar a explicar el método de estimación por MCO, es necesario recordar que los coeficientes que se pretenden estimar son de la función de regresión muestral, los cuales al cumplir la propiedad de insesgadez se acercan a los parámetros poblacionales. Esto se asume debido a que es imposible estimar parámetros poblacionales, por eso es que se usa una muestra representativa de las variables a relacionar.

3.4.1. Estimación de modelos de regresión simple mediante MCO.

Aunque este modelo es menos frecuente que el modelo de regresión múltiple, se explicará cómo se estima mediante MCO con el objetivo de explicar de manera sencilla los términos que se emplean en la estimación, para posteriormente explicar la estimación en los modelos de regresión múltiple.

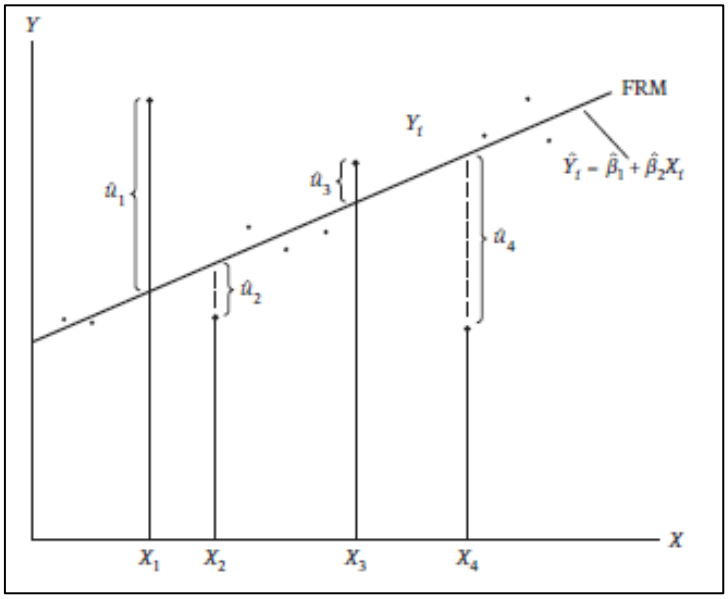
Tal como su nombre lo indica, este método de estimación consiste en básicamente minimizar el valor de los residuos, en palabras más técnicas, (Novales, 1998) Explica textualmente.

> *"El estimador de mínimos cuadrados que introducimos en esta sección utiliza como criterio la minimización de la Suma de los Cuadrados de los Residuos, habitualmente denominada Suma Residual, y denotada por SR."* (Novales, 1998)

(Cid S., Mora C., & Valenzuela H., 1990) Refuerzan la idea expresando que lo que se busca es que la dispersión de los valores muestreados u observados de la endógena sea la más mínima posible con respectivo al valor de su media. Recuerde que a esa *dispersión* en la FRM, se le conoce como término residual y se le expresa de la siguiente manera:

$$\widehat{\mu_i} = Y_i - \widehat{Y_i} \quad (3.4.1.)$$

Pues bien, este es el punto de partida para entender la estimación por Mínimos Cuadrados Ordinarios. (Gujarati & Porter, 2010) Expresan lo anterior en el siguiente gráfico.



Gráfica 3.11. Criterio de mínimos cuadrados.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

El gráfico 3.11. Muestra cómo cada punto de la línea estimada FRM, es el valor estimado de *Y* para cada valor de *X*, alrededor de la línea existen puntos a diferentes dispersiones de cada valor de la línea. **El principio de mínimos cuadrados**, tal como ya se dijo es reducir lo más posible la Suma Residual. La Suma Residual se expresa de la siguiente forma:

$$\sum \hat{\mu}_i = \sum (Y_i - \hat{Y}_i) \text{ (3.4.2.)}$$

(Gujarati & Porter, 2010) Explica la existencia de un problema en la ecuación (3.4.2.).

> *"En otras palabras, a todos los residuos se les da la misma importancia sin considerar cuán cerca o cuán dispersos estén de las observaciones individuales de la FRM."* (Gujarati & Porter, 2010)

Por esto es que la suma residual en la mayoría de los casos es igual a 0. Entonces ¿Cómo se logra anular este problema? La solución es elevando al cuadrado a los residuos. De tal manera que (3.4.2.) ahora se expresa cómo:

$$\sum \hat{\mu}_i^2 = \sum (Y_i - \hat{Y}_i)^2 \text{ (3.4.3.)}$$

Al elevar al cuadrado los residuos, permitimos que la suma residual sea la más mínima posible sin importar cuan distribuidos están los residuos de la línea estimada. La forma (3.4.3.) también puede ser descrita como:

$$\sum \hat{\mu}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad (3.4.4.)$$

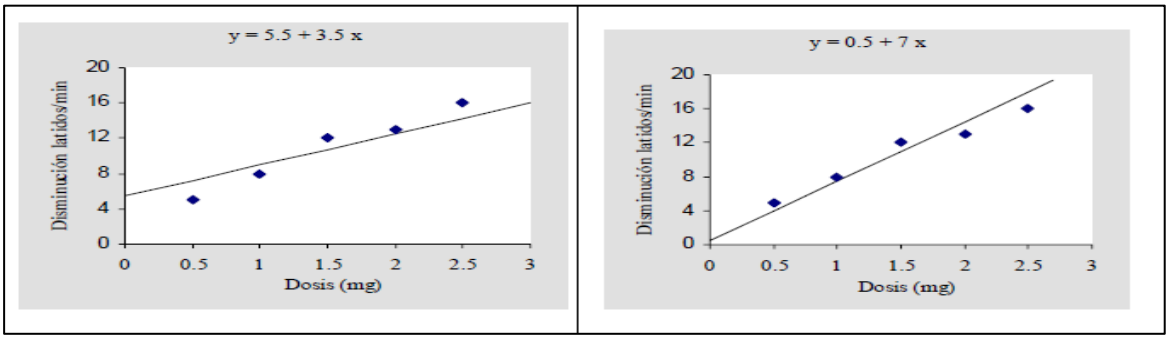
(Novales, 1998) Aclara que estos coeficientes de regresión, se les conoce como estimadores de MCO y se debe escoger la recta que minimiza la suma de los cuadrados de los residuos (**SCR**). (Orellana, 2008) Aclara la idea anterior con el siguiente ejemplo. El siguiente cuadro muestra la información sobre 5 sujetos de prueba sometidos a ser suministrados cada uno con una dosis en mg de cierta droga y también muestra la máxima disminución de la FC (DFC) de cada uno de ellos, Siendo el modelo especificado: $DFC = \beta_0 + \beta_1 * DOSIS + \mu$, podemos darnos cuenta que la variable DOSIS explica a la DFC y que además se trata de una función de regresión poblacional, por lo que se debe encontrar los estimadores.

Dosis(mg)	Máxima disminución de la FC (DFC)
0.5	5
1.0	8
1.5	12
2.0	13
2.5	16

Tabla 3.8. Datos de DOSIS y DFC
Elaboración (Orellana, 2008)
Fuente (Orellana, 2008)

(Orellana, 2008)

Intenta ajustar o estimar la recta de regresión lineal otorgando valores a los coeficientes. Siendo $\widehat{DFC} = 5.5 + 3.5 * DOSIS + \mu$ y $\widehat{DFC} = 0.5 + 7.0 * DOSIS + \mu$ las rectas *estimadas* otorgando valores estimados de la variable dependiente. Veamos cómo se



distribuye cada uno.

Ya anteriormente se dijo que el modelo con menor suma residual cuadrática es preferible sobre otros. Calculemos la SRC de cada línea de regresión. Para ello primero se empieza calculando los valores de la línea estimada de regresión, es decir los valores que conforman dicha línea de regresión, para lograrlo se reemplaza en las ecuaciones con los estimadores otorgados para cada valor de la regresora, de tal forma que las tablas (3.9.) y (3.10.) muestran el procedimiento. Posteriormente, se hallan los residuos restando los valores de la variable dependiente *Y* observada con los valores de la variable *Y* estimada. Los residuos hallados los elevamos al cuadrado y finalmente sumamos sus potencias y habremos obtenido los SRC de ambas ecuaciones. Observe (3.9.) y (3.10.).

Y	X	$\widehat{Y}$	$\mu = Y - \widehat{Y}$	μ^2
5	0.5	7.3	-2.3	5.1
8	1	9.0	-1.0	1.0
12	1.5	10.8	1.3	1.6
13	2	12.5	0.5	0.3
16	2.5	14.3	1.8	3.1
SUMA				10.9

→ Esta tabla se calcula con la ecuación:

$$\widehat{DFC} = 5.5 + 3.5 * DOSIS + e$$

Siendo su SRC expresada a partir de la ecuación:

$$\sum (Y_i - 5.5 - 3.5X_i)^2$$

Tabla 3.9. Datos de DOSIS y DFC para la primera ecuación.
Elaboración propia
Fuente (Orellana, 2008)

Y	X	$\widehat{Y}$	$\mu = Y - \widehat{Y}$	μ^2
5	0.5	4.0	1.0	1.0
8	1	7.5	0.5	0.25
12	1.5	11.0	1.0	1.0
13	2	14.5	-1.5	2.25
16	2.5	18.0	-2.0	4.0
SUMA				8.50

→ Esta tabla se calcula con la ecuación:

$$\widehat{DFC} = 0.5 + 7.0 * DOSIS + e$$

Siendo su SRC expresada a partir de la ecuación:

$$\sum (Y_i - 0.5 - 7X_i)^2$$

Tabla 3.10. Datos de DOSIS y DFC para la segunda ecuación.
Elaboración propia
Fuente (Orellana, 2008)

La segunda ecuación tiene una SRC menor que la primera ecuación, por lo tanto es preferible usar la segunda ecuación para medir el efecto de la regresora sobre la endógena, sin embargo, debido a que la población es compleja de explicar podrían haber

otros estimadores que tengan SRC menores, por lo que para encontrarlos se ejecuta la estimación por Mínimos Cuadrados Ordinarios. Para ello, se hace uso de un sistema de ecuaciones conformado por **las ecuaciones normales.** Las cuales para ser calculadas primero se somete a derivadas la Sumatoria Residual Cuadrática, que es la expresión (3.4.4.), con respecto a cada uno de sus estimadores e igualadas a cero. En el caso de la regresión simple, se derivaran la pendiente y el intercepto por lo que solamente se generarán dos ecuaciones normales. A continuación (Novales, 1998) Muestra el proceso:

$\frac{\partial SR}{\partial \widehat{\beta_1}} = -2\sum\left(Y_i - \widehat{\beta_1} - \widehat{\beta_2}X_i\right) = 0$ (3.4.5.)

$\frac{\partial SR}{\partial \widehat{\beta_2}} = -2\sum\left(Y_i - \widehat{\beta_1} - \widehat{\beta_2}X_i\right)x_i = 0$ (3.4.6.)

Estas ecuaciones también puedes escritas como $-2\sum \hat{\mu}_i$ y $-2\sum \hat{\mu}_i X_i$ respectivamente, las cuales son las condiciones de primer orden. Con (3.4.5.) y (3.4.6.) se tomarán sus segundas derivadas respecto a los parámetros, con el fin de construir una matriz Hessiana, la cual al ser un modelo simple será de *2x2*.

$H_{2x2} = \begin{pmatrix} 2n & 2\sum X_i \\ 2\sum X_i & 2\sum {X_i}^2 \end{pmatrix}$ (3.4.7.)

La determinante de la matriz se calcula de la siguiente forma:

$|H| = 4(n\sum {X_i}^2 - ((\sum X_i)^2) = n^2\left(\frac{\sum {X_i}^2}{n} - \overline{X_i}^2\right) = n^2\frac{\sum(X_i - \overline{X_i})^2}{n} = n^2 S_X^2$ (3.4.8.)

Donde S_x^2 es la varianza muestral de *X*. Con la matriz podemos resolver el sistema de ecuaciones, de tal forma que ahora podemos construir las **ecuaciones normales**.

(Novales, 1998) Interpreta que la solución a las ecuaciones (3.4.5.) y (3.4.6.) serán los valores numéricos de los parámetros, las siguientes ecuaciones son las ecuaciones normales y por lo tanto la solución al sistema de ecuaciones:

$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum Y_i X_i$ (3.4.9.)

$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum {X_i}^2$ (3.4.10.)

Finalmente, para obtener la fórmula con la cual hallar el valor de los parámetros, primero despejamos $\hat{\beta}_1$ en (3.4.9.)

$\hat{\beta}_1 = \frac{\sum Y_i - \hat{\beta}_2 \sum X_i}{n} = \overline{Y}_i - \hat{\beta}_2 \overline{X}_i$ (3.4.11.)

Donde, $\overline{Y}_i$ y $\overline{X}_i$ son los promedios de Y y X respectivamente. Y para hallar el valor de $\hat{\beta}_2$ sustituiremos el estimador $\hat{\beta}_1$ en (3.4.10.)

$$\hat{\beta}_2 = \frac{\sum Y_i X_i - \frac{1}{n}(\sum X_i)(\sum Y_i)}{\sum {X_i}^2 - \frac{1}{n}((\sum Y_i)^2)} = \frac{\sum (X_i - \overline{X}_i)(Y_i - \overline{Y}_i)}{\sum (X_i - \overline{X}_i)^2} = \frac{\sum x_i y_i}{\sum x_i^2} \text{ (3.4.12.)}$$

(Gujarati & Porter, 2010) Llaman media muestral a $x_i = \left(X_i - \overline{X}_i\right)$ y $y_i = (Y_i - \overline{Y}_i)$, es decir a la diferencia entre el valor observado con su media. Para que quede claro proseguiremos con el ejemplo de (Orellana, 2008) Pero ahora estimaremos los estimadores muestrales mediante MCO usando las fórmulas (3.4.11.) y (3.4.12.)

$$\hat{\beta}_1 = \overline{Y}_i - \hat{\beta}_2 \overline{X}_i = 10.8 - 5.4 * 1.5 = 2.7 \text{ (3.4.13.)}$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{13.5}{2.5} = 5.4 \text{ (3.4.14.)}$$

Con los parámetros calculados en (3.4.13.) y (3.4.14.) podemos construir la recta de regresión.

$$\widehat{DFC} = 2.7 + 5.4 * DOSIS + \mu \text{ (3.4.15.)}$$

Donde la ecuación (3.4.15.) es el modelo de regresión muestral estimado, observe que la variable dependiente tiene un gorrito, la cual indica que corresponde a un modelo estimado. Donde 2.7 es el intercepto y 5.4 la pendiente. La tabla 3.11. Muestra las sumatorias necesarias para calcular los estimadores muestrales.

N	Y_i	X_i	$y_i = (Y_i - \overline{Y}_i)$	$x_i = \left(X_i - \overline{X}_i\right)$	x_i^2	$x_i y_i$
1	5	0.5	-5.8	-1	1	5.8
2	8	1	-2.8	-0.5	0.25	1.4
3	12	1.5	1.2	0	0	0
4	13	2	2.2	0.5	0.25	1.1
5	16	2.5	5.2	1	1	5.2
Sumatoria	54	7.5	0	0	2.5	13.5
Promedio	10.8	1.5				

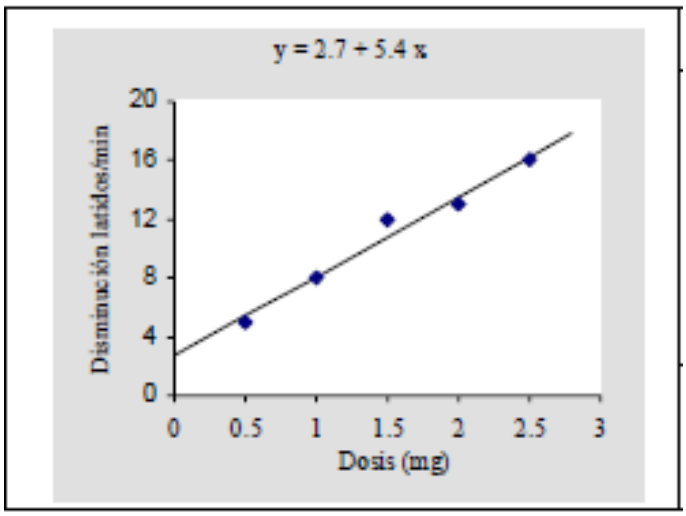
Tabla 3.11. Ejemplo de estimación de un modelo simple mediante MCO.
Elaboración (Orellana, 2008)
Fuente (Orellana, 2008)

Usando la ecuación (3.4.15.) podemos calcular la *Y* estimada y los errores, posteriormente se gráfica los valores estimados, en algunos textos se les conoce como valores ajustados. Solamente reemplazamos el valor de los estimadores en la ecuación (3.4.15.) para cada valor de X_i. Por ejemplo, el primer valor, seria: $\widehat{Y_1} = 2.7 + 5.4 * 0.5 = 5.4$ y así sucesivamente. La tabla 3.12. Muestra el resto de los valores de la *Y* estimada, los residuos y los residuos al cuadrado.

N	Y_i	X_i	$\widehat{Y_i} = 2.7 + 5.4 * X_i + \widehat{\mu_i}$	$\widehat{\mu_i} = (Y_i - \widehat{Y_i})$	$\widehat{\mu^2}_i$
1	5	0.5	5.4	-0.4	0.16
2	8	1	8.1	-0.1	0.01
3	12	1.5	10.8	1.2	1.44
4	13	2	13.5	-0.5	0.25
5	16	2.5	16.2	-0.2	0.04

Tabla 3.12. Ejemplo de estimación de un modelo simple mediante MCO (2).
Elaboración (Orellana, 2008)
Fuente (Orellana, 2008)

Con los datos calculados que se muestran en la tabla 3.12. Podemos construir un gráfico de regresión que contenga la línea de regresión, y alrededor de esta línea estarán los residuos. Tal como se muestra a continuación.



Gráfica 3.13. Gráfico de regresión del ejemplo.
Elaboración: (Orellana, 2008)
Fuente: (Orellana, 2008)

Finalmente, podemos interpretar la ecuación $\widehat{DFC} = 2.7 + 5.4 * DOSIS + \mu$. El intercepto, es decir 2.7, indica el punto en el que la línea de regresión choca con el eje vertical, por lo que se puede interpretar como la disminución de la frecuencia cardiaca esperada cuando la dosis es cero. Por otro lado, la pendiente, 5.4, indica que por cada aumento en una unidad de mg de dosis suministrada a los sujetos de prueba, la DFC aumentó en 5.4 pulsaciones/min. También observe que $\widehat{\mu^2}_\iota = 1.90$, esta es la Suma Residual Cuadrática, y mediante MCO se ha elegido la SRC mínima del modelo la cual es 1.90.

Usando los principios de estimación de MCO para el modelo simple, se puede estimar los modelos múltiples de regresión. En el siguiente apartado se explica el procedimiento.

3.4.2. Estimación del modelo de regresión múltiple mediante MCO.

El mismo principio usado para estimar el modelo simple se repite para estimar el modelo múltiple de regresión. Comencemos especificando el modelo de regresión múltiple con la función de regresión muestral FRM.

$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki} + \widehat{\mu_\iota}$ (3.4.16.)

Recordando que al usar MCO se intentará minimizar la Suma Residual Cuadrática. En un modelo múltiple se expresa de la siguiente forma.

$$SRC = min \sum(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki} + \hat{\mu}_i)^2 \text{ (3.4.17.)}$$

La metodología es la misma que en el modelo anterior, comenzamos derivando a la SRC con respecto a los parámetros de tal forma que conseguimos el siguiente sistema de ecuaciones:

$$\frac{\partial SRC}{\partial \hat{\beta}_1} = 2 \sum(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki}) = 0 \text{ (3.4.18.)}$$

$$\frac{\partial SRC}{\partial \hat{\beta}_2} = 2 \sum(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki}) X_{2i} = 0 \text{ (3.4.19.)}$$

$$\frac{\partial SRC}{\partial \hat{\beta}_3} = 2 \sum(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki}) X_{3i} = 0 \text{ (3.4.20.)}$$

$$\vdots$$

$$\frac{\partial SRC}{\partial \hat{\beta}_k} = 2 \sum(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki}) X_{ki} = 0 \text{ (3.4.21.)}$$

Al ser igualadas a cero las anteriores expresiones, se consigue el siguiente sistema de ecuaciones normales.

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \cdots + \widehat{\beta_k} \sum X_{ki} \text{ (3.4.22.)}$$

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum {X_{2i}}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} + \cdots + \widehat{\beta_k} \sum X_{2i} X_{ki} \text{ (3.4.23.)}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum {X_{3i}}^2 + \cdots + \widehat{\beta_k} \sum X_{3i} X_{ki} \text{ (3.4.24.)}$$

$$\vdots$$

$$\sum Y_i X_{ki} = \hat{\beta}_1 \sum X_{ki} + \hat{\beta}_2 \sum X_{2i} X_{ki} + \hat{\beta}_3 \sum X_{3i} X_{ki} + \cdots + \widehat{\beta_k} \sum {X_{ki}}^2 \text{ (3.4.25.)}$$

Debido a que resulta complicado y sumamente difícil despejar los parámetros y calcular las sumatorias, más aún cuando se trabajan con muestras demasiado grandes, se hace uso del álgebra matricial para estimar MCO.

3.4.2.1. Estimación MCO mediante el uso de matrices.

(Pérez L., 2012) Explica que para la notación matricial se adopta la forma:

$$y = X\beta + \mu \text{ (3.4.26.)}$$

Donde:

- y es una matriz vector de *nx1* que representa los valores de la variable dependiente.

- X es una matriz de *nxk* que contiene las variables independientes, las cuales el número de filas es el número de observaciones y el número de columnas son el número de parámetros tomando en cuenta el intercepto, por lo que *k-1* es el número de variables explicativas.
- β es una matriz vector de *kx1* donde el número de filas es el número de parámetros tomando en cuenta al intercepto.
- μ es una matriz vector de *nx1* que contiene el número de residuos.

(Uriel & Aldás, 2005) Manifiestan que esta forma matricial, parte de un sistema de ecuaciones, siendo más específicos de las funciones de regresión poblacionales. Recuerde que:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu_i \text{ (3.4.27.)}$$

Equivale a decir:

$$Y_1 = \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + \mu_1 \text{ (3.4.28.)}$$

$$Y_2 = \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + \mu_2 \text{ (3.4.29.)}$$

$$\vdots$$

$$Y_k = \beta_1 + \beta_2 X_{2k} + \beta_3 X_{3k} + \cdots + \beta_k X_{kk} + \mu_k \text{ (3.4.30.)}$$

Las ecuaciones anteriores pueden representarse en una forma matricial:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \text{ (3.4.31.)}$$

La forma matricial (3.4.26.) representa a las matrices (3.4.31.), donde en la matriz X la primera columna se agrega una columna compuesta por 1 para calcular el intercepto. Teniendo esto como base, ahora podemos estimar la FRP con la ayuda de la FRM. Una vez más se muestra el modelo especificado de la función de regresión muestral.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \widehat{\beta_k} X_{ki} + \widehat{\mu_i} \text{ (3.4.32.)}$$

Cuya forma matricial es:

$$y = X\hat{\beta} + \hat{\mu} \text{ (3.4.33.)}$$

Observe cómo incluso en su forma matricial, se vuelve a colocar el *sombrero* tanto al vector que representa a los estimadores como al vector que representa los residuos. La expresión (3.4.33.) también puede ser ampliada como la expresión (3.4.31.). De tal manera que se denota:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_k \end{bmatrix} \text{ (3.4.34.)}$$

Y al igual que en el método anterior, para estimar el vector de los estimadores se trata de minimizar la SRC, que cuya forma matricial es:

$$\hat{\mu}'\hat{\mu} = [\hat{\mu}_1 \quad \hat{\mu}_2 \quad \cdots \quad \hat{\mu}_k] \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_k \end{bmatrix} = \sum \hat{\mu}_i{}^2 = SRC \text{ (3.4.35.)}$$

Al tomar en cuenta que el término residual es la diferencia entre el valor observado o muestreado de la variable dependiente y el valor estimado de la variable dependiente. Podemos denotar la siguiente expresión matricial:

$\hat{\mu} = y - \hat{y} = y - X\hat{\beta}$ (3.4.36.)

Si reemplazamos (3.4.36.) en (3.4.35.) entonces estamos denotando la Suma Cuadrática Residual en su forma matricial aún más extensa:

$$SRC = (y - X\hat{\beta})'(y - X\hat{\beta}) \text{ (3.4.37.)}$$

Por consiguiente al recordar que: $(X\hat{\beta})' = \hat{\beta}'X'$ y la propiedad: $y'X\hat{\beta} = \hat{\beta}'X'y$ se obtiene:

$SRC = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$ (3.4.38.)

Donde al derivar SRC con respecto a su vector columna de los estimadores e igual a cero obtenemos:

$$\frac{\partial SRC}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \rightarrow X'X\hat{\beta} = X'y \text{ (3.4.39.)}$$

Donde (3.4.39.) corresponde a la forma matricial de las ecuaciones normales. Es decir, las expresiones:

$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_{2i} + \hat{\beta}_3 \sum X_{3i} + \cdots + \widehat{\beta_k} \sum X_{ki}$ (3.4.22.)

$$\sum Y_i X_{2i} = \hat{\beta}_1 \sum X_{2i} + \hat{\beta}_2 \sum {X_{2i}}^2 + \hat{\beta}_3 \sum X_{2i} X_{3i} + \cdots + \widehat{\beta_k} \sum X_{2i} X_{ki} \text{ (3.4.23.)}$$

$$\sum Y_i X_{3i} = \hat{\beta}_1 \sum X_{3i} + \hat{\beta}_2 \sum X_{2i} X_{3i} + \hat{\beta}_3 \sum {X_{3i}}^2 + \cdots + \widehat{\beta_k} \sum X_{3i} X_{ki} \text{ (3.4.24.)}$$

$$\vdots$$

$$\sum Y_i X_{ki} = \hat{\beta}_1 \sum X_{ki} + \hat{\beta}_2 \sum X_{2i} X_{ki} + \hat{\beta}_3 \sum X_{3i} X_{ki} + \cdots + \widehat{\beta_k} \sum {X_{ki}}^2 \text{ (3.4.25.)}$$

Es igual a la siguiente forma matricial:

$$\begin{bmatrix} n & \sum X_{2i} & \sum X_{3i} & \cdots & \sum X_{ki} \\ \sum X_{2i} & \sum {X^2}_{2i} & \sum X_{3i} X_{2i} & \cdots & \sum X_{ki} X_{2i} \\ \sum X_{3i} & \sum X_{2i} X_{3i} & \sum {X^2}_{3i} & \cdots & \sum X_{ki} X_{3i} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum X_{ki} & \sum X_{2i} X_{ki} & \sum X_{3i} X_{ki} & \cdots & \sum {X^2}_{ki} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} =$$

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & X_{31} & \cdots & X_{2n} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{k1} & X_{k2} & X_{k3} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_k \end{bmatrix} \text{ (3.4.40.)}$$

Siendo (3.4.40.) equivalente a (3.4.39.), donde $(X'X)$ tiene características importantes ya que posteriormente será usado para calcular la varianza de los estimadores. En palabras de (Gujarati & Porter, 2010) La diagonal principal son las sumas simples de los cuadrados, mientras que los elementos que no conforman la diagonal principal son las sumas simples de productos cruzados. Ahora para *despejar* el vector columna de los estimadores, aplicaremos la inversa de $(X'X)$. De esta forma:

$$(X'X)^{-1} X'X\hat{\beta} = (X'X)^{-1} X'y \text{ (3.4.41.)}$$

Donde $(X'X)^{-1} X'X$ equivale a la matriz identidad, es decir $(X'X)^{-1} X'X = I$, por lo que:

$$\hat{\beta} = (X'X)^{-1} X'y \text{ (3.4.42.)}$$

Donde: $\hat{\beta}$ es una matriz vector columna *kx1*, $(X'X)^{-1}$ es una matriz de *kxk*, *y* es una matriz *nx1* y X'es una matriz de *kxn.*

3.4.3. El valor esperado y la varianza de los estimadores en el modelo de regresión simple y en el modelo de regresión múltiple.

3.4.3.1. Esperanza de los estimadores y el cumplimiento del insesgamiento.

Recordando teoría estadística, existen dos medidas de dispersión: la esperanza o media y la varianza. La esperanza es el valor central de una variable aleatoria en una gráfica de dispersión mientras que la varianza mide la dispersión al cuadrado de cada valor observado con respecto al valor central.

Entonces cuando estimamos los estimadores de la función muestral, estamos suponiendo que el valor estimado de los estimadores es igual al valor de los parámetros poblacionales, y esto es una propiedad que debe ser cumplida para obtener buenos estimadores, pero ¿de dónde sale esta suposición? La respuesta es muy fácil de encontrar. Si recordamos la teoría mostrada anteriormente, los parámetros poblacionales especificados en la función poblacional son totalmente desconocidos e imposibles de calcular, entonces hacemos uso de una muestra para estimar los parámetros poblacionales a partir de estimadores. En otras palabras estamos suponiendo que los estimadores son **insesgados**. Al hablar de insesgamiento de los estimadores, podemos denotar mediante:

$$E(\hat{\beta}) = \beta \quad (1.5.4.)$$

Dónde (1.5.4.) equivale a $E(\hat{\beta}) - \beta = 0$ y se lee textualmente: "*el valor esperado de los estimadores es igual al verdadero valor poblacional"* o *"el valor esperado de beta estimado es igual al verdadero valor del beta poblacional."* Ambas formas son válidas de leer (1.5.4.).

Puede resultar confusa la expresión anterior debido a no tratarse de una variable en sí, sino de los estimadores entonces ¿Cómo puede darse el caso de suponer insesgamiento? La respuesta puede ser inferida teniendo en cuenta que, la población es todo el universo que se quiere estudiar, pero como no es posible abarcar toda la población en una investigación hacemos uso de la muestra, que a diferencia de la población que es solo una sola, se pueden utilizar varias muestras para estimar a la misma población. Al momento de estimar, mediante MCO o cualquier otro método, lo que se pretende es que el valor estimado sea lo más cercano al valor verdadero de la población. En otras palabras

se busca que el valor esperado de beta estimado sea igual que el verdadero valor del beta poblacional.

La propiedad de insesgamiento de los estimadores es demostrable usando matrices, (De Grange C., 2005) Detalla el procedimiento para comprobarlo tomando en cuenta que $\hat{\beta} = (X'X)^{-1}X'y$, se expresa:

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \mu) \text{ (3.4.43.)}$$

$$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mu \text{ (3.4.44.)}$$

$$\hat{\beta} = \beta + (X'X)^{-1}X'\mu \text{ (3.4.45.)}$$

Donde X es una matriz fija y $\hat{\beta}$ es una matriz vector fija, por lo que aplicando esperanzas a ambos lados, se obtiene:

$$E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'\mu] \text{ (3.4.46.)}$$

$$E(\hat{\beta}) = \beta + (X'X)^{-1}E(X')E(\mu) \text{ (3.4.47.)}$$

Al recordar que la media del término de perturbación es 0, entonces:

$$E(\hat{\beta}) = \beta \text{ (3.4.48.)}$$

Una aclaración, tanto las expresiones (1.5.4.) y (3.4.48.) representan la propiedad de insesgamiento, la diferencia está marcada en que (1.5.4.) está denotada en forma ecuacional mientras (3.4.48.) está en forma matricial. Algo similar ocurre con $E(\mu_i) = 0$ que representa la media del término de perturbación en su forma ecuacional, mientras que $E(\mu) = 0$ también denota el mismo supuesto, pero en su forma matricial.

3.4.3.2. La varianza y el error estándar de la regresión.

Al hablar de la existencia de la esperanza de los estimadores, también puede admitirse la existencia de las **varianzas** y los **errores estándares** de los estimadores.

Previamente se debe dar a conocer que son las varianzas y los errores estándares de los estimadores. La siguiente cita textual, podría aclarar los conceptos:

> *"Además de saber que la distribución muestral de $\hat{\beta}$ está centrada en β ($\hat{\beta}$ es insesgado), también es importante saber que tanto puede esperarse que $\hat{\beta}$ se aleje, en promedio, de β."* (Wooldrige, 2009)

Si tomamos en cuenta que se puede tomar varias muestras para explicar a la misma población, podríamos hallar diferentes estimadores provenientes de todas esas muestras, y todos estos estimadores estarían dispersos alrededor de la esperanza del estimador muestral usado para hacer la estimación en la población. Por lo que, como si de una variable se tratase, es necesario conocer cuán alejados o dispersos en promedio están esos estimadores de la esperanza del estimador muestral. Este concepto presentado por (Wooldrige, 2009) Hace referencia al error estándar y a la varianza del estimador. **Podríamos hacer un paralelismo entre el error estándar y la desviación estándar o típica. La primera mide la dispersión en promedio de los valores de todos los estimadores provenientes de un número indeterminado de muestras alrededor de su valor esperado (promedio),** mientras la desviación típica mide la dispersión en promedio de los valores de una variable con respecto a su valor esperado. En el caso de la varianza del estimador o de una variable, ambas miden la dispersión anteriormente mencionada al cuadrado. Otra característica similar es que, a menor dispersión, los valores están más cercanos a su valor medio, lo cual es preferible a una dispersión mayor en la que los valores están más alejados de su valor medio. Este se cumple tanto para la desviación estándar y el error estándar.

Para explicar cómo calcular la varianza de los estimadores y su posterior error estándar, es necesario entender el concepto de **varianza del error** y **desviación del error**. La desviación del error, es también llamada **error cuadrático medio**, **error estándar de la regresión** y otros. En realidad, no importa el nombre con el que se le conozca sino entender cómo se calcula y que es lo que significa.

Retomemos el concepto de la función de regresión poblacional, al igual que los parámetros poblacionales, la varianza poblacional representada con σ^2 es totalmente desconocida, es por tanto que al hacer uso de la muestra se pretende estimar la varianza del error. Previamente a explicar el cálculo de la varianza del error, es necesario aclarar cualquier duda sobre la expresión **término de error** y **término residual**, ambos tienen significados parecidos, pero son diferentes, el primero es propio de la función de regresión poblacional mientras que el segundo es propio de la función de regresión muestral. Donde el término residual pretende ser la variable que estime al término de error, ya que la FRP es totalmente desconocida, no solo en sus parámetros sino también en los valores de sus variables. Por lo tanto: los errores son inobservables mientras que los residuos son

completamente observables. De esta manera, al pretender estimar la varianza del error, en lugar de usar a los mismos errores se usarán a los residuos.

(Wooldrige, 2009) Detalla cómo se calcula la varianza del error. Teniendo en cuenta que el valor esperado del término de perturbación al cuadrado es igual a la varianza del error, de tal forma que:

$$var(\mu) = E(\mu^2) = \sigma^2 \text{ (3.4.49.)}$$

La expresión (3.4.49.) es el supuesto de homocedasticidad, por lo que para estimar correctamente a la varianza del error se hace uso del supuesto que la varianza del término de error es constante. De esta manera para estimar la varianza del error poblacional se hace uso del supuesto de homocedasticidad, la cual puede ser reemplazada y podríamos usar la fórmula:

$$n^{-1} \sum \mu^2 = \sigma^2 \text{ (3.4.50.)}$$

Sin embargo, al ser el término de error totalmente inobservable se pueden usar los residuos, después de todo el término residual es el estimador del término de perturbación, de esta manera (3.4.50.) puede escribirse como:

$$n^{-1} \sum \hat{\mu}^2 = \hat{\sigma}^2 \text{(3.4.51.)}$$

Si observamos con cuidado (3.4.51.) notaremos que la sumatoria residual cuadrática está explícitamente en la fórmula, por lo que (3.4.51.) equivale a:

$$\frac{SRC}{n} = \hat{\sigma}^2 \text{ (3.4.52.)}$$

(Pérez L., 2012) Advierte que (3.4.52.) generaría un estimador sesgado, es decir que el estimador de la varianza del error sería diferente de la varianza poblacional. Este sesgamiento se origina según (Uriel & Aldás, 2005) ya que no se ha tomado en cuenta las restricciones presentes en las ecuaciones normales sobre los residuos, (Wooldrige, 2009) Menciona estas restricciones, en el caso del modelo de regresión simple, serían las siguientes dos restricciones:

$$\sum \hat{\mu}_i = 0 \text{ (3.4.53.)}$$

$$\sum X_i \hat{\mu}_i = 0 \text{ (3.4.54.)}$$

Por lo tanto (3.4.52.) genera el estimador insesgado cuando:

$\frac{SRC}{n-2} = \hat{\sigma}^2$ (3.4.55.)

El denominador en la anterior formula, se le conoce como **grados de libertad.** La importancia de denominar el concepto de grados de libertad, se debe a Sir Ronald Fisher. (De la Cruz-Ore, 2013) Brevemente explica la definición propuesta por Fisher, sustentado en los trabajos de Gauss, que los grados de libertad hacen referencia a la diferencia entre el número de observaciones y el número de parámetros desconocidos a estimar incluido el intercepto. De ser así, entonces la fórmula (3.4.55.) para la regresión múltiple sería:

$\frac{SRC}{n-k} = \hat{\sigma}^2$ (3.4.56.)

Donde *k* es el número de estimadores en el modelo incluido el intercepto. Existen muchas formas de plasmar el denominador a la hora de calcular la varianza del error, sea cual sea la expresión encontrada en los libros serios de econometría debe entenderse que el numerador siempre será la sumatoria residual cuadrática y el denominador es la diferencia entre el **número de observaciones y el número de parámetros a estimar (incluido el intercepto).**

Al igual que la varianza de una variable, la cual al calcular su raíz cuadrática se obtiene la desviación estándar, por ello es que, a partir de la varianza del error podemos calcular el error estándar de regresión usando:

$$\hat{\sigma} = \sqrt{\frac{SRC}{n-k}} \text{ (3.4.57.)}$$

Aunque pueda parecer un simple estimador calculado a partir de la varianza del error, el concepto que tiene el error de regresión lo hace uno de los estimadores más importantes al momento de elegir modelos econométricos. Por lo general, se recomienda que el modelo econométrico con el error estándar de regresión mínimo sea elegido sobre otros. (Wooldrige, 2009) Explica con esta cita textual.

> *"La $\hat{\sigma}$estimada es interesante porque es una estimación de la desviación estándar de los factores no observables que afectan a Y; de manera equivalente, es una estimación de la desviación estándar de Y después de haber eliminado el efecto de X."* (Wooldrige, 2009)

Lo que se intenta explicar en la cita anterior, es que el error estándar de regresión mide cómo la variable dependiente es afectada por los factores no observados, que son

especificados por la variable término residual en la función de regresión muestral, siendo más precisos, este estimador pretende dar una medición sobre cómo los residuos hacen variar a la variable dependiente. Por ello es que se elige o se prefiere un modelo econométrico con el error estándar de regresión lo más mínimo posible.

A continuación, vamos a demostrar el cálculo de la varianza del error mediante matrices. Para ello (Uriel, 2013) Detalla que tomando en cuenta que $\hat{\mu} = y - \hat{y}$ y $\hat{\beta} = (X'X)^{-1}X'y$ se puede expresar:

$\hat{\mu} = y - X\hat{\beta}$ (3.4.36.)

$\hat{\mu} = y - X(X'X)^{-1}X'y$ (3.4.58.)

Al utilizar la matriz identidad para reducir la expresión, obtenemos:

$\hat{\mu} = [I - X(X'X)^{-1}X']y$ (3.4.59.)

$\hat{\mu} = My$ (3.4.60.)

De esta manera, en (3.4.60.) se expresa al vector de los residuos en función a la variable explicada y además *M* es una matriz idempotente. Pero podemos ir más allá y expresarla en función al vector de las perturbaciones.

$\hat{\mu} = [I - X(X'X)^{-1}X']y$ (3.4.59.)

$\hat{\mu} = [I - X(X'X)^{-1}X'](X\beta + \mu)$ (3.4.61.)

$\hat{\mu} = X\beta - X(X'X)^{-1}X'X\beta + \mu - X(X'X)^{-1}X'\mu$ (3.4.62.)

$\hat{\mu} = X\beta - X\beta + [I - X(X'X)^{-1}X']\mu$ (3.4.63.)

$\hat{\mu} = [I - X(X'X)^{-1}X']\mu$ (3.4.64.)

$\hat{\mu} = M\mu$ (3.4.65.)

Recordando que la suma cuadrática de los residuos es: $SCR = \hat{\mu}'\hat{\mu}$, entonces:

$\hat{\mu}'\mu = \mu'M'M\mu = \mu'M\mu$ (3.4.66.)

Ahora aplicando esperanzas:

$E(\hat{\mu}'\hat{\mu}) = E(\mu'M\mu)$ (3.4.66.)

Lo que se busca ahora es calcular la traza de la esperanza, por lo que:

$E(\hat{\mu}'\hat{\mu}) = trE(\mu'M\mu)$ (3.4.67.)

$E(\hat{\mu}'\hat{\mu}) = E(tr\mu'M\mu)$ (3.4.68.)

Al reordenar:

$E(\hat{\mu}'\hat{\mu}) = E(trM\mu\mu')$ (3.4.69.)

Al no ser M un vector aleatorio, se obtiene:

$E(\hat{\mu}'\hat{\mu}) = trME(\mu\mu')$ (3.4.70.)

En este punto se hará un paréntesis, ya que

$E(\mu\mu') = \sigma^2 I$ (3.4.71.)

(3.4.71.) supone el cumplimiento del supuesto de homocedasticidad. Esto es fácil de demostrar matricialmente. Teniendo:

$$E(\mu) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E(\mu_1) \\ E(\mu_2) \\ \vdots \\ E(\mu_n) \end{bmatrix} \text{(3.4.72.)}$$

Por lo que al pretender expresar la forma matricial de $E(\mu\mu')$ tenemos:

$$E(\mu\mu') = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n] \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = E \begin{bmatrix} \mu_1^2 & \mu_1\mu_2 & \cdots & \mu_1\mu_n \\ \mu_2\mu_1 & \mu_2^2 & \cdots & \mu_2\mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n\mu_1 & \mu_n\mu_2 & \cdots & \mu_n^2 \end{bmatrix} \text{(3.4.73.)}$$

Donde al aplicar las esperanzas a cada elemento del producto, tenemos:

$$E(\mu\mu') = \begin{bmatrix} E(\mu_1^2) & E(\mu_1\mu_2) & \cdots & E(\mu_1\mu_n) \\ E(\mu_2\mu_1) & E(\mu_2^2) & \cdots & E(\mu_2\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\mu_n\mu_1) & E(\mu_n\mu_2) & \cdots & E(\mu_n^2) \end{bmatrix} \text{(3.4.74.)}$$

Donde si recordamos los supuestos de homocedasticidad del término de perturbación y la ausencia de autocorrelación de las perturbaciones, representadas con: $E(\mu_i^2) = \sigma^2$ y $E(\mu_i\mu_j) = 0$ respectivamente, por lo que podemos reemplazar en (3.4.74.)

$$E(\mu\mu') = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \text{(3.4.75.)}$$

De esta manera se demuestra $E(\mu\mu') = \sigma^2 I$, por lo que terminando el paréntesis, se puede proseguir en (3.4.70.)

$E(\hat{\mu}'\hat{\mu}) = trME(\mu\mu')$ (3.4.70.)

$E(\hat{\mu}'\hat{\mu}) = trME\sigma^2 I$ (3.4.76.)

Finalmente mediante la propiedad de la traza de que $tr(AB) = tr(BA)$, entonces al tomar en cuenta que $M = [I - X(X'X)^{-1}X']$, deducimos:

$trM = tr[I_{nxn} - X(X'X)^{-1}X'] = trI_{nxn} - trX(X'X)^{-1}X' = trI_{nxn} - trI_{kxk} = n - k$ (3.4.77.)

Al reemplazar (3.4.77.) en (3.4.76.) obtenemos lo siguiente:

$E(\hat{\mu}'\hat{\mu}) = \sigma^2(n-k) \rightarrow \sigma^2 = \frac{E(\hat{\mu}'\hat{\mu})}{n-k}$ (3.4.78.)

De esta manera al ser $\hat{\sigma}^2$ un estimador insesgado de σ^2, finalizamos en:

$\hat{\sigma}^2 = \frac{\hat{\mu}'\hat{\mu}}{n-k}$ (3.4.79.)

Para demostrar su insesgadez, retomamos (3.4.79.) y aplicamos esperanzas:

$E(\hat{\sigma}^2) = E\left(\frac{\hat{\mu}'\hat{\mu}}{n-k}\right) = \frac{E(\hat{\mu}'\hat{\mu})}{n-k} = \frac{\sigma^2(n-k)}{n-k} = \sigma^2$ (3.4.80.)

3.4.3.3. Varianza y error estándar de los estimadores.

Una vez entendido la varianza del error y el error estándar de la regresión, podremos entender cómo hallar la varianza y el error estándar de los estimadores mediante MCO usando matrices. Una vez más todo parte de $\hat{\beta} = (X'X)^{-1}X'y$, donde:

$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mu$ (3.4.44.)

$\hat{\beta} = \beta + (X'X)^{-1}X'\mu$ (3.4.45.)

Donde al pasar a restar el vector de los estimadores con el vector de los parámetros, obtenemos:

$\hat{\beta} - \beta = (X'X)^{-1}X'\mu$ (3.4.81.)

Entonces al tener en cuenta que la matriz *var-cov* de los estimadores es

$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$ (3.4.82.)

Entonces reemplazamos:

$var - cov(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$ (3.4.83.)

$var - cov(\hat{\beta}) = E\{[(X'X)^{-1}X'\mu][(X'X)^{-1}X'\mu]'\}$ (3.4.84.)

$var - cov(\hat{\beta}) = E[(X'X)^{-1}X'\mu\mu'X(X'X)^{-1}]$ (3.4.85.)

$$var - cov(\hat{\beta}) = [(X'X)^{-1}X'E(\mu\mu')X(X'X)^{-1}] \text{ (3.4.86.)}$$

Al recordar el supuesto de homocedasticidad, el cual es $E(\mu\mu') = \sigma^2 I$, entonces:

$$var - cov(\hat{\beta}) = [(X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1}] \text{ (3.4.87.)}$$

Donde reduciendo términos, finalmente obtenemos la matriz *var-cov* de los estimadores.

$$var - cov(\hat{\beta}) = \sigma^2(X'X)^{-1} \text{ (3.4.88.)}$$

En su forma matricial se observa:

$$var - cov(\hat{\beta}) = \begin{bmatrix} var(\hat{\beta}_1) & cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ cov(\hat{\beta}_2, \hat{\beta}_1) & var(\hat{\beta}_2) & \cdots & cov(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_1) & cov(\hat{\beta}_k, \hat{\beta}_2) & \cdots & var(\hat{\beta}_k) \end{bmatrix} \text{ (3.4.89)}$$

Donde la diagonal de la matriz son las varianzas de los estimadores y los elementos fuera de la diagonal son las covarianzas de las variables explicativas. Debido a que los elementos de la matriz triangular inferior y superior se repiten, solo es necesario calcular uno de ellos para hallar las covarianzas. Finalmente, al igual que la varianza del error, al aplicar la raíz cuadrada a la varianza del estimador en (3.4.89.) obtenemos el error estándar del estimador.

3.4.4. Bondad de ajuste en el modelo de regresión simple y múltiple.

Recordando lo dicho anteriormente, es preferible que nuestro modelo estimado tenga estimadores con un error estándar de regresión lo más mínimo posible, sin embargo, esta no es la única medida para elegir modelos estimados. Una medida fundamental para elegir un modelo econométrico estimado mediante MCO es sin lugar a dudas la bondad de ajuste, de hecho, la bondad de ajuste puede ser el determinante que nos ayude a decidir cuál es el modelo econométrico más apropiado.

La siguiente cita expone la idea anterior.

> *"El interés del EER [error estándar de regresión] como indicador del grado de ajuste de un modelo de regresión disminuye cuando queremos comparar la bondad del ajuste de dos modelos que tienen una variable dependiente diferente. En tal caso, no es en absoluto cierto que el modelo con menor EER sea el modelo con mejor ajuste, de hecho, no podríamos afirmar nada al respecto, salvo que*

establezcamos alguna medida relativa de grado de ajuste, que es lo que hacemos en esta sección." (Novales, 1998)

Lo que (Novales, 1998) Plantea, es que si bien es cierto el EER es importante para determinar cuál modelo econométrico es mejor para explicar, no es un indicador necesariamente determinante que señala cual es el mejor modelo, lo que se busca en la econometría es que los estimadores que miden la influencia de la(s) variable(s) regresora(s) sean MELI.

Por ello, es que al momento de comprobar la bondad de ajuste de un modelo se está buscando medir cómo se ajusta el modelo con los datos observados. Esto quiere decir, en palabras de (Cid S., Mora C., & Valenzuela H., 1990) Que lo que se busca con la bondad de ajuste es determinar cuánto es la proporción de la variabilidad de la variable dependiente que está explicada por la(s) variable(s) regresora(s) con los datos usados para el modelo. Por ello es que se dice que la bondad de ajuste mide como los datos se ajustan con el modelo sin tomar en cuenta a los residuos, que son datos inobservables.

Sin embargo, previamente a la explicación del cálculo para hallar el valor del coeficiente de determinación, (Cid S., Mora C., & Valenzuela H., 1990) Exponen una diferencia sutil en el coeficiente de determinación en el modelo de regresión simple y múltiple. Cuando se trata de un modelo de regresión múltiple, el coeficiente de determinación pasa a ser conocido como el **coeficiente de determinación múltiple** y depende del número de variables explicativas, de tal forma que a medida que se le agreguen más variables explicativas al modelo, el coeficiente de determinación múltiple no decrece, por el contrario, aumentará. El coeficiente de determinación múltiple mide la proporción de la variación de la endógena provocada por las variables exógenas.

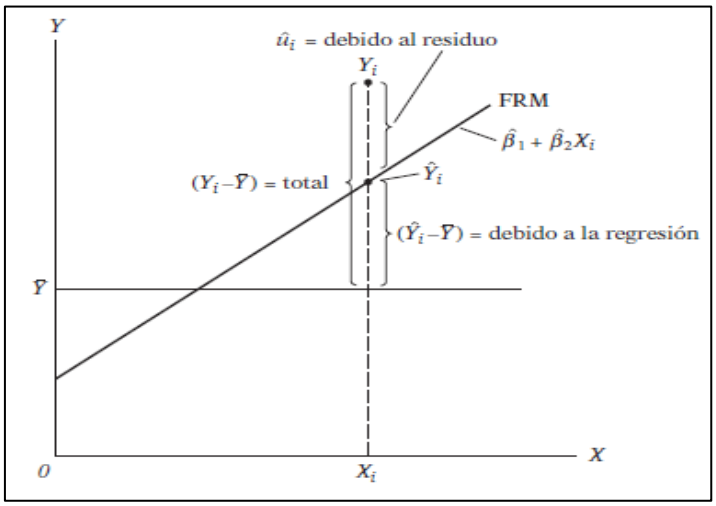
Veamos ahora cómo (Novales, 1998) expone la forma para deducir la fórmula que permite calcular el coeficiente de determinación.

$$Y_i - \overline{Y} = \left(\widehat{Y}_i - \overline{Y}\right) + \left(Y_i - \widehat{Y}_i\right) \text{ (3.4.90.)}$$

Donde al recordar que $\left(Y_i - \widehat{Y}_i\right) = \widehat{\mu}_i$ entonces

$$Y_i - \overline{Y} = \left(\widehat{Y}_i - \overline{Y}\right) + \widehat{\mu}_i \text{ (3.4.91.)}$$

Lo que (3.4.91.) expone se puede observar en el gráfico presentado a continuación, recogido de (Gujarati & Porter, 2010).



Gráfica 3.14. Partición de la varianza de Y_i en dos componentes.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

Visto de esta manera, se logra ver que existen dos diferencias más, aparte del ya explicado residuo representado con $(Y_i - \hat{Y}_i) = \hat{\mu}_i$, en donde lo que se pretende es medir la parte que varía de la variable endógena producto a la regresión, representado con $(\hat{Y}_i - \overline{Y})$. (Pérez L., 2012) Detalla los siguientes conceptos de sumatorias:

$$Suma\ Cuadrática\ Total = \sum(Y_i - \overline{Y})^2 \quad (3.4.92.)$$

$$Suma\ Cuadrática\ Explicada = \sum(\hat{Y}_i - \overline{Y})^2 \quad (3.4.93.)$$

$$Suma\ Cuadrática\ Residual = \sum(Y_i - \hat{Y}_i)^2 \quad (3.4.94.)$$

Donde:

$$\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum \hat{\mu}_i \quad (3.4.95.)$$

(Novales, 1998) Explica cómo se llega a (3.4.95.) a partir de (3.4.91.) Tomando en cuenta que al tener (3.4.91.) se debe elevar al cuadrado tenemos:

$$(Y_i - \overline{Y})^2 = [(\hat{Y}_i - \overline{Y}) + \hat{\mu}_i]^2 \quad (3.4.96.)$$

$$(Y_i - \overline{Y})^2 = (\hat{Y}_i - \overline{Y})^2 + 2(\hat{Y}_i - \overline{Y})\hat{\mu}_i + \hat{\mu}_i^2 \quad (3.4.97.)$$

Ahora sumaremos para toda la muestra:

$$\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + 2\sum(\hat{Y}_i - \overline{Y})\hat{\mu}_i + \sum \hat{\mu}_i^2 \quad (3.4.98.)$$

Hagamos un breve paréntesis, al tener en cuenta que $\sum(\hat{Y}_i - \overline{Y})\hat{\mu}_i = \sum \hat{\mu}_i \hat{Y}_i - \overline{Y}\sum \hat{\mu}_i$ y recordar que $\sum \hat{\mu}_i = 0$, entonces solamente nos queda: $\sum \hat{\mu}_i \hat{Y}_i =$

$\sum \widehat{\mu_i}(\widehat{\beta_0} + \widehat{\beta_1}X_1) = \widehat{\beta_0}\sum \widehat{\mu_i} + \widehat{\beta_1}\sum \widehat{\mu_i}X_1$, sin embargo ya se ha planteado que $\sum \widehat{\mu_i} = 0$ y $\sum \widehat{\mu_i}X_1 = 0$, entonces reemplazamos y al final obtenemos: $\widehat{\beta_0}(0) + \widehat{\beta_1}(0) = 0$, en consecuencia resolvemos en:

$$\sum(Y_i - \overline{Y})^2 = \sum(\widehat{Y_i} - \overline{Y})^2 + \sum \widehat{\mu_i} \quad (3.4.95.)$$

Llegado a este punto, es fácil deducir la forma que permite calcular el coeficiente de regresión. Debido a que lo que se intenta medir es la proporción explicada de la variabilidad de la endógena por el modelo de regresión entonces podemos finalmente entender la fórmula para hallar el coeficiente de determinación.

$$R^2 = \frac{SCE}{SCT} \; o \; R^2 = 1 - \frac{SCR}{SCT} \quad (3.4.96.)$$

Cualquiera de las dos formas que (3.4.96.) expone es válida para hallar el coeficiente de determinación representado con R^2. Tendrá un valor mayor a 0 y menor a 1, y mientras más cercano de 1 se encuentre, entonces sería mejor para el modelo, puesto que la endógena sería explicada enormemente por el modelo especificado. Finalmente, esta medida de bondad de ajuste tiende a usarse con más importancia cuando se trata de una regresión múltiple, para entender el motivo se presenta a continuación algunas consideraciones que (Uriel & Aldás, 2005) Detallan para su interpretación:

- Como ya se mencionó, cuando se agregan variables explicativas al modelo, el coeficiente de determinación aumenta, sin embargo, esto ocurre aunque no exista una relación con la variable endógena.
- Cuando el modelo no tiene un intercepto, el coeficiente de determinación no tiene una interpretación, ya que $\sum(Y_i - \overline{Y})^2 = \sum(\widehat{Y_i} - \overline{Y})^2 + \sum \widehat{\mu_i}$ no se cumple generando que se puedan calcular coeficientes de determinación que no corresponde al intervalo [0,1], es decir menor a 0 y mayor a 1.
- Cuando se trata de un modelo de datos de series temporales, el coeficiente de determinación suele ser elevado aunque no exista una relación causal.
- El coeficiente de determinación no es comparable cuando se trata de elegir cuál es la forma funcional más eficiente para explicar al modelo.

Aunque lo ideal sería que el coeficiente de determinación sea lo más cercano a 1 posible, en algunos casos tener un coeficiente de determinación tan elevado puede ser producto de errores en la especificación del modelo, debido a que el número de regresoras

puede hacer aumentar el valor del coeficiente de determinación sin que necesariamente exista una relación causal con la variable endógena. Por lo tanto, tener el coeficiente de determinación tan elevado cuando se tienen pocas regresoras, debería ser tratado más como una sospecha que el modelo presenta algún sesgo que como un acierto cuando se busca modelar correctamente. La siguiente cita expone lo dicho anteriormente.

> *"Por otra parte, la adición de nuevas variables en el modelo, nunca significará una disminución en el valor de R^2, debido a que el valor de SCR nunca aumentará con la adición de nuevas variables independientes y SCT es siempre el mismo para un conjunto dado de respuestas. Por esta razón se sugiere, a veces, que una medida modificada de R^2 se emplee en lugar de la ya descrita, de modo tal que ella sea sensible al número de variables en el modelo, esta medida se llama* ***Coeficiente de determinación múltiple ajustado****, se denota por R_a^2 (...)"* (Cid S., Mora C., & Valenzuela H., 1990)

Lo que la cita anterior intenta explicar es que el coeficiente de determinación no toma en cuenta si se agregan variables al modelo, sobre todo si son relevantes para el modelo o no, por ello, con el coeficiente de determinación ajustado se pretende incluir las variables para calcularlo. De tal forma que se halla con:

$$R_a^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)} \quad (3.4.97.)$$

En (3.4.97.) al igual que para hallar la varianza del error, se divide entre sus grados de libertad, donde el denominador tiene *k* que significa el número de variables dependientes. (Uriel & Aldás, 2005) Detallan algunas consideraciones para interpretar al coeficiente de determinación ajustado, que también puede denotarse como $\overline{R^2}$.

- A diferencia del R^2, el $\overline{R^2}$ puede tomar valores negativos cuando el ajuste del modelo sea muy malo y al igual que R^2, cuando se acerca a 1 entonces tiene una excelente bondad de ajuste.
- Debido a que toma en cuenta la relevancia de las variables regresoras, **al momento de incluir una nueva variable y esta medida aumenta, entonces se ha logrado acertar en la correcta especificación del modelo**, caso contrario sucede si en vez de aumentar, disminuye. Por esto, es que en algunos libros se puede encontrar que el $\overline{R^2}$ puede comparar modelos con diferentes números de regresoras.

- Similar al R^2, no tiene una interpretación clara cuando no se toma en cuenta al intercepto.
- Lo mismo sucede con los modelos de datos temporales, cuando se estiman estos el $\overline{R^2}$ suele ser elevado. Pero claro que nunca podrá superar al R^2. La razón de esto, es que las series temporales presentan componentes los cuales son: la tendencia, el componente cíclico, el componente estacionario y el componente irregular. Los cuales tienen una enorme influencia sobre el comportamiento de las variables en el tiempo.
- Finalmente, tampoco se debería usar para comparar distintas formas funcionales.

3.4.5. Tabla ANOVA.

Finalmente, se presenta la tabla ANOVA, como último tema para entender la estimación de los modelos de regresión lineal clásicos. Esta tabla muestra un análisis de la varianza de la variable dependiente mostrando cuáles son los factores más influyentes. Sin embargo, por lo general solo se muestran tres partes: **Sumas Cuadráticas, Grados de Libertad y Medias Cuadráticas**. Ya que en algunas tablas no solo se muestran estos tres componentes sino también los factores que influyen de cada variable del modelo de regresión sobre la variable dependiente. A continuación, se muestra la tabla y sus componentes.

Fuente de variación	**Suma de Cuadrados**	**Grados de Libertad**	**Media Cuadrática**
Regresión	$SCE = \sum(\widehat{Y}_i - \overline{Y})^2$	$k - 1$	$MCE = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{k - 1}$
Residual	$SCR = \sum(Y_i - \widehat{Y}_i)^2$	$n - k$	$MCR = \frac{\sum(Y_i - \widehat{Y}_i)^2}{n - k - 1}$
Total	$SCT = \sum(Y_i - \overline{Y})^2$	$n - 1$	$MCT = \frac{\sum(Y_i - \overline{Y})^2}{n - 1}$

Tabla 3.13. Tabla ANOVA.
Elaboración (Uriel & Aldás, 2005)
Fuente (Uriel & Aldás, 2005)

Si recordamos que la varianza del error es calculado mediante$\frac{\sum(Y_i-\hat{Y}_i)^2}{n-k-1}$, podemos inferir que la **Media Cuadrática Residual es sinónimo de la varianza del error**. La tabla 3.13. También puede ser escrita en su forma matricial.

Fuente de variación	**Suma de Cuadrados**	**Grados de Libertad**	**Media Cuadrática**
Regresión	$SCE = \hat{\beta}'X'Y - n\overline{Y}^2$	$k-1$	$MCE = \frac{\hat{\beta}'X'Y - n\overline{Y}^2}{k-1}$
Residual	$SCR = Y'Y - \hat{\beta}'X'Y$	$n-k$	$MCR = \frac{Y'Y - \hat{\beta}'X'Y}{n-k-1}$
Total	$SCT = Y'Y - n\overline{Y}$	$n-1$	$MCT = \frac{Y'Y - n\overline{Y}}{n-1}$

Tabla 3.14. Tabla ANOVA en su forma matricial.
Elaboración (Cid S., Mora C., & Valenzuela H., 1990)
Fuente (Cid S., Mora C., & Valenzuela H., 1990)

Recuerde: *n* es el número de observaciones que emplea el modelo estimado, *k* es el numero estimadores a estimar incluido el intercepto según lo expuesto ya anteriormente.

3.5. Inferencia del Modelo por Mínimos Cuadrados Ordinarios

Hasta este punto, se ha intentado explicar la estimación de los estimadores mediante MCO, pero la elaboración de los modelos econométricos no solo consiste en estimar los estimadores puntuales de la regresión muestral, sino también demostrar que verdaderamente existe una relación causal y no es producto de la casualidad, es decir demostrar que el modelo empleado y las variables que lo conforman son válidas para explicar el comportamiento de la variable endógena. En términos más propios, demostrar la **significancia estadística** mediante las llamadas **prueba de hipótesis o intervalos de confianza**.

3.5.1. Significancia individual.

Cuando se tiene una regresión simple o múltiple, y se pretende demostrar la existencia de la relación causal de la variable explicada con una variable explicativa, se está tratando de demostrar que la variable explicativa tiene un coeficiente que es

significativo. La prueba de hipótesis empleada estaría conformado por una hipótesis nula y una hipótesis alternativa, las cuales representan una prueba sobre los **parámetros poblacionales.** Para realizar la prueba de hipótesis sobre su significancia toma la siguiente estructura:

$H_0: \beta_k = 0$ (3.5.1.)

$H_1: \beta_k \neq 0$ (3.5.2.)

(Lind, Marchal, & Wathen, 2015) Analiza la importancia de esta demostración de la significancia, al comprobar que el coeficiente es distinto de cero entonces el modelo aumenta su capacidad predictiva ya que la variable al lado del estimador significativo debe ser incluido en el modelo, caso contrario ocurre con aquellas variables que tienen estimadores no significativos ya que al ser igual a cero son descartadas del modelo de regresión y por lo tanto el modelo pierde capacidad predictiva. En caso de los modelos de regresión simple, si la pendiente no es significativa, entonces al ser descartado la variable independiente, la media de la variable dependiente se usará como factor de predicción.

Por lo tanto, lo que se busca es que se rechace la hipótesis nula y no se rechace la hipótesis alternativa, ya que así comprobamos que los estimadores son significativos.

Para realizar el contraste de hipótesis de significancia individual, el cual es un sinónimo de la prueba de hipótesis, se realizará con la distribución **t de Student.** Con esta distribución se usará los **estadísticos t calculados y t tabulados**. La **regla de aceptación**, es decir para decidir si rechazar la hipótesis nula y asumir la existencia de la significancia estadística individuales, es que el estadístico t calculado debe ser mayor al estadístico t tabulado.

Para hallar el estadístico *t* calculado se utiliza la fórmula:

$$t_{calculado} = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \text{ (3.5.3.)}$$

Donde:

$\hat{\beta}_k$: Es el estimador muestral del parámetro poblacional que se pretende testear.

β_k: Es el parámetro poblacional que se quiere testear.

$\hat{\sigma}_{\hat{\beta}_k}$: Es el error estándar del estimador muestral.

Ya que se pretende testear la significancia del parámetro poblacional, y además $H_0: \beta_k = 0$ entonces al reemplazar en (3.5.3.) se convierte en:

$$t_{calculado} = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \text{ (3.5.4.)}$$

Para hallar el estadístico t tabulado, también llamado valor crítico, es necesario tomar en cuenta algunos aspectos: **el nivel de significancia, los grados de libertad** y **el número de colas en la prueba de hipótesis.**

- **Nivel de significancia**

Para entender su significado, es necesario mostrar una tabla que muestra los tipos de errores que se pueden cometer en la prueba de hipótesis.

	Investigador	
Hipótesis nula		
	No rechaza H_0	Rechaza H_0
H_0 es verdadera	**Decisión correcta**	**Error tipo I**

Tabla 3.15. Tipo de error.
Elaboración (Lind, Marchal, & Wathen, 2015)
Fuente (Lind, Marchal, & Wathen, 2015)

Es muy común que cuando se pretende hacer una prueba de hipótesis, encontrarnos la existencia de algún factor que ponga en evidencia que la prueba ha estado equivocada. Por ello, es que siempre se tiene en cuenta una probabilidad de que se ha cometido un error, ya sea rechazar una hipótesis nula verdadera o aceptar una hipótesis falsa. Teniendo esto en cuenta, podemos definir al nivel de significancia como **la probabilidad de haber cometido error tipo I, es decir, la probabilidad de rechazar una hipótesis nula verdadera.**

En este caso, cuando se contraste la significancia individual de los parámetros, básicamente el error tipo I seria rechazar que el parámetro poblacional no es significativo cuando verdaderamente no lo es. En términos menos confusos, **el error tipo I en este tipo de contrastes de hipótesis es aceptar que el parámetro es significativo cuando**

en realidad no lo es. El nivel de significancia está representado por lo general con α. A demás no existe un valor establecido, sino que el mismo investigador debe elegir el nivel de significancia de manera subjetiva, pero por lo general el valor más empleado por la mayoría de programas estadísticos y econométricos y en investigaciones es el 0.05 o 5%. En STATA se emplea un nivel de significancia del 5% por defecto, pero en algunos comandos se encuentra que usa además el 1% y 10%. (Lind, Marchal, & Wathen, 2015) Expande esta idea, señalando que por lo general se usa el 5% cuando se trata de investigaciones más aun referidas al consumidor mientras que se recomienda usar el 1% cuando se trate de control de calidad y el 10% si se quiere realizar encuestas políticas.

Como último dato, en la teoría estadística, la probabilidad de cometer error tipo II se denota con β, pero no es usado en este tipo de contrastes.

- **Grado de libertad**

Anteriormente ya se definió los grados de libertad, y son los mismos que se usan para dividir las sumas cuadráticas y obtener las medias cuadráticas.

Siendo el número de observaciones de la muestra menos el número de estimadores a estimar: $n - k$ incluido el intercepto. Esto se cumple tanto para las regresiones simples como para las regresiones múltiples.

- **Número de colas**

Para explicar a qué se refiere con colas en las pruebas de hipótesis, veamos la siguiente figura.

Pruebas de hipótesis

Gráfica de t de Student.

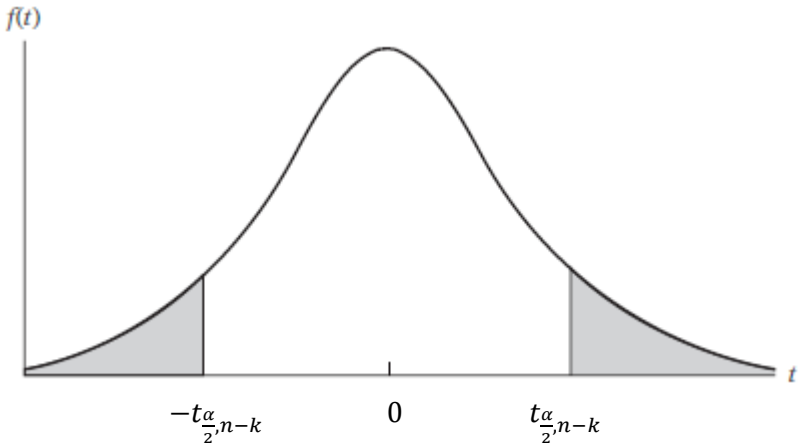
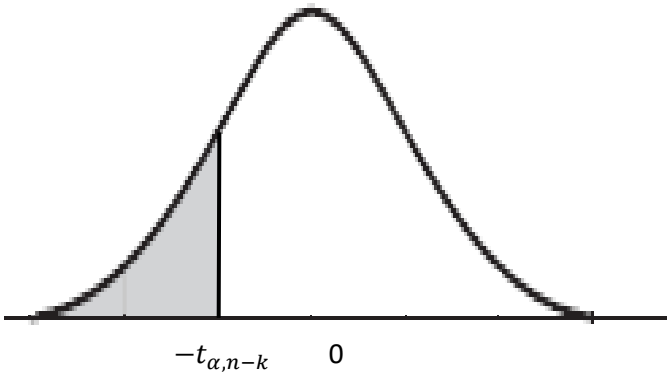
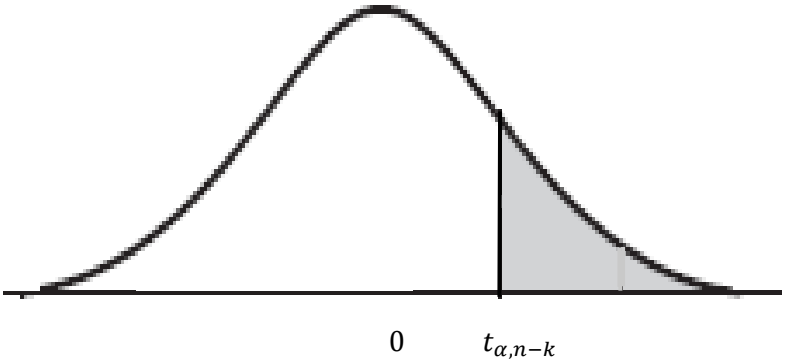
$H_0: \beta_k = 0$ $H_1: \beta_k \neq 0$ **Bilateral o dos colas**	
$H_0: \beta_k = 0$ $H_1: \beta_k < 0$ **Unilateral, una cola a la izquierda**	
$H_0: \beta_k = 0$ $H_1: \beta_k > 0$ **Unilateral, una cola a la derecha**	

Tabla 3.16. Gráficas de *t* de student.
Elaboración (Gujarati & Porter, 2010)
Fuente (Gujarati & Porter, 2010)

El número de colas en estadística, hace referencia al número de regiones de rechazo que puede tener un gráfico de distribución al momento de querer realizar un contraste de hipótesis. En el caso de un contraste de significancia, la hipótesis nula siempre será el parámetro poblacional igual a cero, mientras que la hipótesis alternativa

es que el mismo parámetro poblacional es distinto a cero. Es esta diferencia lo que provoca que se tomen en cuenta una prueba de hipótesis de dos colas o bilateral. Sin embargo, en algunos trabajos e investigaciones no basta con que sea distinto de cero, sino que tenga un signo esperado y que deberá cumplirse. (Novales, 1998) Explica que cuando se quiere hallar un estadístico t tabulado se toma en cuenta el número colas partiendo de la teoría económica o la teoría que se este empleado. Para explicarlo, imagine el siguiente modelo: $Y = \beta_1 + \beta_2 X + \mu$, donde según la teoría empleada y las evidencias, el signo esperado que debe cumplir β_2 es positivo, por lo tanto cuando se quiere contrastar la significancia mediante pruebas de hipótesis la estructura de la hipótesis nula es: $H_0: \beta_k = 0$ mientras que la hipótesis alternativa ya no llevaría el signo de la diferencia sino tendría que ser: $H_1: \beta_k > 0$, por lo tanto el número de colas sería unilateral a la izquierda, sin embargo, de no tener especificado el signo esperado de β_2, entonces al comprobar la significancia la hipótesis alternativa seria: $H_1: \beta_k \neq 0$ y de esta manera se haría uso de dos colas. Cuando se quiere denotar cómo se distribuye el estadístico t calculado con el estadístico t tabulada se expondría: $tc \sim t_{\frac{\alpha}{2},gl}$ cuando se usa una prueba de hipótesis con dos colas y $tc \sim t_{\alpha,gl}$ cuando es una cola.

De esta manera se podrá buscar en la tabla del estadístico *t* de Student los valores críticos usando el nivel de significancia y de grados de libertad. Así, podremos determinar si aceptar o rechazar la hipótesis nula, siguiendo la regla de decisión: si el estadístico *t* calculado es menor al estadístico *t* tabulado o crítico entonces no se rechaza a la hipótesis nula y se asumirá que el estimador no tiene significancia individual y por lo tanto la variable que lo acompaña debería ser descartada del modelo.

3.5.1.1. Estimación por intervalos.

Tal como su nombre indica, ahora se construirá intervalos de confianza para estimar el valor de los estimadores con el uso de probabilidades, la idea surge debido a que se tiene que utilizar una muestra para estimar valores desconocidos poblacionales. Tal como explican (Gujarati & Porter, 2010) La estimación puntual, que es la se ha venido explicando hasta este punto, no puede ser tomada como fiable en su totalidad y esa desconfianza se crea porque las muestras que pueden ser usadas para estimar el mismo modelo poblacional son diferentes entre sí; por lo que conocer cuáles son los valores del intervalo no solo es fundamental sino también necesario para comprender más sobre se relación entre las variables.

Los estimadores tienen errores estándares, y siguiendo con lo expuesto anteriormente, los errores estándares miden la fiabilidad de los estimadores, y se espera a que estos sean lo menor posible. Por lo tanto, para construir intervalos de confianza será necesario hacer uso de los errores estándares de los estimadores. Siguiendo la fórmula:

$$\Pr\left[\hat{\beta}_k - t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)\right] = 1 - \alpha \text{ (3.5.5.)}$$

La fórmula (3.5.5.) hace uso de algunos elementos que ya se han visto anteriormente: $\boldsymbol{1-\alpha}$ es el nivel de confianza y $\boldsymbol{\alpha}$ es el nivel de significancia y tal como se dijo anteriormente es la probabilidad de cometer error tipo I y como consecuencia que se escoge de manera arbitraria debido al tipo de investigación o ya sea porque el programa estadístico usa cierto nivel de significancia, el nivel de confianza también es escogida de forma arbitraria, por lo general se escoge el 95% de nivel de confianza y un 5% de significancia. También se hacen presentes el estadístico *t* tabulado o crítico hallado desde la tabla del estadístico t de Student y los errores estándares que provienen de la raíz al cuadrado de sus varianzas. (Gujarati & Porter, 2010) Explican que la fórmula anterior se puede hallar mediante el uso de la distribución *t* para construir los intervalos de confianza. De esta manera se tiene

$$\Pr\left[-t_{\frac{\alpha}{2},n-k} \leq t \leq t_{\frac{\alpha}{2},n-k}\right] = 1 - \alpha \text{ (3.5.6.)}$$

El valor del centro corresponde al estadístico t calculado, por lo que al reemplazarse en (3.5.6.) se obtiene:

$$\Pr\left[-t_{\frac{\alpha}{2},n-k} \leq \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \leq t_{\frac{\alpha}{2},n-k}\right] = 1 - \alpha \text{ (3.5.7.)}$$

El símbolo $\hat{\sigma}_{\hat{\beta}_k}$ corresponde al error estándar del estimador $\hat{\beta}_k$ pero puede ser confundido como el símbolo $\hat{\sigma}$ que es el estimador del error estándar de la regresión, dos conceptos parecidos pero distintos tal como ya se explicó anteriormente, por lo que para evitar alguna confusión el símbolo $\hat{\sigma}_{\hat{\beta}_k}$ se reemplazará por $ee(\hat{\beta}_k)$ para referirse al error estándar del estimador $\hat{\beta}_k$. Volviendo al tema central, si reorganizamos (3.5.7.) se obtiene

$$\Pr\left[\hat{\beta}_k - t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k) \leq \beta_k \leq \hat{\beta}_k + t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)\right] = 1 - \alpha \text{ (3.5.5.)}$$

Sin embargo, este intervalo es un intervalo fijo y no aleatorio, (Gujarati & Porter, 2010) Definen que al ser un valor desconocido el valor del parámetro poblacional, se tiene que hacer uso de un estimador muestral, por lo que el parámetro poblacional se convierte en un valor fijo que puede estar o no en el intervalo construido, por ello es que para interpretarse se sigue la siguiente sintaxis, por ejemplo si utilizamos un 5% de significancia entonces interpretamos como: la probabilidad de construir un intervalo que contenga el valor verdadero del parámetro poblacional en un intervalo desde $\hat{\beta}_k - t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$ hasta $\hat{\beta}_k + t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$ es del 95%. Esto es muy distinto a decir que la probabilidad de que el valor verdadero este incluido en un intervalo desde $\hat{\beta}_k - t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$ hasta $\hat{\beta}_k + t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$ sea del 95%, ya que la primera hace referencia a la **probabilidad de construir** un intervalo que contenga el valor verdadero del parámetro poblacional mientras que la segunda haría referencia a la probabilidad que el verdadero valor del estimador esté en el intervalo construido. **Por lo que al hacer uso de un intervalo fijo se debería interpretar de la primera forma**, es decir a la probabilidad de construir un intervalo que contenga el valor verdadero es del 95%. Otra forma de interpretar un intervalo fijo considerando lo anteriormente dicho sería: En 95 de 100 intervalos construidos el verdadero valor estará contenido en un intervalo desde $\hat{\beta}_k - t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$ hasta $\hat{\beta}_k + t_{\frac{\alpha}{2},n-k} * ee(\hat{\beta}_k)$.

Finalmente, cuando se quiere comprobar si el estimador tiene significancia individual también resulta útil revisar el intervalo construido, si el valor 0 no se encuentra en el intervalo construido, entonces se puede rechazar la hipótesis nula y asumir que el estimador es significativo.

Entonces, cuando se tiene que $tc > t_{\frac{\alpha}{2},k-1}$ se rechaza la hipótesis nula y se asume que el estimador contrastado tiene significancia individual y debe ir en el modelo. Mientras que $tc < t_{\frac{\alpha}{2},k-1}$ entonces no se rechaza la hipótesis nula y el estimador contrastado no tiene significancia individual y podría considerarse ser descartado del modelo ya que la variable independiente realmente no explica a la variable endógena.

3.5.2. Significancia global.

Anteriormente se había explicado, que la importancia que una variable regresora tenga un estimador con significancia individual radica en que permite a la variable regresora ser tomada en cuenta para estar en el modelo ya que se ha demostrado que explica a la variable endógena. Sin embargo, la significancia no sólo se concentra en la individualidad de cada variable, sino también en verificar que el conjunto de todas las variables especificadas tiene significancia. Es decir, también se debe considerar si el modelo especificado para explicar las variaciones de la variable endógena tiene significancia. En palabras de (Court & Rengifo, 2011) Lo que se quiere verificar es que exista significancia global en el modelo para determinar si el modelo realmente puede ser usado para explicar la variabilidad de la variable endógena.

Y al igual que en la significancia individual, también se hará uso de la prueba de hipótesis para verificar la existencia o no de la significancia global en un modelo estimado. Para entender esta sección se debe revisar el análisis de la varianza del modelo, siendo más precisos se debe revisar la tabla ANOVA. A continuación se reproduce la tabla que anteriormente ya se había mostrado.

Fuente de variación	**Suma de Cuadrados**	**Grados de Libertad**	**Media Cuadrática**
Regresión	$SCE = \sum(\widehat{Y}_i - \overline{Y})^2$	$k - 1$	$MCE = \frac{\sum(\widehat{Y}_i - \overline{Y})^2}{k - 1}$
Residual	$SCR = \sum(Y_i - \widehat{Y}_i)^2$	$n - k$	$MCR = \frac{\sum(Y_i - \widehat{Y}_i)^2}{n - k - 1}$
Total	$SCT = \sum(Y_i - \overline{Y})^2$	$n - 1$	$MCT = \frac{\sum(Y_i - \overline{Y})^2}{n - 1}$

Tabla 3.13. Tabla ANOVA.
Elaboración (Uriel & Aldás, 2005)
Fuente (Uriel & Aldás, 2005)

Usando esta tabla se espera contrastar una prueba de hipótesis, (Hanke & Wichern, 2006) señalan cuál sería:

$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$ (3.5.6.)

$H_1: por\ lo\ menos\ una\ \beta_k\ es\ distinto\ a\ 0$ (3.5.7.)

(Court & Rengifo, 2011) También muestra otra forma de representar esta hipótesis nula:

H_0: $No\ existe\ una\ relación\ lineal\ entre\ Y\ con\ todas\ las\ variables\ exógenas$. (3.5.8.)

H_1: $Existe\ una\ relación\ lineal\ entre\ Y\ con\ todas\ las\ variables\ exógenas$. (3.5.9.)

De cualquier forma, se debe tener en cuenta que en la hipótesis (3.5.6.) no se toma en cuenta el estimador del intercepto, es decir no se incluye β_1. Para contrastar esta hipótesis, se hará uso del estadístico *F* calculado el cual tiene una distribución *F* de Fisher, se puede denotar al estadístico *F* tabulado como $F_{\alpha,k-1,n-k}$. La distribución que sigue el estadístico *F* calculado se representa:

$Fc \sim F_{\alpha,k-1,n-k}$ (3.5.10.)

En algunos libros de econometría se encuentra simplemente: $F \sim F_{\alpha,k-1,n-k}$ pero se hará de (3.5.10) para evitar confusiones con el estadístico *F* tabulado, también llamado valor crítico al igual que la distribución *t* de Student que ya se explicó en la sección anterior. El estadístico *F* calculado se puede hallar utilizando la tabla ANOVA o el coeficiente de determinación. Estas son las dos formas para hallarlo:

Con la tabla ANOVA	Con el Coeficiente de Determinación
$Fc = \frac{MCE}{MCR} = \frac{\frac{SCE}{k-1}}{\frac{SCR}{n-k}}$	$Fc = \frac{\frac{R^2}{(k-1)}}{\frac{1-R^2}{(n-k)}}$

Tabla 3.17. F calculado.
Elaboración propia
Fuente (Gujarati & Porter, 2010)

(Gujarati & Porter, 2010) Ponen al descubierto la relación entre el estadístico *F* calculado y el coeficiente de determinación, por lo que ambas tienen un tipo de relación directa, cuando una incrementa su valor la otra también. Si R^2 es igual a 1 entonces el estadístico *F* calculado tiende hacia el infinito.

Por otro lado, al igual que para hallar los valores críticos del *t* tabulado se debe revisar la tabla *t* de Student, para hallar el valor crítico del estadístico *F* tabulado se debe revisar la tabla *F* de Fisher, la cual se hace uso del nivel de significancia y de los grados de libertad. Algo en que difieren ambas distribuciones es que la distribución *F* siempre

será de una cola, por lo general se escoge la izquierda. Y la **regla de decisión** es la misma, **si: $Fc > F_{\alpha,k-1,n-k}$ se rechaza la hipótesis nula y se asume que el modelo es significativo para explicar a la variable endógena, caso contrario sucede si $Fc < F_{\alpha,k-1,n-k}$ donde no se puede rechazar la hipótesis nula y se asume que el modelo no puede ser empleado para explicar a la variable endógena.**

Ya sea en el modelo simple o múltiple, la significancia global usando la prueba *F* es la misma, y ambas tienen validez al momento de contrastarse.

Otra forma muy útil al momento de contrastar hipótesis de significancia ya sea individual o global, es utilizando el ***valor p***, el cual es la probabilidad de que la hipótesis nula sea verdadera y esta se compara con el nivel significancia, que por lo general toma el valor de 5%. Los programas estadísticos calculan el *valor p* de cada una de los estimadores y de la significancia global cuando se ejecuta el comando respectivo. De esta manera cuando $valor\ p < 0.05$ entonces la hipótesis nula sea de significancia individual o global será rechazada y se asume que la variable o el modelo son significativos, según la prueba de hipótesis contrastada. Recordemos que el nivel de significancia puede ser de 1% a 10% por lo que dado un nivel de significancia el *valor p* puede o no mostrar la existencia de significancia individual o global. El *valor p* puede tomar valores desde 0 a 1, donde lo que se prefiere es que sea lo más cercano a 0 posible, de esta forma se podrá inferir que existe significancia. Llegado a este punto solo queda comentar sobre ¿Qué pasaría si alguna variable no tiene significancia individual o si el modelo carece de significancia global? Una pregunta válida tomando en cuenta que ambos tipos de significancia corresponden a la significancia estadística.

Empecemos con la significancia estadística individual, tomando en cuenta que las variables se toman en cuenta para construir el modelo a partir de una teoría económica que las respalda e indica cómo es el comportamiento que uno esperaría que tengan las variables regresoras para que sean estadísticamente significativas, pero no siempre ocurre así, de hecho (Wooldrige, 2009) Menciona a la **significancia económica** como otro aspecto importante al momento de revisar el modelo. Básicamente, la significancia económica es descrita como la magnitud que tiene la relación de una variable explicativa con la explicada expresada con los estimadores, más específicamente el signo esperado de los estimadores. En algunos modelos econométricos, el signo esperado de los estimadores puede ser un indicio que la variable es o no significativa en el modelo para

explicar a la variable endógena. (Wooldrige, 2009) También advierte que, si bien la significancia estadística es importante, no debe ser tomada en cuenta dejando de lado a la significancia económica para verificar o no si el modelo está correctamente especificado, ya que haciendo esto se podría llegar a conclusiones equivocadas. Por lo tanto, el uso de una muestra grande y de niveles de significancia menores hace que ambas significancias coincidan. Lo ideal sería que el modelo tenga variables significativas, que el estimador cumpla con el signo esperado y la magnitud sea lo suficientemente grande como para decir que es muy influyente. En el caso que la muestra sea grande y la variable no tiene significancia, pero si cumple con el signo esperado y además tiene un efecto grande medido por el estimador entonces podría aceptarse en el modelo. Por el contrario, si tuviese una muestra pequeña entonces debería considerarse aumentar la muestra para obtener estimaciones seguras. Si, por el contrario, la variable es significativa pero no tiene ni el signo esperado ni un efecto grande, entonces podría resolverse revisando la especificación del modelo, y es que esto por lo general es ocasionado por un error de especificación, quizá una variable ha sido omitida o una variable no debería estar ahí.

Si se tiene la sospecha que una variable debe ser sacada del modelo, la significancia global podría ser de ayuda para esto; ya que cuando se acepta la hipótesis nula de la prueba de hipótesis sobre la significancia global, lo aconsejable es volver a especificar el modelo, con otras o quitando algunas variables.

En conclusión, el investigador deberá tener criterio y deberá seguir el juicio de su investigación para considerar una variable que no tiene significancia individual estadística en el modelo debido a que la teoría económica puede ser más fuerte cuando se quiere comprobar esto. Caso contrario sucede cuando el modelo no tiene significancia global estadística, es mejor considerar replantear el modelo completamente.

3.6. Diagnósticos y Corrección de Violación de los Supuestos de la Estimación mediante Mínimos Cuadrados Ordinarios

Cuando se estima mediante MCO se espera a que la estimación cumpla con los supuestos debido a que el incumplimiento de los supuestos de MCO ocasiona que los estimadores dejan de ser MELI conduciéndonos a resultados equivocados. Por lo tanto, para estar seguros que los estimadores son los correctos entonces se debe evaluar si el modelo cumple con los supuestos establecidos.

Aunque no es el tema principal de la presenta guía de estudios, se presentara algunos métodos para corregir y detectar si el modelo especificado tiene o no alguna violación en los supuestos de MCO y los métodos empleados serán ejemplificados usando los comandos de STATA.

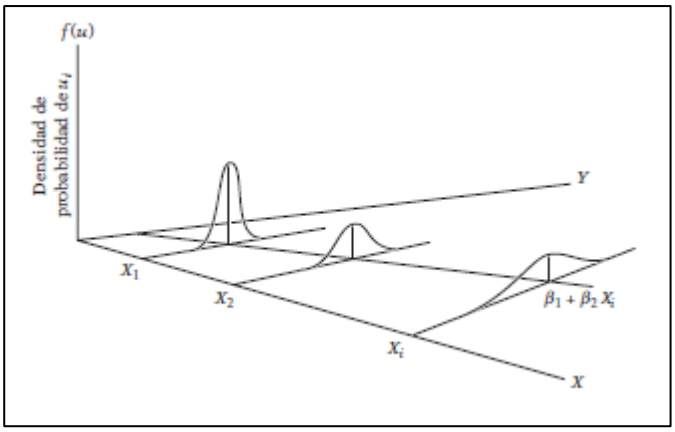
3.6.1. Test de detección y métodos correctivos de heterocedasticidad.

Hagamos brevemente un repaso de la naturaleza de la estimación bajo heterocedasticidad, sus causas y consecuencias.

La heterocedasticidad es la violación del supuesto de homocedasticidad, el cual indica que la varianza del término de error deja de ser constante para cada valor de la variable independiente. Tal como anteriormente se indicó:

$var(\mu_i|X_i) = \sigma_i^2$(3.3.10.)

La expresión (3.3.10.) tiene el subíndice *i* para cada valor de μ dado su respectivo valor de la variable independiente, *X*, demostrando que no tiene la misma varianza para otro valor de μ, lo anterior se puede expresar con el grafico que ya se expuso anteriormente:



Gráfica 3.7. Varianza no constante.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

Observe como la curva de distribución en *X1* es más alta que *X2*, cuando algo así sucede es porque la estimación no tiene una varianza constante y debido a esto es que la varianza condicional de *Y* dado *X*, $var(Y|X)$ tampoco es constante. El problema fundamental de la heterocedasticidad es que los estimadores ya no tienen varianza mínima, ya que la homocedasticidad no influye en el momento de estimar los estimadores.

Sin embargo, las pruebas de significancia, el coeficiente de determinación y los errores estándares de los estimadores ya no tienen sentido de interpretación y podrían llevarnos a conclusiones falsas sobre el modelo. Estas son las principales consecuencias según la tabla 3.6. Se extrae la sección que habla de la heterocedasticidad y se mostrará a continuación:

Heterocedasticidad	• Los estimadores del modelo conservan su insesgamiento, sin embargo dejan de ser eficientes, por lo tanto el estimador por MCO ya no tiene varianza mínima haciendo que los estimadores ya no sean MELI. Al perder la eficiencia de los estimadores ya no es posible estimar mediante MCO.
	• Muy diferente a los estimadores, la **varianza del error** estimada del modelo se vuelve una varianza sesgada, esto quiere decir que la varianza del error estimada del modelo es diferente a la varianza poblacional, por lo tanto ya no es posible hacer inferencias sobre la población desde la muestra debido a que sólo arrojaría conclusiones equivocadas. Este es el principal problema de la heterocedasticidad, ya que al ser la varianza del error sesgada generaría un **error estándar de la regresión** ineficiente por lo que el error estándar de la regresión estaría subestimado o sobreestimado, es decir el error estándar de la regresión estaría equivocado derivado de ello, probar las hipótesis de **significancia individual** y **global** estarían equivocadas.
	• Debido a que el error estándar de la regresión es ineficiente, el **coeficiente de determinación**, que mide cuanto explican la(s) variable(s) explicativa(s) a la endógena también estaría equivocado.
	• Una vez más, debido al error estándar de la regresión estimado del modelo ineficiente, la **matriz de varianza y covarianza** de los estimadores mostraría valores incorrectos.
	• Los pronósticos y predicciones que se quieran realizar a partir del modelo ajustado pueden estar equivocados.

Tabla 3.6. Consecuencias de la violación a los supuestos del modelo de regresión lineal con estimación por MCO.
Elaboración propia
Fuente: (Pérez L., 2012) (Hanke & Wichern, 2006) (Novales, 1998)

El cuarto punto habla sobre una matriz de varianza y covarianza de los estimadores incorrectamente estimada. Para explicar este punto que habla sobre la estimación de MCO bajo heterocedasticidad primero se hará un repaso sobre la

estimación con varianza homocedástica y posteriormente se contrastará con una estimación con varianza heterocedástica para explicar sus diferencias. La varianza homocedástica en su forma matricial es $E(\mu\mu') = \sigma^2 I$, cuya demostración es:

$$E(\mu\mu') = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n]\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix} = E\begin{bmatrix}\mu_1^2 & \mu_1\mu_2 & \cdots & \mu_1\mu_n\\ \mu_2\mu_1 & \mu_2^2 & \cdots & \mu_2\mu_n\\ \vdots & \vdots & \ddots & \vdots\\ \mu_n\mu_1 & \mu_n\mu_2 & \cdots & \mu_n^2\end{bmatrix} \text{(3.4.73.)}$$

Donde al aplicar las esperanzas a cada elemento del producto, tenemos:

$$E(\mu\mu') = \begin{bmatrix}E(\mu_1^2) & E(\mu_1\mu_2) & \cdots & E(\mu_1\mu_n)\\ E(\mu_2\mu_1) & E(\mu_2^2) & \cdots & E(\mu_2\mu_n)\\ \vdots & \vdots & \ddots & \vdots\\ E(\mu_n\mu_1) & E(\mu_n\mu_2) & \cdots & E(\mu_n^2)\end{bmatrix} \text{(3.4.74.)}$$

Sí que $E(\mu_i^2) = \sigma^2$ y $E(\mu_i\mu_j) = 0$ podemos reemplazar en (3.4.74.)

$$E(\mu\mu') = \begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{bmatrix} = \sigma^2\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix} \text{(3.4.75.)}$$

De esta manera se demuestra $E(\mu\mu') = \sigma^2 I$, (Pérez L., 2012) Manifiesta que según (3.4.75.) podemos decir que el término de error tiene una distribución normal con media cero y una matriz de varianza y covarianza idéntica. Sin embargo, cuando la varianza no es constante, la matriz ya no es idéntica, es decir la diagonal ya deja de ser 1, para ser concretos (De Grange C., 2005) Señala su forma matricial tomando en cuenta que $var(\mu_i|X_i) = E(\mu_i^2) = \sigma_i^2$.

$$E(\mu\mu') = \begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix} = \sigma^2\Omega \text{ (3.6.1.)}$$

Cuando la varianza no es homocedástica, entonces la matriz de varianza y covarianza del término de error ya no es idéntica, por el contrario, ahora depende de una matriz Ω. Esta matriz tiene consecuencias en la estimación de la varianza y errores estándares de los estimadores. Recordemos que cuando se estima con varianza homocedástica, la varianza de los estimadores en su forma matricial es:

$$var - cov(\hat{\beta}) = \sigma^2(X'X)^{-1} \text{ (3.4.88.)}$$

De forma más extensa:

$$var - cov(\hat{\beta}) = \begin{bmatrix} var(\hat{\beta}_1) & cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ cov(\hat{\beta}_2, \hat{\beta}_1) & var(\hat{\beta}_2) & \cdots & cov(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_1) & cov(\hat{\beta}_k, \hat{\beta}_2) & \cdots & var(\hat{\beta}_k) \end{bmatrix} \text{(3.4.89)}$$

Ahora veamos cómo es la matriz varianza-covarianza de los estimadores con una varianza heterocedástica, comencemos en:

$$var - cov(\hat{\beta}) = [(X'X)^{-1}X'E(\mu\mu')X(X'X)^{-1}] \text{ (3.4.86.)}$$

Al aplicar $E(\mu\mu') = \sigma^2\Omega$ entonces:

$$var - cov(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1} \text{ (3.6.2.)}$$

(Greene, 2012) Comenta que debido a que (3.6.2.) halla las varianzas y los errores estándares ineficientes provoca que la inferencia usando las pruebas *t* y *F* pierden el sentido de ser interpretadas ya que demostraran conclusiones falsas, además que los estimadores estimados por MCO ya no son los mejores estimadores porque su varianza no es mínima. Por lo tanto, detectar la heterocedasticidad en un modelo resulta importante para comprobar que los estimadores cumplen con la propiedad de eficiencia y también realizar conclusiones verdaderas sobre las pruebas de significancia. A continuación, se explicará brevemente los métodos formales e informales para detectarla.

3.6.1.1. Métodos para detectar la existencia de heterocedasticidad.

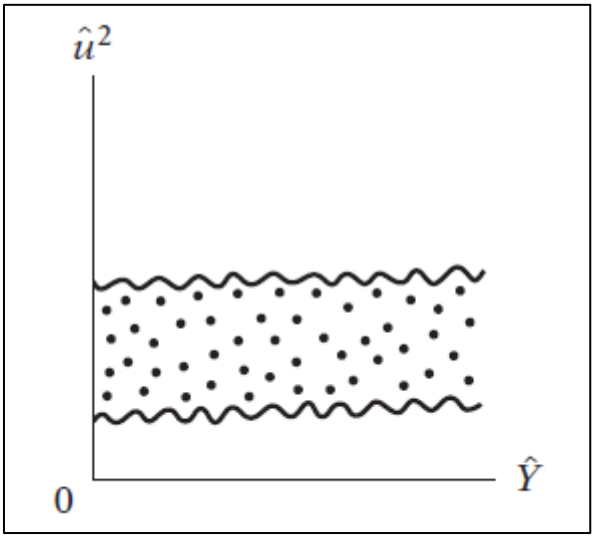
La heterocedasticidad en el modelo puede detectarse mediante los métodos informales que son gráficos de nubes de dispersión y métodos formales con el empleo de prueba de hipótesis. De hecho, la detección de autocorrelación, la no normalidad de los residuos y el error de especificación pueden detectarse mediante métodos formales e informales. Y tanto la heterocedasticidad como las demás violaciones a los supuestos de MCO referidos al término de error, se tomara en cuenta al término residual $\hat{\mu}_i$ para verificar la existencia o no de estos problemas en el modelo puesto que el término residual $\hat{\mu}_i$ es el estimador del término de error.

3.6.1.1.1. Métodos informales.

En este punto, uno podría preguntarse ¿Cómo se puede validar tal contraste gráfico si la varianza poblacional σ^2 es desconocida y el término del error μ_i también lo

es? Recordando lo que (Wooldrige, 2009) Explico: es que se intenta buscar la variable culpable de la heterocedasticidad ya que la varianza del término de error está en función de la variable independiente o de alguna variable independiente si el modelo fuese simple o múltiple respectivamente, (Novales, 1998) Justifica porque se usa al estimador del término de error el cual es el término residual, es decir se usa $\hat{\mu}_i$ y también el estimador de la varianza del termino de error que es $\hat{\sigma}^2$; debido a que ambas son aproximaciones a sus valores poblacionales respectivos es que es válido hacer uso de ellas para el contraste de heterocedasticidad en el modelo.

Luego de este preámbulo, podemos afirmar entonces que tal como ya se dijo anteriormente, los métodos informales son los gráficos de nube de dispersión el cual relaciona los valores residuales al cuadrado $\hat{\mu}_i^2$ con los valores estimados, también llamados ajustados o predichos, de la variable dependiente $\hat{Y}_i$ y se busca que no haya ningún patrón definido en los gráficos. Si hubiese algún patrón establecido es que podemos sospechar que el modelo presenta problemas de heterocedasticidad. El siguiente gráfico recogido de (Gujarati & Porter, 2010) Ponen de manifiesto cómo debería ser un gráfico libre de heterocedasticidad.



Gráfica 3.15. Grafica de dispersión entre $\hat{\mu}_i^2$ y $\hat{Y}_i$ libre de heterocedasticidad.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

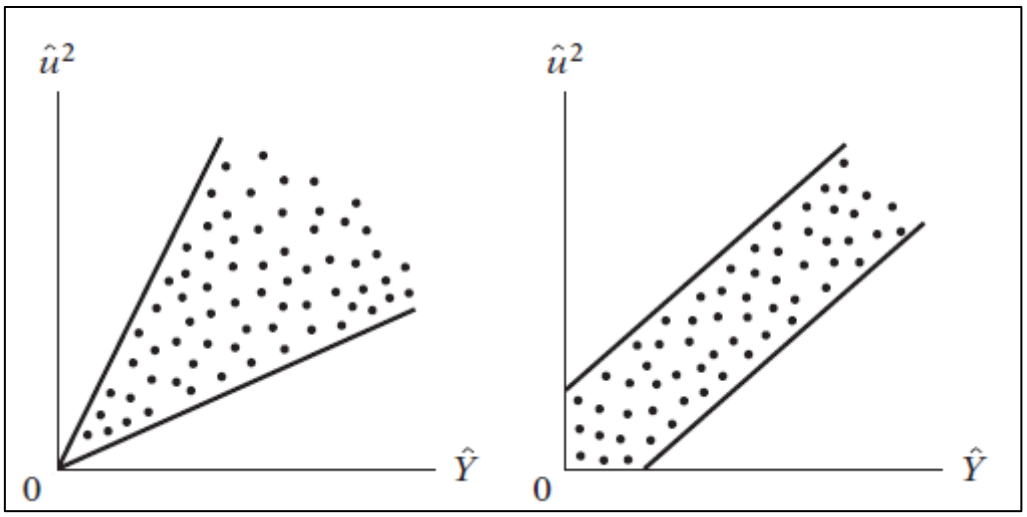
Lo que la gráfica 3.15. Quiere decir en palabras de (Gujarati & Porter, 2010) Es que no existe una relación sistemática entre los residuales al cuadrado y los valores estimados de la variable dependiente. Se puede llegar a esa conclusión ya que no se observa un patrón de crecimiento o decrecimiento ni tampoco valores atípicos que

podrían indicar señales de heterocedasticidad, de hecho, las líneas en forma de ondas que están en la parte superior e inferior del gráfico indican que la nube de puntos no esta tan dispersa. El mencionado patrón que muestra un crecimiento o decrecimiento entre ambas variables puede ser explicado según la siguiente cita textual:

> *"Dado que las series económicas presentan casi siempre una tendencia definida (positiva o negativa), la simple gráfica de error [se refiere al término de error] puede servir para conocer intuitivamente si el mero transcurso del tiempo da lugar a un incremento/decremento continuado del error, lo que sería significativo de una relación entre la evolución de las variables del modelo y los valores cada vez mayores o cada vez menores de este."* (De Grange C., 2005)

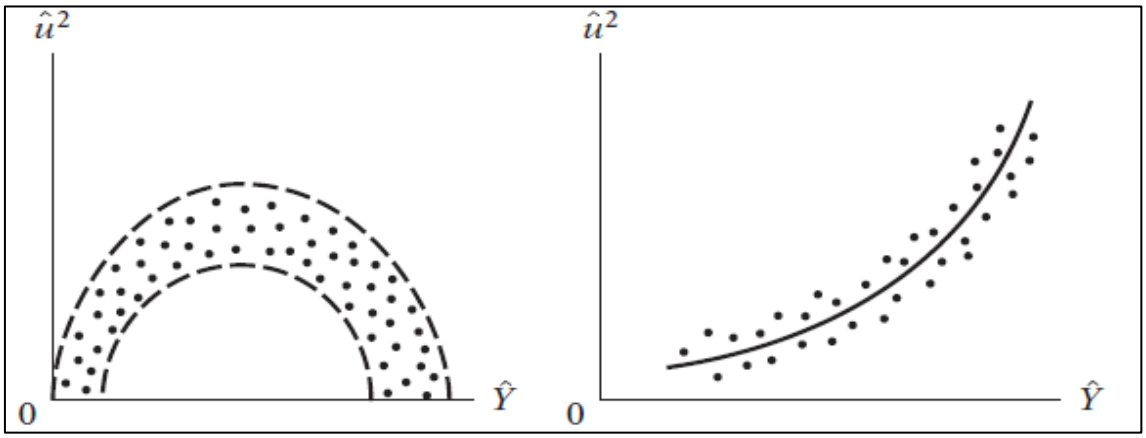
La cita anterior sugiere que a lo largo tiempo, las variables tienden a mostrar una tendencia la cual puede ser creciente o decreciente; esta tendencia es propia de las variables económicas y de datos de series temporales; no por ello la heterocedasticidad es exclusiva de las series temporales, de hecho la heterocedasticidad es más frecuente en los datos de corte transversal que en las series temporales, sin embargo lo que la cita indica es que usando un concepto tan sencillo como la evolución del tiempo se puede justificar la existencia de patrones. Pero ¿Cómo puede explicarse si se utiliza datos de corte transversal? La respuesta es fácil de intuir: suponga que se estudia los ingresos de una población en una ciudad determinada, el cual obtiene datos desde las más humildes viviendas hasta las más ostentosas viviendas entonces debido a una brecha sumamente profunda es que la varianza en el modelo aumentará; en términos más propios de la teoría econométrica la introducción de datos atípicos al modelo causa que existan patrones de crecimiento o decrecimiento en estos gráficos. Una última aclaración: los datos atípicos también pueden existir en las series temporales, pero son frecuentes a encontrarse en los datos de corte transversal.

Veamos entonces cómo son los gráficos de dispersión entre los residuos al cuadrado y los valores estimados de la variable dependiente que indican posible heterocedasticidad. Los siguientes gráficos han sido tomados de (Gujarati & Porter, 2010).



Gráfica 3.16. Grafica de dispersión entre $\hat{\mu}_i^2$ y $\hat{Y}_i$ con heterocedasticidad.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

La gráfica anterior muestra un claro patrón entre los residuos al cuadrado y los valores predichos de la variable dependiente, cuando se observan estos gráficos podemos sospechar fuertemente que la heterocedasticidad está presenta en el modelo, sin embargo es posible que los patrones no solo sean en forma lineal como es el caso del gráfico de la derecha; puede ser que encontremos una relación cuadrática tal como señala (Gujarati & Porter, 2010). Los siguientes gráficos lo representan.

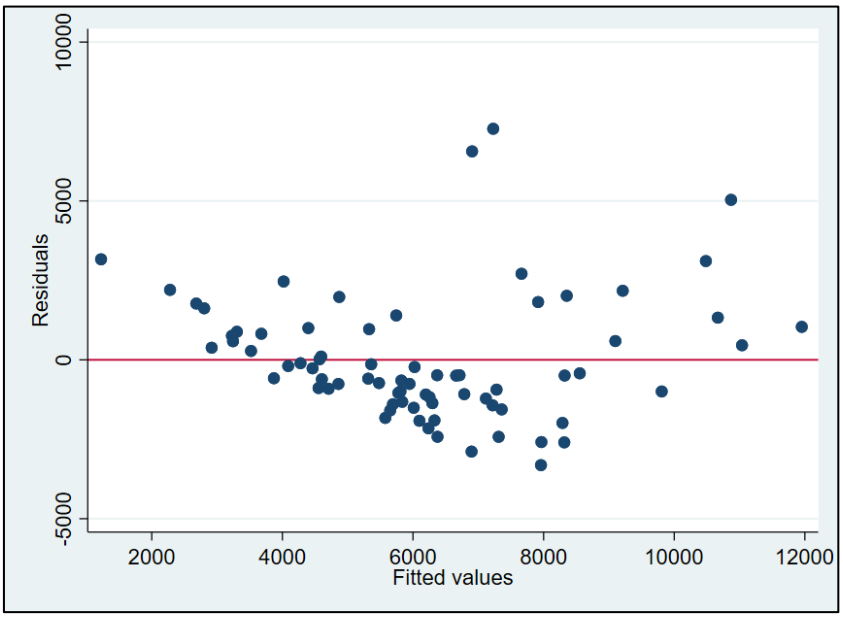


Gráfica 3.17. Grafica de dispersión entre $\hat{\mu}_i^2$ y $\hat{Y}_i$ con heterocedasticidad y una relación cuadrática.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

Por lo general cuando se tienen este tipo de gráficos de dispersión la heterocedasticidad ha sido provocado por un error en la forma funcional en el modelo y podría bastar con transformar la variable dependiente al cuadrado para corregir este problema. Con estos gráficos ha quedado claro que los patrones indican existencia de heterocedasticidad, pero recordemos que la heterocedasticidad es provocada por la existencia de datos atípicos en el modelo, por lo que para terminar de esclarecer las dudas veamos los siguientes gráficos recogidos del ejemplo que brinda (Pérez L., 2012) Ofreciendo otro punto de vista sobre esta parte de los métodos informales. Se tiene el siguiente modelo:

$$price_i = \hat{\beta}_1 + \hat{\beta}_2 weight + \hat{\beta}_3 mpg + \hat{\beta}_4 forxmpg + \hat{\beta}_5 foreign + \hat{\mu}_i \quad (3.6.3.)$$

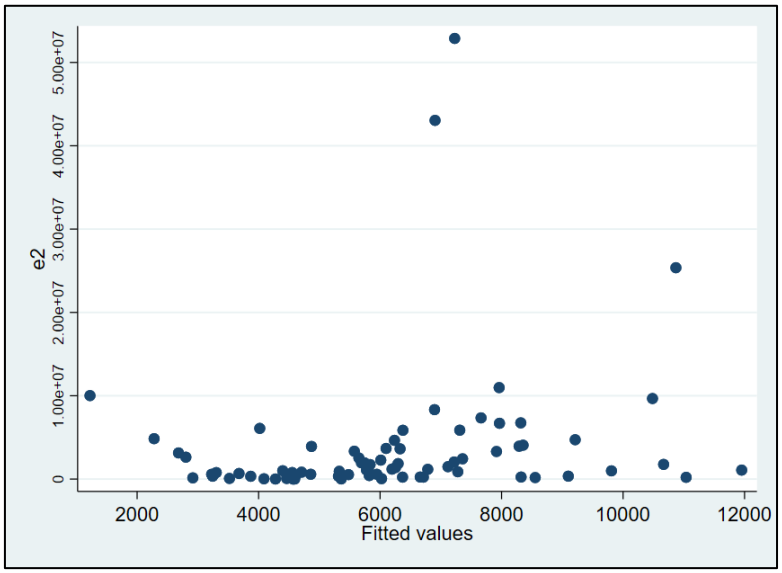
Veamos como es el gráfico realizado en el programa STATA sobre la dispersión entre los valores predichos de la variable dependiente y los residuos.



Gráfica 3.18. Grafica de dispersión entre $\widehat{\mu}_i^2$ y $\widehat{Y}_i$ con heterocedasticidad.
Elaboración propia
Fuente: (Pérez L., 2012)

En el gráfico se observa un patrón, el cual concentra los puntos por debajo de la línea roja, mientras que existen algunos puntos que están alejados. Se trata de los datos atípicos presenten en el modelo anteriormente estimado, además se puede observar que los puntos disminuyen y en cierto valor de la eje horizontal vuelven a crecer por lo que se podría apreciar una curvatura, el patrón sugiere que la forma funcional apropiada podría ser cuadrática tal como interpreta (Pérez L., 2012).

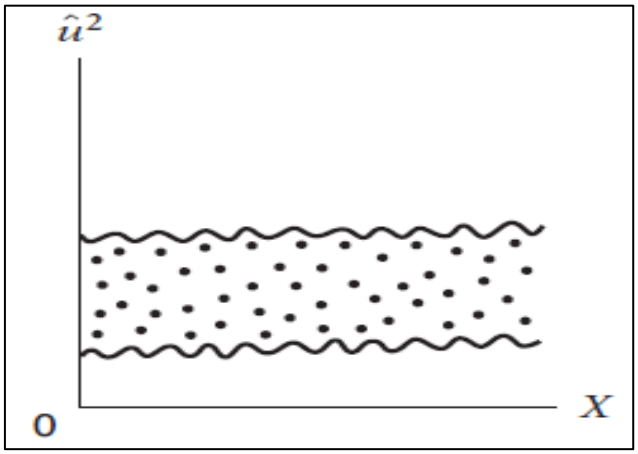
Tomando todo esto en cuenta podemos sospechar la existencia de heterocedasticidad, no obstante, a diferencia de los gráficos anteriores no ha tomado los residuos elevados al cuadrado. A continuación, veamos el siguiente gráfico que relaciona los valores al cuadrado de los residuos y los valores predichos de la variable dependiente.



Gráfica 3.19. Grafica de dispersión entre $\widehat{\mu_i^2}$ y $\widehat{price_i}$ con heterocedasticidad. Elaboración propia Fuente: (Pérez L., 2012)

Lo primero que se puede notar es que los datos atípicos persisten en el modelo por lo que la presencia de heterocedasticidad en el modelo es fuertemente sospechosa, incluso es más notorio que al grafico 3.18. Ya que la nube de puntos está altamente concentrada en la parte inferior del gráfico y además no se aprecia la supuesta curvatura en el gráfico. En resumen, lo que se intenta explicar con las gráficas 3.16., 3.17., 3.18. Y 3.19. Es que al momento de relacionar un gráfico de puntos entre los residuos ya sean al cuadrado o no con la variable dependiente se tiene que buscar al patrón en específico o la existencia de datos atípicos como en los dos últimos gráficos. Es recomendable realizar gráficos tanto con $\hat{\mu}_i$ y $\hat{\mu}_i^2$ con $\hat{Y}_i$.

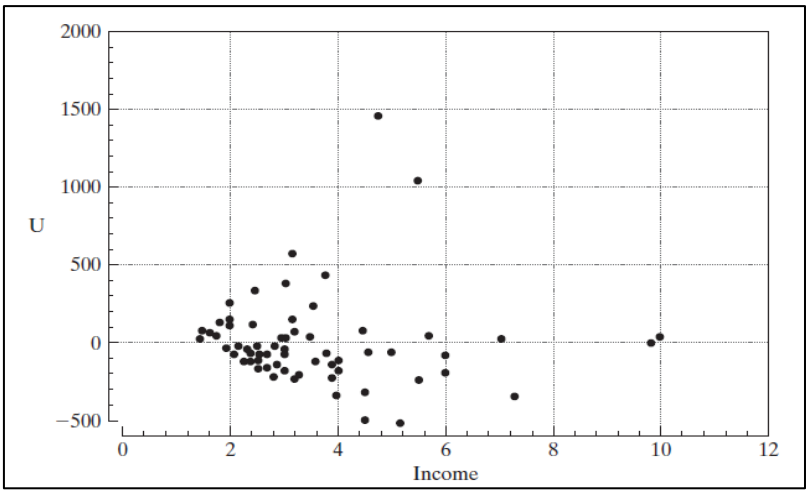
Aunque la sospecha no sea fuerte es conviene realizar gráficos de dispersión entre $\hat{\mu}_i$ y los valores de la(s) variable(s) explicativa(s). Con el fin de identificar la posible variable explicativa que cause el problema de heterocedasticidad en el modelo. Se presentan los gráficos de (Gujarati & Porter, 2010) Y al igual que los gráficos anteriores se espera a que no haya un patrón sistemático ni mucho menos datos atípicos.



Gráfica 3.20. Grafica de dispersión entre $\hat{\mu}_i^2$ y X_i libre de heterocedasticidad.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

El supuesto de homocedasticidad define que no puede existir una dependencia de la varianza del término de error y la(s) variable(s) explicativa(s), y cuando este supuesto se rompe se reconocen fácilmente patrones en gráficos de dispersión entre $\hat{\mu}_i$ y la(s) variable(s) explicativa(s).

(Greene, 2012) Muestra un patrón claro en la siguiente gráfica la cual muestra cómo la variable explicativa *income*$_i$ es causante de la heterocedasticidad en un modelo planteado por el autor.



Grafica 3.21. Grafica de dispersión entre $\hat{\mu}_i$ y $income_i$ con heterocedasticidad.
Elaboración: (Greene, 2012)
Fuente: (Greene, 2012)

Podemos observar en la gráfica 3.21. Existe una concentración en la nube de dispersión y al mismo tiempo se pueden ver datos que están muy alejados, siendo estos los ya mencionados datos atípicos. Cuando se encuentra este tipo de gráficos entre $\hat{\mu}_i$ y X_i podríamos señalar cuál es la variable causante de heterocedasticidad; no obstante, los métodos informales no son determinantes, por lo que es necesario contrastar con métodos formales los cuales haciendo uso de la prueba de hipótesis serán decisivos para demostrar la existencia o no de heterocedasticidad en el modelo.

3.6.1.1.2. Métodos formales.

Los contrastes formales, los cuales se refieren a los métodos formales, se emplean para saber con exactitud cómo se comportan los residuos con las variables explicativas. En palabras de (Novales, 1998) Estos métodos consisten en explorar la posibilidad que la varianza de los residuos dependan directamente por alguna variable explicativa, el autor justifica que esta situación es frecuente en las variables económicas y termina de señalar que cuando se puede encontrar alguna capacidad predictiva desde las variables explicativas hacia el termino residual, entonces existe heterocedasticidad en el modelo.

Existen muchas pruebas que pueden ser empleadas al momento de verificar la existencia o no de heterocedasticidad, por lo que solamente se explicara las más utilizadas: Test de White y Test de Breush-Pagan.

- **Prueba de White o test de White.**

La prueba de White para heterocedasticidad es una de las más extendidas y de mayor uso cuando se requiere verificar mediante un contraste formal la existencia o no de heterocedasticidad. (Galán F., y otros, 2016) Llama a este test como **una prueba robusta**, ya que no requiere realizar asumir previamente si los residuos son normales, es decir si siguen una distribución normal, o si se tiene en cuenta de forma a priori que alguna variable puede ser la causante de heterocedasticidad. Por ello es que se le conoce como **la prueba general de heterocedasticidad de White**, de hecho (Greene, 2012) Señala que este es el motivo por el cual es tan empleada, pero al mismo tiempo señala que esta podría ser su principal desventaja debido a que al ser tan general no solo puede probar la existencia o no de heterocedasticidad sino también de un sesgo de especificación.

La prueba de White debe ser tratada con cuidado y tomando consciencia que pueden existir otras pruebas que son mejores que esta. Pese a esto, su importancia radica

en la idea con la cual se puede verificar la presencia o no de heterocedasticidad. (De Grange C., 2005) Detalla en la siguiente cita como White consigue verificar la existencia o no de heterocedasticidad.

> *"La idea subyacente es determinar si las variables explicativas del modelo, sus cuadrados y todos sus cruces posibles no repetidos sirven para determinar la evolución del error al cuadrado."* (De Grange C., 2005)

Esto quiere decir que con la prueba de White se busca determinar cuál variable explicativa tiene significancia individual al momento de explicar la varianza muestral de los errores. Para ello se realizan los siguientes pasos teniendo el siguiente modelo múltiple:

$Y = \hat{\beta}_1 + \hat{\beta}_2 X_1 + \hat{\beta}_3 X_2 + \hat{\mu}$ (3.6.4.)

Paso 1. Realizar la regresión de (3.6.4.) mediante MCO.

Paso 2. Calcular los residuos después de haber realizado la regresión en el paso anterior.

Paso 3. Elevar al cuadrado los residuos previamente calculados, es decir obtener $\hat{\mu}^2$.

Paso 4. Calcular los productos de las variables explicativas y sus respectivos cuadrados.

Paso 5. Realizar mediante MCO la siguiente regresión auxiliar.

$\hat{\mu}^2 = \hat{\alpha}_1 + \hat{\alpha}_2 X_1 + \hat{\alpha}_3 X_2 + \hat{\alpha}_4 X_2 X_3 + \hat{\alpha}_5 X_2^2 + \hat{\alpha}_6 X_3^2 + \hat{v}$ (3.6.5.)

Tal vez pueda generarse la pregunta ¿Por qué White tomó en cuenta $\hat{\mu}^2$ en vez de $\hat{\mu}$? (Court & Rengifo, 2011) Explica el motivo. Debido a que la esperanza de los errores es igual a 0 entonces la varianza es: $var(\mu) = E(\mu^2) - E(\mu)^2 = E(\mu^2)$ por ello es que si se examina el comportamiento del error al cuadrado con las variables explicativas se logra determinar la existencia o no de errores homocedásticos.

Paso 6. Realizar la prueba de hipótesis, la cual puede ser descrita siguiendo las siguientes estructuras:

$H_0: no\ existe\ homocedasticidad$ (3.6.6.)

$$H_1: existe\ homocedasticidad$$

$H_0: \hat{\alpha}_2 = \hat{\alpha}_3 = \hat{\alpha}_4 = \hat{\alpha}_5 = \hat{\alpha}_6$ (3.6.7.)

$$H_1: \hat{\alpha}_2 \neq \hat{\alpha}_3 \neq \hat{\alpha}_4 \neq \hat{\alpha}_5 \neq \hat{\alpha}_6$$

Paso 7. Contrastar la anterior prueba de hipótesis utilizando la siguiente distribución: $n * {R^2}_{auxiliar} \sim X^2_{0.05,gl}$.

Al igual que cuando se pretende especificar si existe o no significancia estadística, cuando el estadístico calculado asintótico que es $n * {R^2}_{auxiliar}$ supera al estadístico crítico X^2_{gl}, el cuál es el estadístico ji cuadrado crítico con los grados de libertad determinados por el número de estimadores que tiene la regresión auxiliar omitiendo el intercepto, se puede rechazar la hipótesis nula y encontramos que existe heterocedasticidad en el modelo. (Baum, 2006) Advierte que debido al consumo de tantos grados de libertad por el número de regresores en el modelo auxiliar es que esta prueba puede no ser la recomendada para detectar la heterocedasticidad. Incluso (Gujarati & Porter, 2010) También comentan sobre la naturaleza de la prueba de White para heterocedasticidad, afirmando que debido a estos grados de libertar en ocasiones tan elevados la heterocedasticidad no necesariamente es la causante de rechazar la hipótesis nula sino también la presencia de un error de especificación está generando problemas en el modelo. Para tener una prueba de White que se ajuste solamente a verificar la existencia o no de heterocedasticidad (Gujarati & Porter, 2010) Recomienda excluir los productor cruzados en el modelo auxiliar.

Finalmente, lo que llevó a White a proponer esta prueba de heterocedasticidad, se debe según explica (Wooldrige, 2009) A tomado la prueba de Breush-Pagan, le agregó los productos cruzados y los cuadrados de las variables explicativas. Por lo tanto, se puede deducir que la prueba que se verá a continuación inspiró a la creación de este test.

- **Prueba de Breush-Pagan.**

También llamada prueba de multiplicador de Breush-Pagan. Esta prueba de heterocedasticidad tiene un procedimiento parecido al de White para verificar la existencia o no de heterocedasticidad, el cual en resumen es plantear una prueba de hipótesis, realizar una regresión auxiliar y contrastar la prueba de hipótesis.

(Colin C. & Trivedi, 2005) Explican que la prueba estándar de Breusch-Pagan dependía fuertemente del supuesto de que los errores se distribuyen normalmente, sin embargo, tiempo después se logró desarrollar una versión de la misma prueba la cual propone que ya no es necesario el supuesto de la normalidad de los errores según afirma (Greene, 2012).

La idea para realizar esta prueba se recoge en la siguiente cita.

"La idea del contraste es comprobar si se puede encontrar un conjunto de variables Z, que sirvan para explicar la evolución de la varianza de las perturbaciones aleatorias, estimada está a partir del cuadrado de los errores del modelo inicial sobre el que se pretende comprobar si existe o no heterocedasticidad." (De Grange C., 2005)

Este test tiene los siguientes pasos, los cuales serán explicados con el siguiente modelo:

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \cdots + \hat{\beta}_k X_k + \hat{\mu} \quad (3.6.8.)$$

Paso 1. Realizar la regresión de (3.6.8.) mediante MCO.

Paso 2. Obtener los residuos después haber realizado la regresión y elevarlos al cuadrado. Es decir obtenga $\widehat{\boldsymbol{\mu}}^{\boldsymbol{2}}$.

Paso 3. Calcular $\tilde{\sigma}^2 = \frac{\sum \hat{\mu}^2}{n}$, el cual (Gujarati & Porter, 2010) **lo identifican como el estimador de máxima verosimilitud de σ^2.**

Paso 4. Mediante MCO estimar la siguiente regresión auxiliar:

$$\frac{\hat{\mu}^2}{\tilde{\sigma}^2} = \hat{\alpha}_1 + \hat{\alpha}_2 Z_2 + \hat{\alpha}_3 Z_3 + \cdots + \hat{\alpha}_p Z_p + \hat{v} \quad (3.6.9.)$$

Paso 5. Plantear la prueba de hipótesis:

$$H_0: no\ existe\ homocedasticidad$$
$$H_1: existe\ homocedasticidad \quad (3.6.6.)$$

Paso 6. El estadístico calculado será: *SCE/2* el cual sigue una distribución de X^2_{p-1}. Por lo que se denota como $\frac{SCE}{2} \sim X^2_{p-1}$ el cual si el estadístico calculado que es *SCE/2* supera al estadístico X^2_{p-1} crítico el cual tiene $p-1$ grados de libertad, donde p es el número de estimadores en la regresión auxiliar

entonces podemos rechazar la hipótesis nula y asumir la existencia de heterocedasticidad caso contrario entonces no se rechaza la hipótesis nula y se asume que el modelo no presenta heterocedasticidad.

La pregunta entonces es ¿Cuáles son las variables independientes en el modelo auxiliar? En teoría, las variables explicativas del modelo auxiliar son las variables explicativas del modelo original. Para ser más precisos, si recordamos que el supuesto de homocedasticidad se viola cuando la varianza del término de error está en función de alguna(s) variable(s) explicativa(s), por lo que Z podrían ser todas o algunas variables explicativas de las cuales se sospecha que genera el problema de heterocedasticidad, similar a la prueba de White. Según lo anterior descrito la función general de la prueba de BP seria:

$$\sigma^2 = f(\hat{\alpha}_1 + \hat{\alpha}_2 Z_2 + \hat{\alpha}_3 Z_3 + \cdots + \hat{\alpha}_p Z_p) \text{ (3.6.10.)}$$

Lo que (3.6.9.) significa es que se han tomado a todas las variables explicativas en (3.6.8.) para determinar que la varianza del término de error depende de una función dada en (3.6.10.) De hecho, según este planteamiento (Wooldrige, 2009) Sugiere que si la heterocedasticidad solo es producida por algunas variables entonces la función podría ser:

$$\sigma^2 = f(\hat{\alpha}_1 + \hat{\alpha}_2 Z_1 + \hat{\alpha}_3 Z_3) \text{ (3.6.11.)}$$

Cuya diferencia de (3.6.10.) es que la función solo toma en cuenta a las dos primeras variables explicativas y se realiza siguiendo el mismo procedimiento.

Finalmente (Wooldrige, 2009) Propone usar los siguientes estadísticos calculados, el primero:

$$Fc = \frac{\frac{R_n^2}{k}}{\frac{(1-R_n^2)}{n-k-1}} \sim F_{n-k-1}^{k} \text{ (3.6.12.)}$$

El cual R_n^2 y k corresponden a datos conseguidos del modelo:

$$\hat{\mu}^2 = \hat{\alpha}_1 + \hat{\alpha}_2 Z_2 + \hat{\alpha}_3 Z_3 + \cdots + \hat{\alpha}_p Z_p + \hat{v} \text{ (3.6.13.)}$$

El primero es el coeficiente de determinación y el segundo el número de regresores del modelo, y sigue una distribución de $F_{0.05,n-k-1}^{k}$ donde se somete a la misma prueba de hipótesis y se consigue contrastar con la misma regla de decisión.

Y el segundo es:

$n * R_n^2 \sim X_k^2$ (3.6.14.)

Donde R_n^2 y k se consiguen del modelo descrito en el párrafo anterior y siguen una distribución de ji cuadrado con *k* grados de libertad, la cual también sigue la misma regla de decisión.

3.6.1.2. Métodos para corregir la existencia de heterocedasticidad.

Tal como indica el título de esta sección, se explicara un breve repaso sobre cómo corregir el problema de heterocedasticidad en un modelo estimado mediante MCO.

3.6.1.2.1. Mínimos Cuadrados Generalizados.

Este método de corrección es preferible al método de estimación por MCO, debido a que los estimadores hallados mediante MCO pierden la propiedad de tener la varianza mínima provocando estimadores no MELI. Por lo que es necesario hacer uso de otro método de estimación capaz de asegurarnos la estimación libre de heterocedasticidad. Un método ampliamente usado con el fin de corregir la heterocedasticidad en el modelo es el **método de los mínimos cuadrados generalizados (MCG)** de hecho este método también es empleado para corregir el problema de autocorrelación.

¿En qué consiste el método de estimación mediante MCG? Para comprender esto tenemos que recordar el principio de estimación en el que se basa el MCO el cual es el causante que la estimación MCO arroje un modelo con heterocedasticidad.

Anteriormente, cuando se habló sobre el **principio de mínimos cuadrados** se explicó que es el principio en el cual se basa la estimación de MCO que surge un problema que acorde a (Gujarati & Porter, 2010) El modelo de regresión **pondera** los residuos de igual forma, es decir no toma en cuenta la distancia de cada uno de ellos con respecto en la línea de regresión tal como se muestra en el grafico 3.10.

Siguiendo la teoría de (Novales, 1998) Se concluye que el problema radica en esa forma de estimación usando el principio de mínimos cuadrados. La explicación se expresa en la siguiente cita.

> *"En efecto, al estimar por MCO tratamos de minimizar la Suma de Cuadrados de los Residuos, tratando a todos ellos igualmente. Pero si la varianza correspondiente a cada observación muestral es diferente, esto no parece muy*

adecuado: cuanto mayor sea la varianza, mayor tenderá a ser el componente no explicable de la variable dependiente y más errática o menos fiable será dicha observación." (Novales, 1998)

En resumen, el principio de mínimos cuadrados genera un problema porque trata a todos los residuos de forma igual sin importar la distancia hacia la línea de regresión y debido a la posible aparición de datos atípicos en el modelo es que la varianza de los residuos es más volátil y pierde la capacidad de ser constante para cada observación.

En este punto vale hacer una aclaración para prevenir confusiones posteriores, cuando se intenta aplicar el método de MCG para corregir la presencia de heterocedasticidad en el modelo, el nombre de MCG cambia y se le conoce como **Mínimos Cuadrados Generalizados Ponderados** o simplemente **Mínimos Cuadrados Ponderados (MCP).** (Novales, 1998) Explica que este método es un caso particular del método de estimación de MCG, se podría decir que es una extensión del MCG. A partir de este punto se referirá a este método como MCP y recibe este nombre porque aplicará una ponderación distinta a cada una de las observaciones de tal forma que en palabras de (Novales, 1998) se busca minimizar la suma cuadrática ponderada haciendo que los residuos que corresponden a una observación con mayor varianza tengan una menor ponderación.

Para explicar cómo corregir la heterocedasticidad mediante MCP primero explicaremos la estimación mediante MCG y posteriormente se explicará la del MCP.

Para lograr minimizar la suma cuadrática ponderada se debe realizar el MCG, el cual lo que hace es transformar todo el modelo original, dividiendo cada observación desde la variable dependiente hasta las variables independientes entre la ponderación $w_i = \frac{1}{\sigma_i^2}$, para lograrlo obviamente se debe conocer el valor de las varianzas, pero tal como (Escobar M., Fernández M., & Bernardi, 2012) Señalan, lo que pondera realmente es a los residuos cuadráticos por lo que se debe realizar la siguiente modificación $\sqrt{w_i} = \frac{1}{\sigma_i}$, de esta manera teniendo el modelo original $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu_i$ el cual está expresado en sus términos poblacionales se consigue la transformación cuando lo dividimos entre σ_i y el modelo queda expresado de la siguiente forma.

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_{2i}}{\sigma_i} + \beta_3 \frac{X_{3i}}{\sigma_i} + \cdots + \beta_k \frac{X_{ki}}{\sigma_i} + \frac{\mu_i}{\sigma_i} \text{ (3.6.15.)}$$

(Greene, 2012) Explica el uso de esta ponderación usando las siguientes matrices, para empezar recuerde que:

$$E(\mu\mu'|X) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \sigma^2\Omega = \sigma^2 \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \text{ (3.6.1.)}$$

Además, del modelo original expresado en matrices $Y = X\beta + \mu$ cuya varianza es (3.6.1.) es decir heterocedástica, se obtienen los estimadores eficientes según el método de MCG cuando se asume que la matriz Ω es conocida y además simétrica. El autor asume que existe una matriz no singular P que contiene las ponderaciones, la cual es:

$$\Omega^{-1} = P'P \text{ (3.6.16.)}$$

Si multiplicamos P en el modelo original con heterocedasticidad obtenemos:

$PY = PX\beta + P\mu$ (3.6.17)

El cual tiene su equivalente a:

$Y^* = X^*\beta + \mu^*$ (3.6.18.)

Donde la varianza del término de error es: $E(\mu^*\mu^{*\prime}|X) = P\sigma^2\Omega P'$ y si aplicamos (3.6.16.) entonces obtenemos que la varianza del término de error es $E(\mu^*\mu^{*\prime}|X) = \sigma^2\mathrm{I}$. Así, habremos obtenido un error sin una varianza heterocedástica, por lo que en el modelo transformado el problema está resuelto.

Para hallar los estimadores de (3.6.18.) se sigue la siguiente fórmula matricial $\hat{\beta} = (X^{*\prime}X^*)^{-1}(X^{*\prime}Y^*)$, entonces al reemplazar lo que se describió en (3.6.17.) se reescribe de tal forma que los estimadores conseguidos por MCG son:

$\widehat{\beta_{MCG}} = (X^{*\prime}X^*)^{-1}(X^{*\prime}Y^*)$ (3.6.19.)

$\widehat{\beta_{MCG}} = [(PX)'(PX)]^{-1}[(PX)'(PY)]$ (3.6.20.)

$\widehat{\beta_{MCG}} = (X'P'PX)^{-1}(X'P'PY)$ (3.6.21.)

$\widehat{\beta_{MCG}} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y)$ (3.6.22.)

Y además la varianza del estimador es homocedástica y se halla matricialmente mediante:

$Var(\widehat{\beta_{MCG}}) = \sigma^2(X^{*\prime}X^*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$ (3.6.23.)

De (3.6.22.) y (3.6.23.) se deduce que la matriz Ω debe ser conocida tal como anteriormente se explicó, y conocer Ω implica también conocer los valores de σ_i^2, pero esto en la práctica no es posible, entonces **¿Cómo aplicar MCG para corregir la heterocedasticidad? La respuesta es utilizar MCP** ya que permite aproximar el valor de σ_i^2 a una función de las variables independientes, de esta manera se puede representar que $\sigma_i^2 = f(Z_i)$ donde Z_i hace referencia a las variables independientes que puedan generar problemas de heterocedasticidad. (Pérez L., 2012) Identifica alguna de las funciones más comunes, entre ellas son: $\sigma_i^2 = \sigma^2 Z$, $\sigma_i^2 = \sigma^2 Z^2$ y estas son las matrices de $var - cov(\mu)$, respectivamente:

$$\Omega = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_n \end{bmatrix} \text{ Y } \Omega = \begin{bmatrix} Z_1^2 & 0 & \cdots & 0 \\ 0 & Z_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_n^2 \end{bmatrix} \text{ (3.6.24.)}$$

Las ponderaciones según estas funciones serán $\frac{1}{Z_n}$ y $\frac{1}{Z_n^2}$ respectivamente y al igual que en (3.6.15.) se dividió entre la desviación estándar, según sea la función que sigue la varianza del término de error, se debe dividir entre la raíz cuadrada de la variable explicativa que genera heterocedasticidad. Por ejemplo:

Si la función fuese $\sigma_i^2 = \sigma^2 X_2$ entonces se debe realizar la siguiente transformación.

$$\frac{Y_i}{\sqrt{X_2}} = \beta_1 \frac{1}{\sqrt{X_2}} + \beta_2\sqrt{X_2} + \beta_3 \frac{X_{3i}}{\sqrt{X_2}} + \cdots + \beta_k \frac{X_{ki}}{\sqrt{X_2}} + \frac{\mu_i}{\sqrt{X_2}} \text{ (3.6.25.)}$$

O si fuese $\sigma_i^2 = \sigma^2 X_2^2$ entonces se sigue la siguiente transformación.

$$\frac{Y_i}{X_2} = \beta_1 \frac{1}{X_2} + \beta_2 + \beta_3 \frac{X_{3i}}{X_2} + \cdots + \beta_k \frac{X_{ki}}{X_2} + \frac{\mu_i}{X_2} \text{ (3.6.26.)}$$

Una función de poco uso es $\sigma_i^2 = \sigma^2 E[Y]^2$, cuya transformación es.

$$\frac{Y_i}{E[Y]} = \beta_1 \frac{1}{E[Y]} + \beta_2 \frac{X_{2i}}{E[Y]} + \beta_3 \frac{X_{3i}}{X_2} + \cdots + \beta_k \frac{X_{ki}}{E[Y]} + \frac{\mu_i}{E[Y]} \text{ (3.6.27.)}$$

En las transformaciones anteriores, el error es homocedástico ya que a diferencia de MCO, la estimación por MCP le pone una ponderación mayor a las observaciones de menor varianza mientras que a aquellas observaciones con mayor varianza se le pone una

ponderación menor o en su defecto ni siquiera se le pone una ponderación lo que hace que desaparezca.

Sin embargo, el problema de usar MCP radica en primer lugar que se debe conocer la naturaleza de la heterocedasticidad, es decir la función de la cual depende la varianza heterocedástica y además que en algunas funciones los resultados no se pueden interpretar, por ejemplo (Greene, 2012) Indica que en estos modelos de regresión es difícil o en su defecto imposible interpretar el coeficiente de determinación R^2 cuando la función es $\sigma_i^2 = \sigma^2 X_2^n \; n > 2$, es decir cuando depende de alguna potencia mayor a 2, porque el modelo carecería de intercepto, de hecho el coeficiente de determinación no debería ser tomado en cuenta si es mayor en el modelo transformado que en el modelo original. Si bien es cierto, los estimadores que se hallan usando MCP son eficientes y consistentes otro problema de esta estimación surge cuando se usan pesos que están correlacionados, ya que al usarse pesos correlacionados los estimadores son ineficientes y además incorrectos.

Los estimadores de MCP se hallan empleando matrices. Primero conviene recordar que:

$$E(\mu\mu'|X) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \sigma^2\Omega = \sigma^2 \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} \text{ (3.6.1.)}$$

Lo cual se puede resumir en $\sigma_i^2 = \sigma^2 w_i$ (Greene, 2012) Explica que esta función sigue una distribución normalizada por lo que: $tr(\Omega) = \sum w_i = n$, con esto en cuenta entonces las ponderaciones que se encuentran en la diagonal de Ω^{-1} sería igual a $\frac{1}{w_i}$. Entonces el modelo transformado es:

$$PY = \begin{bmatrix} \frac{Y_1}{\sqrt{w_1}} \\ \frac{Y_2}{\sqrt{w_2}} \\ \vdots \\ \frac{Y_i}{\sqrt{w_i}} \end{bmatrix} \text{ Y } PX = \begin{bmatrix} \frac{X_1}{\sqrt{w_1}} \\ \frac{X_2}{\sqrt{w_2}} \\ \vdots \\ \frac{X_i}{\sqrt{w_i}} \end{bmatrix} \text{ (3.6.28.)}$$

Entonces el estimador MCP sería:

$$\widehat{\beta_{MCP}} = [\sum w_i XX']^{-1}[\sum w_i XY] \text{ (3.6.29.)}$$

En (3.6.29.) acorde a (Greene, 2012) Las ponderaciones son altas en aquellas observaciones con menores varianzas. Sin embargo (3.6.29.) está expresado en una forma muy general, por ello cuando se asume que la varianza depende de alguna regresora lo que se hace es asumir una aproximación hacia la matriz Ω en vez de estimarla. Por lo tanto (3.6.29.) puede reescribirse como:

$$\widehat{\beta_{MCP}} = (X'\mathrm{V}^{-1}X)^{-1}(X'\mathrm{V}^{-1}Y) \text{ (3.6.30.)}$$

Esta nueva forma de expresarla es muy parecida a $\widehat{\beta_{MCG}} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}Y)$ la diferencia radica en que la matriz V, la cual es la matriz que contiene las ponderaciones en su diagonal, está expresando la dependencia que tiene la varianza del término de error con una o más variables independientes. (Colin C. & Trivedi, 2005) Complementan lo anterior afirmando que **en (3.6.30.) no se está asumiendo que $\mathbf{V^{-1} = \Omega^{-1}}$ sino que se le aproxima en función de alguna regresora**. Por ello es que aunque la varianza de los estimadores MCG sea $Var(\widehat{\beta_{MCG}}) = \sigma^2(X'\Omega^{-1}X)^{-1}$ cuando se aplica MCP la varianza se reescribe y se obtiene al resolver:

$$Var(\widehat{\beta_{MCP}}) = \sigma^2(X'\mathrm{V}^{-1}X)^{-1}X'\mathrm{V}^{-1}\Omega\mathrm{V}^{-1}X(X'\Omega^{-1}X)^{-1} \text{ (3.6.31.)}$$

En resumen, lo que se busca hacer con la estimación por MCP es conocer la naturaleza de la heterocedasticidad en el modelo, es decir conocer cuál es la variable independiente que genera heterocedasticidad y transformar el modelo original en función a la raíz de esa variable independiente. En la mayoría de modelos, la función es $\sigma_i^2 = \sigma^2 Z$. Para acabar esta parte, es necesario recalcar que según (Novales, 1998) El modelo transformado no debe ser usado para calcular los estadísticos *t* ni *F*, ni mucho menos calcular los residuos, tal vez solamente para comprobar que el problema está resuelto. Los estimadores del modelo transformado sustituirán a los del modelo original al igual que sus errores estándares y la varianza del término residual.

Finalmente, se hablara un poco sobre la varianza del término de error cuando depende de alguna variable independiente, es decir de $\sigma_i^2 = \sigma^2 Z$ ya que esta es la función más frecuente con la que uno se topa cuando se tiene una modelo con varianza heterocedástica. Cuando $\sigma_i^2 = \sigma^2 X_2$, es porque el modelo original queda transformado en:

$$\frac{Y_i}{\sqrt{X_2}} = \beta_1 \frac{1}{\sqrt{X_2}} + \beta_2\sqrt{X_2} + \beta_3 \frac{X_{3i}}{\sqrt{X_2}} + \cdots + \beta_k \frac{X_{ki}}{\sqrt{X_2}} + \frac{\mu_i}{\sqrt{X_2}} \text{ (3.6.25.)}$$

La cual $\frac{\mu_i}{\sqrt{X_2}} = v$, entonces $E(v^2) = \sigma^2$ por ello es que es válido aplicar MCO a (3.6.25.) Donde la matriz V seria:

$$\sigma^2 V = \sigma^2 \begin{bmatrix} X_{21} & 0 & \cdots & 0 \\ 0 & X_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_{2n} \end{bmatrix} \text{(3.6.32.)}$$

De la cual se encuentra la matriz P dividendo a cada observación entre $\sqrt{X_{2i}}$ ya que así se consigue obtener las ponderaciones.

$$P = \begin{bmatrix} \frac{1}{\sqrt{X_{21}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{X_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{X_{2i}}} \end{bmatrix} \text{(3.6.33.)}$$

Entonces el modelo transformado seria:

$$PY = \begin{bmatrix} \frac{1}{\sqrt{X_{21}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{X_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{X_{2i}}} \end{bmatrix} . \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \end{bmatrix} = \begin{bmatrix} \frac{Y_1}{\sqrt{X_{21}}} \\ \frac{Y_2}{\sqrt{X_{22}}} \\ \vdots \\ \frac{Y_i}{\sqrt{X_{2i}}} \end{bmatrix} = Y^* \text{ (3.6.34.)}$$

$$PX = \begin{bmatrix} \frac{1}{\sqrt{X_{21}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{X_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{X_{2i}}} \end{bmatrix} . \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2i} & X_{3i} & \cdots & X_{ki} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{\sqrt{X_{21}}} & \sqrt{X_{21}} & \frac{X_{31}}{\sqrt{X_{21}}} & \cdots & \frac{X_{k1}}{\sqrt{X_{21}}} \\ \frac{1}{\sqrt{X_{22}}} & \sqrt{X_{22}} & \frac{X_{32}}{\sqrt{X_{22}}} & \cdots & \frac{X_{k2}}{\sqrt{X_{22}}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{X_{2i}}} & \sqrt{X_{2i}} & \frac{X_{3i}}{\sqrt{X_{2i}}} & \cdots & \frac{X_{ki}}{\sqrt{X_{2i}}} \end{bmatrix} = X^*, \begin{bmatrix} \frac{\mu_1}{\sqrt{X_{21}}} \\ \frac{\mu_1}{\sqrt{X_{22}}} \\ \vdots \\ \frac{\mu_1}{\sqrt{X_{2i}}} \end{bmatrix} = \mu^* \quad \text{(3.6.35.)}$$

Y al recordar que $\Omega^{-1} = P'P$ pero V es la aproximación de Ω entonces en MCP tenemos: $\mathrm{V}^{-1} = P'P$, en su forma extensa tenemos:

$$P'P = \begin{bmatrix} \frac{1}{X_{21}} & 0 & \cdots & 0 \\ 0 & \frac{1}{X_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{X_{2I}} \end{bmatrix} = \begin{bmatrix} X_{21} & 0 & \cdots & 0 \\ 0 & X_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_2 \end{bmatrix}^{-1} = \mathrm{V}^{-1} \text{ (3.6.36.)}$$

Lo cual se reemplaza en $\widehat{\beta_{MCP}} = (X'\mathrm{V}^{-1}X)^{-1}(X'\mathrm{V}^{-1}Y)$.

3.6.1.2.2. Errores estándar robustos.

Sin duda alguna la estimación mediante MCP parece la estimación más apropiada para corregir el problema de heterocedasticidad, sin embargo, el método requiere conocer cuál es la variable independiente que influye en la varianza del término de error para lograr corregir el modelo.

Por ello, lo que aparenta ser la solución podría convertirse en un problema si no se conoce la naturaleza heterocedástica de σ_i^2, por lo tanto lo que White propusó es simplemente elegir la estimación que brinde estimadores menos eficientes pero insesgados de MCO, en palabras de (Bravo & Vásquez Javiera, 2008) Esta parece ser la solución más sensata que utilizar MCP.

De esta forma cuando se elige la estimación de MCO que tenga estimadores insesgados y poco eficiente lo que se usa es el **estimador de matriz de varianza-covarianza consistente con heterocedasticidad** que por sus siglas en ingles corresponde a **HCCME** planteada en su momento por Huber y posteriormente redescubierta por White, a este estimador se le conoce también como **Errores estándares de Huber/White** o simplemente **Errores Estándares de White**.

Por lo que los estimadores usando MCO o el método que errores robustos son iguales, lo único que cambiará serán las varianzas de los estimadores y con estos, las pruebas *t*, los errores estándares y la prueba *F*. Pero ¿Cómo White corrige la heterocedasticidad en el modelo sin conocer la naturaleza de σ_i^2? Todo empieza tomando en cuenta que los estimadores en MCO continúan siendo insesgados, consistentes y asintóticamente normal distribuidos, por lo que la matriz asintótica de las varianzas de los estimadores es:

$$Asy.var(\beta) = \frac{\sigma^2}{n}(plim\frac{1}{n}X'X)^{-1}(plim\frac{1}{n}X'\Omega X)(plim\frac{1}{n}X'X) \text{ (3.6.37.)}$$

Y la estimación de la matriz de covarianzas asintóticas podría estar basada en

$var(\beta) = (X'X)^{-1}(\sigma^2 \sum w_i XX')(X'X)^{-1}$ (3.6.38.)

Cabe aclarar que el término "**asintótica**", en términos simples, se refiere a muestras grandes que tienden hacia el infinito y tiene que ver con la propiedad asintótica de los estimadores, siendo una de esas propiedades la **consistencia** la cual ha sido definida como: a medida que la muestra aumenta los estimadores tienen a acercarse a su valor poblacional. Para (Greene, 2012) La media cuadrática consistente de los estimadores de MCO depende del comportamiento limitante de la matriz:

$Q_n^* = \frac{X'\Omega X}{n} = \frac{1}{n}\sum w_i XX'$ (3.6.39.)

A partir de (3.6.39.) White demostró que es posible obtener un estimador apropiado para la varianza los estimadores de mínimos cuadrados incluso si la heterocedasticidad desconocida dependiera de alguna variable independiente. Se busca un estimador para:

$Q_* = \frac{1}{n}\sum \sigma_i^2 XX'$ (3.6.40.)

Donde la diferencia entre (3.6.39.) y (3.6.40.) es que en la primera ecuación w_i es conocida y en la segunda σ_i^2 es desconocida tal como se asume que debe ser para aplicar la corrección de White, que por cierto White logró demostrar que bajo condiciones generales el estimador es:

$S_0 = \frac{1}{n}\sum \mu_i^2 XX'$ (3.6.41.)

El estimador (3.6.41.) es consistente y además se cumple que $plim S_0 = plim Q_*$, lo que quiere decir que el estimador (3.6.41.) es consistente a (3.6.40.) (Greene, 2012) Aclara que en realidad no se estima Q_* sino que se encuentra una función usando los datos de la muestra de tal forma que sea lo más cercana posible a los parámetros poblacionales aumentando el tamaño de la muestra. De esta forma (3.6.41.) converge en (3.6.40.) usando los datos de la muestra, más específicamente usando los errores al cuadrado de la muestra. La justificación del uso de μ_i^2 es que de esta forma se logra la consistencia de los estimadores. Por lo tanto el resultado final se consigue si mantenemos que $plim S_0 = plim Q_*$lo que equivale a

$plim \frac{1}{n}\sum \mu_i^2 XX' = plim \frac{1}{n}\sum \sigma_i^2 XX'$ (3.6.42.)

Entonces el resultado final que es el **estimador de White** es tal como (Greene, 2012) muestra es:

$$Est.Asy.var(\beta) = \frac{1}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}\sum \mu_i^2 XX'\right)\left(\frac{1}{n}X'X\right)^{-1} \quad (3.6.43.)$$

Lo que equivale a:

$$Est.Asy.var(\beta) = n(X'X)^{-1}S_0(X'X)^{-1} \quad (3.6.44.)$$

El concepto parece complicado y de hecho lo es, por ello para que quede libre de dudas se puede resumir todo lo dicho anteriormente en que el estimador de White tiene errores estándares robustos los cuales han sido calculados asumiendo que la varianza del término de error es heterocedástica y además desconocida, por ello haciendo uso de datos muestrales, más específicamente los errores de la regresión que son μ, se ha logrado construir una matriz conocida, la cual aplicando varianzas asintóticas, leyes de números grandes y el teorema del límite central, se demostró que es correcta la estimación de varianza asintóticas usando los errores. Obviamente este ha sido un resumen, por lo que para entender con profundidad el trabajo de White se recomienda leer su artículo original. Sin embargo, actualmente los programas estadísticos incluyen la opción de calcular estimadores robustos de White y STATA no es la excepción, posteriormente se demostrara cómo usarlos.

Después de haber visto estas formas de corregir la heterocedasticidad, uno podría preguntarse ¿Qué medida correctiva emplear? Si bien es una pregunta difícil de responder, pues no existe un consenso claro sobre cual modelo correctivo emplear. La respuesta sería un rotundo **depende**. La heterocedasticidad en un modelo puede ser o no ser conocida, si se puede conocer con exactitud entonces parecería sensato elegir MCP caso contrario los errores robustos, todo se ajusta al modelo planteado, se podría tomar en cuenta el tamaño de la muestra, la cual si es grande la corrección de White podría ser preferible a la de MCP. (Gujarati & Porter, 2010) Mencionan que la presencia de heterocedasticidad no es razón suficiente para desechar el modelo y volver a plantearse otro con otras variables ya que si bien esta puede ser causada por un error de especificación también puede ser causada por datos atípicos. Es en estos modelos cuando el uso de **criterios de información y el tamaño de la varianza del error pueden ser decisivos para elegir un modelo**. Por lo general los investigadores usan ambas medidas de corrección. Un ejemplo de cómo utilizar ambos métodos podemos verlo en el

programa STATA. En este programa estadístico existe el comando ***regress*** el cual combinado con $[weight = 1/X]$ y su opción ***robust***, se está indicando a STATA que pondere a acorde a X y además que utilice los errores robustos. Sea como sea, está en función de la teoría económica, el modelo planteado y el juicio que tenga el investigador para elegir el modelo más apropiado para corregir la presencia de heterocedasticidad. Para acabar esta parte, (Gujarati & Porter, 2010) Mencionan que la aplicación se logaritmos a cada parte del modelo también corrige la heterocedasticidad, esta práctica es común en variables monetarias y sirve para encontrar la elasticidad y los logaritmos son efectivos porque acortan los datos atípicos, es más, la aplicación de logaritmos se utiliza en la econometría de series de tiempo para generar series estacionarias.

3.6.2. Test y métodos correctivos de multicolinealidad.

Como se dijo anteriormente, la multicolinealidad es la violación al supuesto de independencia, el cual plantea que las regresoras tienen cierto grado de dependencia lineal entre ellas, que puede ser perfecta o imperfecta.

Pero esta característica es propia de las variables económicas usadas en los modelos econométricos, entonces ¿Cómo tratar un problema que es característico a la naturaleza de las variables estudiadas? La solución más simple y anticipada sería retirar las variables que producen multicolinealidad en el modelo. Pero retirar variables podría ocasionar más problemas de lo que se tenía en un inicio. A continuación se mostrarán algunos métodos para detectar la existencia de multicolinealidad.

3.6.2.1. Diagnóstico de multicolinealidad.

La multicolinealidad en los modelos es complicada de detectar, a pesar de la existencia de algunas pruebas de hipótesis para detectarla. Por ello, se presentan algunos indicios de la existencia de multicolinealidad.

- **Análisis de la matriz de datos para las variables explicativas.**

El término multicolinealidad fue introducido por Ragnar Frisch en 1934 quien fue un economista noruego que contribuyó no solo a la econometría sino también a la macroeconomía. En su libro “Análisis de confluencia estadística mediante sistemas regresivos integrales” logró diferenciar la presencia de multicolinealidad y los errores de medición, pues según (Núñez Z., 2007) Ambas tienen las mismas consecuencias en un modelo cuando están presentes.

(Núñez Z., 2007) Explica que el rango de la matriz *X* que es la que contiene los datos de las regresoras, debe ser igual al número de regresoras. Esto se escribe como:

$p(X) = k$ (3.6.45.)

Lo que (3.6.45.) quiere decir es que las columnas de la matriz *X* que son el número de variables explicativas en este ejemplo son independientes linealmente entre ellas. Sin embargo, cuando esto no se cumple entonces (3.6.45.) se escribe como:

$p(X) < k$ (3.6.46.)

(3.4.46.) expresa que una columna es la combinación lineal de otra columna de la matriz, por lo que se está cometiendo una infracción al rango de la matriz establecido anteriormente. Hay que tomar en cuenta que (Núñez Z., 2007) Excluye la primera columna la cual está formada solamente por *1* que hace referencia al intercepto.

Por lo que cuando esto sucede, una forma de detectar la multicolinealidad perfecta es que la matriz $(X'X)$ no pueda invertirse porque el determinante es cero. Y cuando es cercano a cero entonces estamos ante un caso de cuasimulticolinealidad también llamada multicolinealidad imperfecta.

Por lo tanto, una forma de detectar la presencia o no de multicolinealidad en el modelo es calculando la determinante de la matriz $(X'X)$, si $|X'X| = 0$ entonces es seguro que existe multicolinealidad perfecta, si por el contrario $|X'X| \simeq 0$ entonces la multicolinealidad imperfecta está presente en el modelo. (Núñez Z., 2007) Indica que en ambos casos no se puede obtener buenos estimadores de MCO.

- **Regresiones auxiliares.**

Este método de verificar la existencia de multicolinealidad puede ser un tanto agobiante sobre todo si el modelo especificado tiene muchas variables explicativas.

Este método de detección de multicolinealidad se basa en el hecho de comparar el coeficiente de determinación y el estadístico *F* calculado que se usa para la prueba de hipótesis sobre la significancia global de varios modelos, que son el modelo original y modelos donde las variables empleadas sean las regresoras. (Gujarati & Porter, 2010) Mencionan el uso de la **regla de práctica de Klein**, la cual considera la existencia de un serio problema de multicolinealidad si el coeficiente de determinación del modelo auxiliar es más alto que el del modelo original.

Otra forma de emplear a regresiones auxiliares para detectar la presencia de multicolinealidad en el modelo es usando el **efecto de R^2 de Theil.** (Galán F., y otros, 2016) Detallan que mediante la siguiente fórmula se calcula:

$$R^2Theil = R^2 - [\sum R^2 - R_i^2] \text{ (3.6.47.)}$$

Donde R^2 es el coeficiente de determinación de la regresión original mientras que R_i^2 es el coeficiente de determinación de la regresión auxiliar, donde si $R^2\ Thiel$ fura nulo la multicolinealidad no estaría presente en el modelo, cuanto más grande sea $R^2\ Thiel$, mayor será el problema de la multicolinealidad. Algo parecido ocurre con la **prueba *F* para multicolinealidad**, donde según (Gujarati & Porter, 2010) Se realizan regresiones auxiliares donde cada regresora se toma como variable explicada sobre las demás regresoras, se toman sus respectivos coeficientes de determinación, y cada uno será contrastado mediante la siguiente prueba de hipótesis:

$$H_0: No\ existe\ multicolinealidad$$

$$H_0: Existe\ multicolinealidad$$

Y se contrasta mediante el siguiente estadístico calculado: $Fc = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)}$ y sigue la siguiente distribución $F_{\alpha,n-k+1}^{k-2}$, donde *k* es el número de regresoras incluido el intercepto del modelo auxiliar, *n* es el tamaño de la muestra y R_i^2 es el coeficiente de determinación de cada modelo auxiliar y la regla de decisión es: si el estadístico calculado supera al tabulado entonces la variable regresora la cual ha sido tomada como variable dependiente en el modelo auxiliar es generadora de multicolinealidad. No obstante, (Gujarati & Porter, 2010) Recomiendan aplicar la regla de Klein el cual tiene la misma capacidad para determinar la existencia de multicolinealidad.

Para que quede claro cómo se detecta multicolinealidad usando modelos auxiliares, se presentará un ejemplo recogido de (Galán F., y otros, 2016) Donde se tiene el siguiente modelo especificado:

$$lcpr_t = \widehat{\beta_1} + \widehat{\beta_2} lrqr_t + \widehat{\beta_3} lypdr_t + \widehat{\beta_4} ltcr_t + \widehat{\mu_t} \text{ (3.6.48.)}$$

Donde:

- $lcpr_t$: Es el logaritmo del consumo privado real en miles de millones de pesos de 1993.

- $lrqr_t$: Es el logaritmo de la riqueza real, el cual ha sido calculado entre el agregado monetario M4 entre el IPC.
- $lypdr_t$: Es el logaritmo del ingreso nacional disponible real en miles de millones de pesos de 1993.
- $ltcr_t$: Es el logaritmo del tipo de cambio real.

El modelo (3.6.47.) al ser estimado mediante MCO se obtiene los siguientes resultados:

$$lcpr_t = 1.90 + 0.15lrqr_t - 0.03lypdr_t + 0.71ltcr_t + \widehat{\mu_t}, \quad \text{Cuyo} \quad R^2 = 0.97 \quad (3.6.49.)$$

Para aplicar la regla de Klein, los autores del modelo consideran el siguiente modelo auxiliar:

$$lypdr_t = \widehat{\alpha_1} + \widehat{\alpha_2}lrqr_t + +\widehat{\alpha_3}ltcr_t + \widehat{e_t} \quad (3.6.50.)$$

El cual arroja los siguientes resultados:

$$lypdr_t = 7.89 + 0.45lrqr_t + +0.02ltcr_t + \widehat{e_t}, R_a^2 = 0.93 \quad (3.6.51.)$$

El coeficiente de determinación del modelo original (3.6.48.) es 0.97 mientras que el coeficiente de determinación el modelo auxiliar (3.6.50.) es 0.93. Si aplicamos la regla práctica de Klein podríamos asumir la existencia de multicolinealidad en el modelo original a pesar que el coeficiente de determinación del modelo auxiliar no sea mayor al del original, ya que estos valores son muy cercanos. Veamos ahora cómo se aplica la regla de $R^2\ Theil$ en el modelo original. Estas son sus modelos auxiliares:

$$lcpr_t = \widehat{\theta_1} + \widehat{\theta_2}lrqr_t + \widehat{\theta_3}ltcr_t + \widehat{v_t} \quad (3.6.52.)$$

$$lcpr_t = \widehat{\theta_1} + \widehat{\theta_2}lrqr_t + \widehat{\theta_3}lypdr_t + \widehat{v_t} \quad (3.6.53.)$$

$$lcpr_t = \widehat{\theta_1} + \widehat{\theta_2}ltcr_t + \widehat{\theta_3}lypdr_t + \widehat{v_t} \quad (3.6.54.)$$

Y estos son los resultados de cada modelo auxiliar respectivamente:

$$lcpr_t = 7.50 + 0.48lrqr_t - 0.02ltcr_t + \widehat{v_t}, \text{ Cuyo } R_1^2 = 0.9424 \quad (3.6.55.)$$

$$lcpr_t = 1.81 + 0.16lrqr_t + 0.70lypdr_t + \widehat{v_t}, \text{ Cuyo } R_2^2 = 0.9737 \quad (3.6.56.)$$

$$lcpr_t = -0.32 - 0.06ltcr_t + 1.02lypdr_t + \widehat{v_t}, \text{ Cuyo } R_3^2 = 0.9677 \quad (3.6.57.)$$

Para calcular el $R^2\ Theil$ se sigue la fórmula:

$$0.9744 - (0.9744 - 0.9424) - (0.9744 - 0.9737) - (0.9744 - 0.9677) = 0.935 \quad (3.6.58)$$

El resultado indica la existencia de multicolinealidad en el modelo, ya que está muy cercano al coeficiente de determinación del modelo original. Finalmente, revisemos el contraste mediante la prueba F, donde el estadístico calculado: $Fc = \frac{\frac{0.93}{3-2}}{\frac{1-0.93}{95-3+1}} = 1235.57$ cuya distribución es $F^{3-2}_{0.05,95-3+1} = 3.94$, podemos observar que el estadístico Fc calculado es mayor al estadístico F crítico, por lo que se asume la existencia de multicolinealidad en el modelo. Entonces mediante las regresiones auxiliares podemos llegar a la conclusión que existe multicolinealidad en el modelo.

- **Número de condición.**

(Uriel & Aldás, 2005) Afirman que esta detección de multicolinealidad, es la más apropiada en tiempos modernos. Inicialmente fue planteado por Rachudel en 1981 y perfeccionado por Belsley en 1980 y 1982.

Este método de detección se basa en que el número de condición $k(X)$ es igual a la raíz cuadrada de la razón entre la raíz característica máxima y la mínima de la matriz $(X'X)$ donde al ser kxk se obtienen k raíces características. Sigue la siguiente formula:

$$k(X) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad (3.6.59.)$$

(Uriel & Aldás, 2005) Explican que el número de condición mide la sensibilidad de las estimaciones de mínimos cuadrados ante pequeños cambios en los datos. La multicolinealidad se detecta cuando el valor calculado es superior a 30, aunque algunos autores recomiendan que cuando es superior a 20 ya se está presentando problemas de multicolinealidad. Este método de detección puede señalar la regresora que genera problema de multicolinealidad, posteriormente se explicará un ejemplo para que quede libre de dudas.

- **Factor de inflación (FIV) y tolerancia de la varianza (TOL).**

Este método de detección es el método más popular y utilizado para detectar la multicolinealidad, se basa en realizar regresiones auxiliares y tomar en cuenta el

coeficiente de determinación de cada regresión auxiliar en la que se toma una regresora como la dependiente y se hace la regresión sobre las demás regresoras, posteriormente se toma el coeficiente de determinación de cada una y se calcula el FIV con la fórmula:

$$FIV = \frac{1}{(1-R_a^2)} \text{ (3.6.60.)}$$

(Hanke & Wichern, 2006) Explican que cuando el FIV se acerca a 1 entonces no se puede sugerir la existencia de multicolinealidad, de hecho cuando FIV se acerca a 1 las variables son estables y los datos o variables agregados o sacados del modelo no afectan en gran medida a los estadísticos *t*, por otro lado cuando se aleja de 1, entonces la variable empieza a dejar de ser estable y los errores estándares y los estadísticos *t* empiezan a cambiar de forma notoria cuando se agregan o quitan datos o variables del modelo. Por último, cuando ya está muy cercano a 10 o en su defecto supera a 10, entonces la variable explicativa no solo es inestable sino que es redundante en el modelo especificado y se podría considerar ser quitado del modelo, pero tal como señala (Wooldrige, 2009) Esto puede ocasionar un sesgo de especificación, por lo que se debería proceder con cuidado.

Por otro lado, se tiene al factor de tolerancia, el cual es definido como la inversa del factor de inflación de varianza según (Gujarati & Porter, 2010). Siendo su fórmula:

$$TOL = \frac{1}{FIV} = (1 - R_a^2) \text{ (3.6.61.)}$$

Donde si TOL se acerca a 0 entonces el problema de multicolinealidad estará fuertemente presente en el modelo econométrico especificado. Sin embargo, (Gujarati & Porter, 2010) Manifiestan que la incorrecta estimación de los errores de regresión no tiene que estar ocasionado necesariamente por un FIV muy elevado, ya que si recordamos que otros problemas como la heterocedasticidad también puede ocasionar el mismo problema.

- **Matriz de correlación.**

Esta es otro método de detección muy común y muy frecuente cuando se quiere detectar multicolinealidad. La correlación alta entre las variables explicativas muestra la existencia de multicolinealidad en el modelo, sin embargo el problema es que la alta correlación no necesariamente indica multicolinealidad en el modelo, ya que al mostrar la correlación solamente entre dos variables no es suficiente para determinar la existencia o no de multicolinealidad. Por ejemplo, suponga el modelo econométrico:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \mu_i \text{ (3.6.62.)}$$

El cual tiene la siguiente dependencia lineal entre las regresoras: $X_{2i} = \alpha_3 X_{3i} + \alpha_4 X_{4i}$, por lo que la matriz de correlación en palabras de (De Grange C., 2005) No podría detectar la correlación existente entre X_{2i} con X_{3i} y X_{4i}. Lo que esto quiere decir, es que la matriz de correlación no podría detectar combinaciones de dependencia lineal complejas.

- **Trampa de la variable ficticia.**

Una variable ficticia, en términos sencillos, es una variable que no muestra información cuantitativa sino una cualidad o característica y solo puede tener dos valores posibles, también **son conocidas como variables Dummy, variables dicotómicas, variables binomiales**, etc. Por lo general se emplean el *0* y *1*, donde *0* denota la carencia o el incumplimiento de una característica o condición y *1* denota el cumplimiento de la característica o condición. Un ejemplo muy común es la variable ***sexo*** donde podría tomar el valor de *1* para mujeres y *0* para hombres; otro ejemplo sería la variable ***vivpropia*** que toma el valor de *1* cuando la vivienda donde vive la familia es propia, por otro lado podría tomar el valor de *0* si la vivienda no es propia de la familia. En realidad, estos valores son totalmente arbitrarios y se puede utilizar cualquier valor, pero por lo general se usan los valores *0* y *1* en la teoría econometría y en los programas estadísticos.

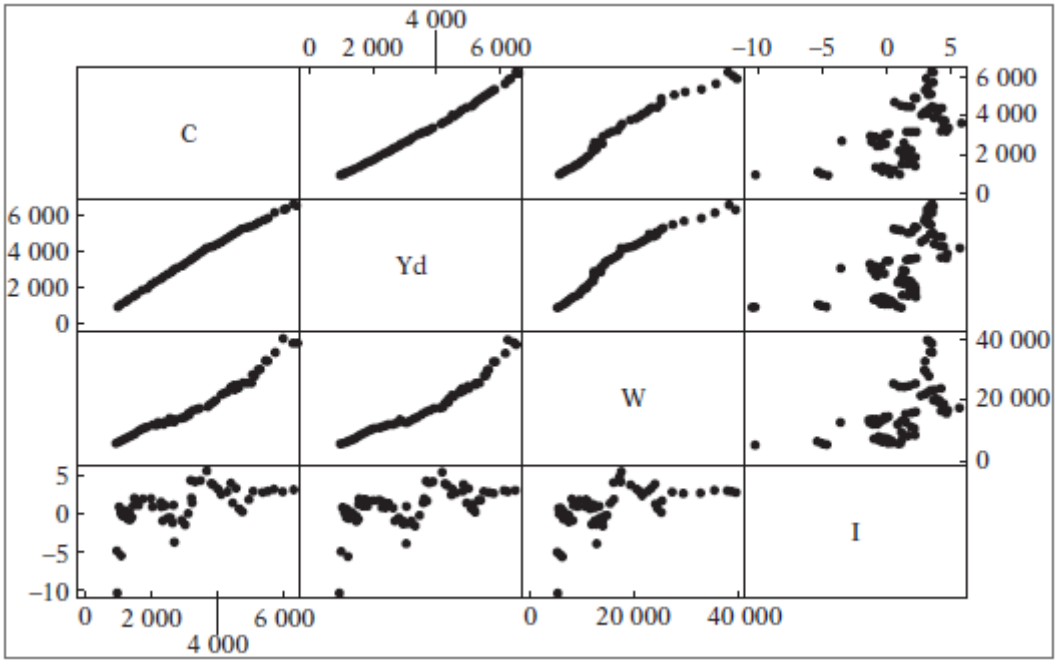
Por lo general, son empleadas en modelos microeconométricos donde se espera capturar el efecto de una característica sobre la variable dependiente. (Stock & Watson , 2012) Explican que el uso excesivo de variables ficticias puede ocasionar multicolinealidad mediante su ejemplo. Suponga que un modelo econométrico busca explicar los gastos monetarios de las familias en una ciudad muy grande donde se divide en tres partes: norte, sur y este, por lo que con el fin de capturar como la ubicación de la vivienda influye en el gasto monetario se crean tres variables dicotómicas: $norte, sur\ y\ este$, al estimar mediante MCO se incluyen las variables ficticias en el modelo. (Stock & Watson , 2012) Indican que si incluyéramos las tres variables sin excluir el intercepto entonces estaríamos cayendo sin lugar a dudas en la multicolinealidad debido a la trampa de las variables ficticias. Por lo que para evitar caer en multicolinealidad por el uso excesivo de variables ficticias entonces se debería o excluir una variable ficticia de las tres o excluir el intercepto; no obstante se recomienda

la exclusión de una variable ficticia. Además según (Stock & Watson , 2012) La multicolinealidad sería perfecta, lo que significa que ni siquiera se podría estimar el modelo ya que las tres variables indican una característica en común, la cual es la ubicación de la vivienda dentro de una ciudad.

- **Gráfica de dispersión.**

Este es posiblemente uno de los métodos de detección menos usados para detectar la presencia de multicolinealidad en el modelo. Similar a los gráficos que conforman los métodos informales para detectar heterocedasticidad y autocorrelación en el modelo, la gráfica de dispersión entre las variables explicativas muestra cómo están correlacionadas las regresoras.

(Gujarati & Porter, 2010) Muestra un ejemplo de esto.



Grafica 3.22. Grafica de dispersión entre las variables independientes y la dependiente.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

(Gujarati & Porter, 2010) Especifican a la variable consumo C como la variable dependiente del modelo y a las variables ingreso personal disponible real Yd, riqueza real W y a la tasa de interés real I. Para interpretar la gráfica, primero ignoremos las gráficas que están por encima de la diagonal y solo fijemos en las gráficas que están por debajo de la diagonal. Identificamos como variables correlacionadas a aquellas que muestran un patrón claro. Por ejemplo, las variables ingreso y riqueza muestran una tendencia

ascendente lo cual podría indicar que estas variables están correlacionadas, por otro lado, la variable tasa de interés no muestra un patrón claro por lo que es menos probable que esté correlacionada con las demás regresoras.

- **Un coeficiente de determinación demasiado alto.**

Por lo general, se espera que el coeficiente de determinación sea lo más alto posible. Sin embargo, esto puede ser un indicio de la existencia de multicolinealidad en el modelo, sobre todo cuando este es demasiado alto y las variables regresoras no tienen significancia individual. Si este fuese el caso, entonces cabría la sospecha que el modelo tiene multicolinealidad.

- **Examen de correlaciones parciales o Test de Farrar-Glauber.**

Este es el método de detección menos usado debido a que es difícil de entender, requiere un procedimiento largo y ha recibido fuertes críticas. Para empezar, cabe recalcar que este test se sostiene en tres pilares para determinar la multicolinealidad, según (Farrar & Glauber R., 1967) Estos son:

- Prueba de presencia y gravedad de multicolinealidad.
- Prueba de dependencia de variables particulares.
- Examinar el patrón de interdependencia entre las regresoras del modelo.

La primera acepción la cual se refiere a esta prueba de multicolinealidad como tal, se contrasta mediante el uso de pruebas de hipótesis. (Pérez L., 2012) Muestra que para detectar la existencia de multicolinealidad mediante la prueba de hipótesis de Farrar-Glauber, se debe calcular el siguiente estadístico calculado:

$$G = -\left[n - 1 - \frac{(2n+5)}{6}\right] \log(|R|) \quad (3.6.63.)$$

El cual sigue la siguiente distribución:

$$G \sim X^2_{\frac{k(k-1)}{2}} \quad (3.6.64.)$$

Donde *n* es el tamaño muestral y *k* es el número de estimadores y *k-1* es el número de regresores y |R| es la determinante de la matriz de correlación. Y se plantea la siguiente prueba de hipótesis:

$$H_0: No\ existe\ multicolinealidad$$

$$H_0: Existe\ multicolinealidad$$

Donde la regla de decisión es igual a las anteriores, según el nivel de significancia si el estadístico calculado supera al tabulado entonces se rechaza la hipótesis nula y se acepta la existencia de multicolinealidad. Pese a la conveniencia de detectar la multicolinealidad mediante una prueba de hipótesis, en realidad este método ha sido fuertemente criticado. El problema con este método para contrastar la existencia de multicolinealidad lo explica (Gujarati & Porter, 2010), a través de la teoría que propone C. Robert, quien demostró que la matriz de correlación no es lo suficientemente convincente para demostrar la multicolinealidad porque no puede medir complejas combinaciones entre las regresoras. Es por esto, que se prefiere evitar el uso de este contraste y es mejor usar los otros indicios para verificar los mismos puntos en los que se centran sus tres pilares. De hecho (Wooldrige, 2009) Aconseja no seguir el contraste de hipótesis ya que al no existir un consenso sobre cuando la correlación se le puede considerar demasiado elevado entonces no puede determinar con exactitud la presencia de multicolinealidad en el modelo.

3.6.2.2. Tratamiento de la multicolinealidad.

Después de mostrar cómo es posible detectar la multicolinealidad, queda hacerse la pregunta: ¿Cómo resolver un modelo con problema de multicolinealidad? Al igual que su detección, el tratamiento que se le debe dar a un modelo con multicolinealidad debe ser ejecutado siguiendo el juicio crítico, al igual que la heterocedasticidad y otras violaciones a los supuestos de MCO. Veamos algunos métodos para corregir la multicolinealidad en un modelo.

- **Retirar variables explicativas.**

Este es el método más fácil para corregir la multicolinealidad, básicamente lo que hace es identificar a las variables regresoras que la causan y retirarlas, el problema surge cuando al momento de retirar una regresora se corre el riesgo de generar un sesgo de especificación por omisión de regresora relevante, también llamado sesgo de especificación por subajuste. Acorde a (Núñez Z., 2007) La exclusión de una regresora no solo debe hacerse con la intención de corregir la multicolinealidad, sino que además es necesario una justificación por parte de la teoría económica. (Wooldrige, 2009) Propone un ejemplo interesante sobre el retiro de variables explicativas en la siguiente cita.

> *"Suponga que se desea estimar el efecto de diversas categorías de gastos en la educación sobre el desempeño de los estudiantes. Es posible que los gastos en sueldos para los profesores, material didáctico, deporte, etc., estén fuertemente correlacionados: las escuelas con mejor situación económica gastan más en todo y las pobres gastan menos en todo. Es claro que es difícil estimar el efecto de una determinada categoría de gastos sobre el desempeño de los estudiantes cuando es poca la variación en una categoría que no puede ser explicada por la variación en las otras categorías (...)"* (Wooldrige, 2009)

La cita anterior ilustra cómo el afán de capturar el efecto de categorías específicas puede conducir a la multicolinealidad porque suelen estar altamente correlacionadas entre ellas. No obstante, al mismo tiempo también menciona como la ausencia de correlación entre las regresoras genera problemas de estimación. En síntesis, la cita menciona que una correlación alta entre las regresoras genera problemas como si hubiera poca correlación entre las regresoras, por lo que ¿realmente conviene excluir regresoras? Como todo en la economía, la respuesta sería un rotundo **depende**: este método de corrección sería más conveniente seguirse cuando la multicolinealidad que presenta el modelo es perfecta, puesto que no permite la estimación exacta de los estimadores por la presencia de la **influencia combinada**.

- **Información a priori.**

Ya que la retirada de una regresora debe ser justificada por la teoría económica, lo que (Gujarati & Porter, 2010) Sugieren que en vez de justificar su exclusión entonces justifiquemos su uso mediante la teoría económica y proponen un ejemplo de un uso correcto de este método de corrección. (Gujarati & Porter, 2010) Especifican el siguiente modelo econométrico:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i \text{ (3.6.65.)}$$

Donde Y_i: consumo, X_{2i}: ingreso y X_{3i}: riqueza, además (Gujarati & Porter, 2010) Especifican a priori que $\beta_3 = 0.10\beta_2$ entonces podremos estimar (3.6.65.) transformándolo en:

$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + \mu_i \text{ (3.6.66.)}$$

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i \text{ (3.6.67.)}$$

Según (Gujarati & Porter, 2010) Se puede estimar β_3 a partir de β_2, pero este método correctivo trae problemas, puesto que su aplicación implica conocer la dependencia lineal de una variable sobre otra. (Gujarati & Porter, 2010) Recomiendan revisar trabajos anteriores para determinar la información a priori, y ejemplifican con la función de Cobb-Douglas que este método correctivo es el idóneo cuando determinamos información a priori brindada por la teoría económica. No obstante, no en todos los modelos econométricos se logrará obtener esta información, y además al igual que con la exclusión de regresoras, esto debe estar justificado por la teoría económica y comprobado mediante una prueba de restricción que tal información a priori es válida para corregir el modelo especificado. (De Grange C., 2005) Llama a este método como **la imposición de restricciones sobre los parámetros.**

- **Transformación de variables.**

Tomando en cuenta que los datos de series temporales suelen estar correlacionados entre sí por la tendencia creciente o decreciente, lo que (Pérez L., 2012) Sugiere es tomar las diferencias de cada variable y realizar la regresión con ellas. (De Grange C., 2005) Explica que esto es válido ya que las variables con datos de series temporales suelen no ser estacionarias y por eso se genera multicolinealidad, pero ¿a qué se refiere la estacionariedad?, en términos sencillos, una variable es estacionaria cuando no tiene crecimiento o decrecimiento en su periodo determinado, lo que implica que la media y la varianza sea constante en el periodo dado; el término es más complejo y profundo de lo que parece y ameritaría otra guía concerniente a las variables estacionarias, como último dato no debe ser confundido con la estacionalidad que es un componente de las series temporales y hace referencia a las oscilaciones ocurridas en periodos menores o iguales a un año.

Retomando el tema principal, (Gujarati & Porter, 2010) Muestran cómo se especifica la transformación de variables por diferencias:

$$Y_t - Y_{t-1} = \beta_2(X_{2t} - X_{2t-1}) + \beta_3(X_{3t} - X_{3t-1}) + \cdots + \beta_k(X_{kt} - X_{kt-1}) + (\mu_t - \mu_{t-1}) \quad (3.6.68.)$$

Sin embargo, la transformación por diferencias también trae consigo algunos puntos en el que se podría cuestionar su efectividad, algunas de estas sería que este método correctivo solo sería aplicable a los datos de series temporales y además (Pérez

L., 2012) Advierte que si bien esto puede corregir la multicolinealidad también puede ser el causante de la autocorrelación.

Ya que la transformación de variables por diferencias es exclusiva de los datos de series temporales, entonces para los de corte transversal (Uriel & Aldás, 2005) Recomienda usar la transformación por ratios o de razón, el cual tiene cierto parecido con los métodos correctivos que se utilizar para tratar la heterocedasticidad en un modelo. Este método se basa en identificar la variable explicativa que tenga mayor correlación y dividir a cada variable del modelo entre la regresora identificada. El modelo transformado seria especificado de la siguiente manera:

$$\frac{Y_i}{X_{2i}} = \beta_1 \left(\frac{1}{X_{2i}}\right) + \beta_2 + \beta_3 \left(\frac{X_{3i}}{X_{2i}}\right) + \cdots + \beta_k \left(\frac{X_{ki}}{X_{2i}}\right) + \left(\frac{\mu_i}{X_{2i}}\right) \text{ (3.6.69.)}$$

No obstante, al igual que la transformación por diferencias, esta transformación también debe tratarse con cuidado, ya que al especificar que $\left(\frac{\mu_i}{X_{2i}}\right)$ se está asumiendo de forma indirecta que la varianza del término de error depende de X_{2i}^2 lo cual se denota como $E(\mu^2) = \sigma^2 X_{2i}^2$, el problema en sí, es que esto no podría ser cierto y de ser así entonces el error tiene varianza heterocedástica por lo que en vez de corregir la multicolinealidad se podría inducir al modelo a la heterocedasticidad. Es recomendado entonces que, de aplicarse este método correctivo, debería hacerse un test de heterocedasticidad y de verificar que los errores no son homocedásticos, entonces descartar este método correctivo.

- **Método de componentes principales (MCP).**

En palabras de (Uriel & Aldás, 2005) Este método permite pasar a un nuevo conjunto de variables que gozan de la ventaja de estar incorrelacionadas entre sí y que puede ordenarse acorde a la información que llevan incorporada, (De Grange C., 2005) Complementa lo anterior afirmando que este método, el cual es una técnica estadística, permite reducir el número de variables regresoras procurando que no se pierda mucha información en el proceso y a las nuevas variables las denomina **componentes principales.**

(Pérez L., 2005) Establece que la importancia de este método radica en que el MCP describe sintéticamente la estructura e interrelaciones de las variables originales a partir de los componentes que se obtienen. Comenzamos la explicación determinando que

en un modelo existen n observaciones con p variables que son: $X_{1i}, X_{2i}, X_{3i}, \dots, X_{pi}$, entonces el primer componente se calcula como una combinación lineal de las demás variables originales. Por lo tanto, se expresa como:

$$Z_{1i} = a_{11}X_{1i} + a_{12}X_{2i} + a_{13}X_{3i} + \cdots + a_{1k}X_{pi} \quad (3.6.70.)$$

Lo que equivale a expresarlo en su forma matricial:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1i} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & X_{3i} & \dots & X_{p1} \\ X_{12} & X_{22} & X_{32} & \dots & X_{p2} \\ & \vdots & \vdots & \ddots & \vdots \\ X_{1i} & X_{2i} & X_{3i} & \dots & X_{pi} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1i} \end{bmatrix} \rightarrow Z = Xa \quad (3.6.71.)$$

El primer componente que se obtiene debe tener la varianza máxima que está sujeta a la restricción: la suma de los pesos (a) al cuadrado es igual a 1 según la condición de identificabilidad, de esta manera se determina que la varianza del primer componente que tiene una media igual a 0, viene dado por:

$$var(Z_i) = \frac{\sum Z_{1i}^2}{n} = \frac{1}{n}Z'Z = \frac{1}{n}a'X'Xa = a'\left[\frac{1}{n}X'X\right]a \quad (3.6.72.)$$

En este punto cabe aclarar que el primer componente se calcula de modo que Z_1 tenga una varianza que sea máxima y que además esté sujeta a la restricción ${a_1}'a_1 = 1$ eligiendo el a_1 que cumpla con lo anterior, el segundo componente se calcula eligiendo a a_2 que cumpla con la condición que Z_2 este incorrelacionada con Z_1, y así sucesivamente. De esta manera los componentes $Z_1, Z_2, \dots, Z_q$ están incorrelacionados. Es necesario señalar que el subíndice q no puede ser igual al número de variables originales, ya que este método reduce el número de variables, se tiene que $q < p$.

Prosiguiendo con la explicación, se asume que $\left[\frac{1}{n}X'X\right]$ es la matriz de covarianzas muestral, a lo que se denomina como V, (Uriel & Aldás, 2005) Detallan que esto implica que las variables originales están expresadas en desviaciones respecto a la media. Si fuesen variables tipificadas entonces $\left[\frac{1}{n}X'X\right]$ sería la matriz de correlaciones y se denota con R, pero este no es el caso. Al usar la matriz de covarianzas, (3.6.72.) se transforma en:

$$var(Z_i) = {a_1}'Va_1 \quad (3.6.73.)$$

Al aplicar la restricción: $a'a = 1$ a (3.6.73.) entonces se forma el lagrangiano:

$L = a_1'Va_1 - \lambda(a_1'a_1 - 1)$ (3.6.74.)

Para maximizar el valor del lagrangiano se deriva respecto a a y se obtiene:

$\frac{\partial L}{\partial a_1} = 2Va_1 - 2\lambda Ia_1 = 0$ (3.6.75.)

Reordenando (3.6.75.) queda:

$(V - \lambda I)a_1 = 0$ (3.6.76.)

Donde para que tenga una solución que no será cero, entonces $|V - \lambda I| = 0$, (Uriel & Aldás, 2005) Al resolverse la ecuación $|V - \lambda I| = 0$ se obtienen p raíces características λ, con lo cual se toma al mayor de ellos y con su correspondiente a_{1i} se halla el vector característico asociado a a_1 usando la regla de normalización $a_1'a_1$. Por lo que, las ponderaciones o pesos usados para hallar el primer componente que están representados en (3.6.70.) están representadas en el vector característico asociado a la raíz característica mayor a *V*.

Para obtener las siguientes componentes, partimos desde $Z = Xa$ y la restricción $a'a = 1$, pero ahora se le agregan las restricciones:

$a_h'a_1 = a_h'a_2 = a_h'a_3 = \ ... \ = a_h'a_{h-1} = 0$ (3.6.77.)

Por lo que se deben imponer tantas restricciones adicionales de que el vector característico está asociado a a h-ésima. En otras palabras, los componentes se calculan como una combinación lineal de las variables originales en las que los coeficientes dados por los pesos o ponderaciones son los vectores característicos correspondientes de la matriz *V*.

Aparentemente este método corrige la multicolinealidad y podría ser la mejor opción, ya que se obtendrán variables que no están correlacionadas. Pero ese es justamente el problema de este método correctivo, ya que si recordamos lo que (Wooldrige, 2009) Planteó que la incorrelación o correlación baja de las variables también es nefasta para la correcta estimación. Entonces una pregunta sale a la luz ¿vale la pena este método si al final es posible no obtener una mejor estimación que la que se consigue bajo multicolinealidad? Podría ser que sí, pero se tendría que revisar meticulosamente el modelo con los componentes, ya que una baja correlación también ocasiona problemas de estimación. Además (Galán F., y otros, 2016) Identifican otro problema con respecto a este método correctivo, y es que al no estar correlacionadas entonces no se podría

interpretar ni darle un sentido económico al modelo de regresión. Como siempre el investigador debe usar su juicio crítico para decidir cuál es el mejor modelo que arroje estimadores MELI. Para concluir, (De Grange C., 2005) Recomienda el uso de este método para la detección de datos outliers o atípicos, revisar la hipótesis de distribución normal multivariada, agrupar elementos de la muestra en subgrupos semejantes y reducción de la dimensión en análisis discriminante.

- **Regresiones de cadena.**

Básicamente se basa en convertir la matriz $(X'X)$ en otra matriz parecida la cual es $(X'X + kI)$, siendo una k la constante adecuada, de esta forma se obtiene una buena bondad de ajuste y significancia individual y global. (De Grange C., 2005) Advierte que esta matriz debe tener la menor perturbación posible con el fin que $|X'X|$ sea distinto a 0, por lo que podemos intuir que este método correctivo es más provechoso cuando se aplica a modelos con multicolinealidad perfecta y además no se puede retirar la regresora que la causa.

Pero como los anteriores métodos correctivos, este método puede presentar problemas en el modelo, siendo el más frecuente calcular estimadores sesgados, y peor aún no tener interpretación económica. Por lo que no es recomendada para corregir la multicolinealidad.

3.6.2.3. Relación entre la micronumerosidad y la multicolinealidad.

Luego de todo lo visto, podemos concluir que para corregir la multicolinealidad, por lo general se corren riesgos de generar otros problemas al modelo especificado entonces ¿Realmente se tiene que corregir la multicolinealidad? Después de todo, las variables económicas están correlacionadas entre sí y la poca correlación entre ellas tampoco es válido para obtener buenos estimadores. Por lo que (Gujarati & Porter, 2010) Recogen la posibilidad de **no hacer nada** cuando la multicolinealidad está presente en el modelo econométrico, de hecho explican que en palabras de Blanchard de tal manera que deja entender que la multicolinealidad es causada por una deficiencia de datos lo que se define como **micronumerosidad**, la cual es producto de la imposibilidad a la cual se enfrentan los economistas de recoger una muestra lo suficientemente grande.

La micronumerosidad es un término acuñado por Goldberger a modo de parodia, quien **sostiene como los economistas se han procurado más por plantear métodos**

correctivos para tratar la multicolinealidad presente en los modelos en vez de preocuparse por la muestra empleada para estimar el modelo, más concretamente el tamaño de la muestra empleada. (Wooldrige, 2009) Expone su punto al afirmar que resulta irónico que en las ciencias sociales, como la economía, se recolecta pasivamente una muestra que podría ocasionar estimadores ineficientes por lo que se recolecta más datos. De hecho, Goldberger sostiene que la micronumerosidad presenta las mismas nefastas consecuencias en el modelo como la multicolinealidad, y esto se debe a la poca variabilidad de las series.

De esta manera, se podría concluir que para resolver la multicolinealidad no solo es posible aplicando medidas correctivas, sino además revisando la muestra empleada y aumentar el número de observaciones en la medida de lo posible. La multicolinealidad no es mala en sí, podría catalogarse como "mala" cuando estamos ante una multicolinealidad exacta; de hecho, cuando STATA detecta que una variable es una combinación exacta lo que hace es no tomarla en cuenta para la regresión. Este es un problema muy difícil de entender y más aún de solucionar debido a que es algo que está implícito en la naturaleza de los regresores y además que no existe un consenso generalizado que determine cuando una correlación puede generar problemas de multicolinealidad. Es casi seguro que en los primeros modelos econométricos que los economistas realizan tengan la multicolinealidad presente en ellos.

3.6.3. Test y métodos correctivos de autocorrelación.

Anteriormente, se había explicado que el supuesto de la ausencia de la autocorrelación en el término de error se debe al hecho que esta es una variable aleatoria obtenida de una muestra aleatoria, por lo tanto, acorde a la aleatoriedad de sus valores deben ser independientes entre sí y no seguir ningún patrón ni tendencia. El supuesto de no autocorrelación se representa como $cov(\mu_i, \mu_j) = 0$ o $E(\mu_i * \mu_j) = 0$, ambas son formas válidas de expresarlo.

Sin embargo, las variables económicas suelen tener autocorrelación en sus datos, sobre todo en los datos de series de tiempo. Según (Hanke & Wichern, 2006), los valores de las series de tiempo dependen fuertemente de los valores pasados, siendo este el motivo por el cual muestran tendencias y patrones, por lo tanto, es difícil considerar a una serie temporal como aleatoria. Los datos de corte transversal tampoco están exentos, en su caso

(Gujarati & Porter, 2010) Denominan que la **correlación espacial** ocurre cuando los datos de las entendidas están correlacionadas entre sí.

La autocorrelación en un modelo genera los mismos problemas que la heterocedasticidad como una varianza incorrecta, conclusiones equivocadas sobre las pruebas de significancia de *t* y *F*, un falso coeficiente de determinación y aunque los estimadores estén insesgados dejan de ser eficientes ya que su varianza ya no es mínima.

En esta sección se explicará cómo detectar la autocorrelación y posteriormente como tratar la presencia de autocorrelación en un modelo econométrico. Previamente, se explicarán algunos conceptos para entender con exactitud cómo detectar la autocorrelación y posteriormente ejecutar un método correctivo.

La autocorrelación se le denota como:

$$E(\mu_i * \mu_j) \neq 0 \quad (3.6.78)$$

Donde los subíndices *i* y *j* indican que se tratan de los datos del término de error. Pero ¿Por qué la autocorrelación no genera estimadores MELI? Para entenderlo debemos tener presente que en un modelo de series temporales con autocorrelación los valores del término de error dependen de sus valores pasados. Por lo tanto al tener el siguiente modelo econométrico: $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$, si asumimos que la autocorrelación está presente entonces podemos especificar:

$$\mu_t = p\mu_{t-1} + e_t \quad (3.6.79.)$$

Donde p se le denomina como el **coeficiente de autocovarianza o autocorrelación**, el cual puede tomar valores desde $-1 < p < 1$.

Para que el modelo se considere libre de autocorrelación p debe estar lo más cercano a 0, de esta forma se asume que el término de error no depende de sus valores pasados. Algo importante a notar es que a (3.6.79.) se le conoce como un **proceso autorregresivo de primer orden** o ***AR(1)*** el cual sugiere que el término de error depende de sí mismo en **un periodo rezagado**, sin embargo (Pérez L., 2012) Advierte que en realidad es una generalización, es decir, la mayoría de modelos econométricos con autocorrelación tienen términos de error que siguen un *AR(1)* pero no necesariamente tiene que ser así, en algunos modelos econométricos el término de error puede depender

de sí mismo en dos, tres o *p* periodos rezagados, pero para fines didácticos asumimos que el modelo sigue un *AR(1).* La forma general de *AR(p)* se escribe como:

$$\mu_t = p_1\mu_{t-1} + p_2\mu_{t-2} + p_3\mu_{t-3} + \cdots + p_p\mu_{t-p} + e_t \text{ (3.6.80.)}$$

Por último, el *AR(p)* se define como un proceso en el que una variable depende de sí misma en *p* periodos rezagados más un término de error. Constituye un tema fundamental en la teoría de econometría de series temporales. Retomando el tema de la autocorrelación, (Wooldrige, 2009) Detalla que (3.6.79.) tiene las siguientes propiedades:

$$E(e_t) = 0 \text{ (3.6.81.)}$$

$$var(e_t) = E(e_t^2) = \sigma_e^2 \text{ (3.6.82.)}$$

$$cov(e_t, e_s) = 0 \text{ (3.6.83.)}$$

El cumplimiento de estas propiedades hace que el término estocástico e_t se le denomine como una variable que sigue un **proceso de ruido blanco**, un término muy empleado en la teoría de econometría de series temporales. (Brooks, 2008) Define al ruido blanco como un proceso que no sigue una estructura perceptible, es decir que tiene media y varianza constante y además las observaciones no están correlacionadas entre sí tal como se muestran en (3.6.81.), (3.6.82.) y (3.6.83.). En (3.6.82.), la varianza es constante y no debe ser confundido con la varianza de un término heterocedástico como se mostró previamente, el subíndice *e* ha sido empleado según la teoría propuesta por (Wooldrige, 2009) para diferenciarlo de la varianza del término de error μ_t.

Asumir que e_t es un proceso de ruido blanco implica a asumir que (3.6.79.) es un proceso estacionario entonces podemos argumentar que (3.6.79.) cumple las siguientes propiedades expuestas por (Gujarati & Porter, 2010).

$$E(\mu_t) = pE(\mu_{t-1)} + E(e_t) = 0 \text{ (3.6.84.)}$$

$$var(\mu_t) = p^2 var(\mu_{t-1}) + var(e_t) \text{ (3.6.85.)}$$

(3.6.85.) equivale a:

$$var(\mu_t) = \frac{\sigma_e^2}{1-p^2} \text{ (3.6.86.)}$$

Para realizar la equivalencia conviene tener en cuenta que la $var(\mu_t)$ en un *AR(1)* es igual a $var(\mu_t) = var(\mu_{t-1}) = \sigma^2$, por lo que al reemplazar en (3.6.85.) y despejar μ_t, obtenemos:

$$var(\mu_t) = p^2 var(\mu_{t-1}) + var(e_t) \text{ (3.6.85.)}$$

$$var(\mu_t) = p^2 var(\mu_t) + \sigma_e^2 \text{ (3.6.87.)}$$

$$var(\mu_t) - p^2 var(\mu_t) = \sigma_e^2 \text{ (3.6.88.)}$$

$$var(\mu_t) = \frac{\sigma_e^2}{1-p^2} \text{ (3.6.86.)}$$

Para hallar la covarianza en (3.6.79.) primero multiplicamos a ambos lados por μ_{t-1} y aplicamos esperanza a ambos lados:

$$cov(\mu_t, \mu_{t-1}) = E(\mu_t * \mu_{t-1}) = E(p\mu_{t-1}^2 + \mu_{t-1} e_t) \text{ (3.6.89.)}$$

Al recordar que $E(e_t) = 0$ entonces (3.6.89.) se transforma en:

$$cov(\mu_t, \mu_{t-1}) = E(\mu_t * \mu_{t-1}) = pE(\mu_{t-1}^2) \text{ (3.6.90.)}$$

Y al aplicar $var(\mu_t) = var(\mu_{t-1}) = var(\mu_t) = \frac{\sigma_e^2}{1-p^2}$ entonces (3.6.90) se reescribe como:

$$cov(\mu_t, \mu_{t-1}) = E(\mu_t * \mu_{t-1}) = p\frac{\sigma_e^2}{1-p^2} \text{ (3.6.91.)}$$

(Gujarati & Porter, 2010) Generalizan la expresión (3.6.92.) para un determinado *AR(p)*.

$$cov(\mu_t, \mu_{t-2}) = E(\mu_t * \mu_{t-2}) = p^2\frac{\sigma_e^2}{1-p^2} \text{ (3.6.92.)}$$

$$cov(\mu_t, \mu_{t-3}) = E(\mu_t * \mu_{t-3}) = p^3\frac{\sigma_e^2}{1-p^2} \text{ (3.6.93.)}$$

$$cov(\mu_t, \mu_{t-4}) = E(\mu_t * \mu_{t-4}) = p^4\frac{\sigma_e^2}{1-p^3} \text{ (3.6.94.)}$$

Y así sucesivamente. Finalmente, el coeficiente de autocorrelación se calcula dividiendo la autocovarianza (3.6.91) entre la varianza $\frac{\sigma_e^2}{1-p^2}$:

$$corr(\mu_t, \mu_{t-1}) = p \text{ (3.6.95.)}$$

Luego de todo lo visto, se puede inferir cual es el problema que un modelo tenga autocorrelación en el término de error.

Recordemos que en un modelo econométrico hemos supuesto que la varianza del término de error es constante y no existe autocorrelación en los datos del término de error, podemos definirlos respectivamente como $E(\mu_i^2) = \sigma^2$ y $E(\mu_i * \mu_j) = 0$ y expresarlos en su forma matricial como $E(\mu\mu') = \sigma^2 I$. Pero si un modelo econométrico tiene autocorrelación, entonces la matriz de *cov-var* tiene los siguientes elementos, según (Greene, 2012).

$$E(\mu\mu') = \sigma^2\Omega = \frac{\sigma_e^2}{1-p^2}\begin{bmatrix} 1 & p & p^2 & p^3 & \cdots & p^{n-1} \\ p & 1 & p & p^2 & \cdots & p^{n-2} \\ p^2 & p & 1 & p & \cdots & p^{n-3} \\ p^3 & p^2 & p & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & p & p \\ p^{n-1} & p^{n-2} & p^{n-3} & \cdots & p & 1 \end{bmatrix} \text{(3.6.96.)}$$

Por lo que, en respuesta de la pregunta: ¿Por qué la autocorrelación no genera estimadores MELI? La matriz (3.6.96.) es la causante de los problemas que genera la autocorrelación, ya que en un modelo econométrico sin autocorrelación la varianza de los estimadores es $var(\hat{\beta}) = \sigma^2 X'X$, pero en presencia de autocorrelación la varianza de los estimadores es:

$$var(\hat{\beta}) = \sigma^2[(X'X)^{-1}(X'\Omega X)^{-1}(X'X)^{-1}] \text{ (3.6.97.)}$$

La expresión (3.6.97.) es muy parecida a la varianza de los estimadores bajo heterocedasticidad representada en (3.6.62.), no obstante, no deben ser tomadas como la misma expresión, ya que en (3.6.97.) se está asumiendo que la varianza del término de error es constantes, en cambio en (3.6.62.) es heterocedástica. Por otro lado, el símbolo Ω empleado en (3.6.97.) y (3.6.62.) tampoco representan las mismas expresiones como ya se ha mostrado anteriormente. Lo que sí está claro es que en la autocorrelación la varianza de los estimadores no es eficiente por lo que trae consigo problemas como falsas conclusiones sobre las pruebas *t* y *F* de los estimadores y del modelo, una incorrecta estimación por intervalos, un error de regresión incorrectamente estimado y un desacertado coeficiente de determinación.

Para finalizar esta parte de la explicación, el coeficiente *p* sigue una restricción según (Gujarati & Porter, 2010) $|p| < 1$, de esta forma se asegura que (3.6.79.) es un

proceso estacionario cuyo término estocástico es un proceso de ruido blanco. Si *p* fuese igual a 1 entonces las varianzas y las covarianzas no podrían ser definidas, por ello es que debe seguir la restricción.

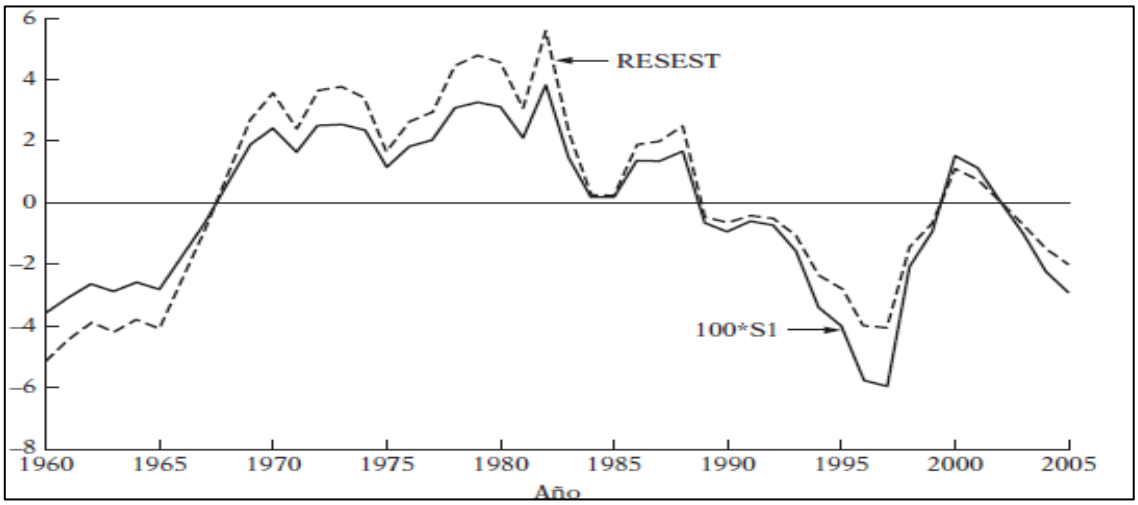
3.6.3.1. Métodos para detectar autocorrelación.

3.6.3.1.1. Métodos informales.

Al igual que con la heterocedasticidad, la autocorrelación también tiene métodos informales en los cuales se utilizan los gráficos para saber el comportamiento de los residuos del modelo. (Gujarati & Porter, 2010) Señalan que se pueden utilizar los **gráficos secuenciales de tiempo** y **gráficos de los residuos estandarizados.** Básicamente en ambos se utilizan los residuos y se grafican respecto al tiempo, la diferencia está en que, los primeros se usan los residuos y en el segundo los residuos estandarizados, los cuales se hallan dividiendo los residuos del modelo sobre el error de la regresión:

$$\hat{\mu}_{est} = \frac{\hat{\mu}_i}{\hat{\sigma}} \text{ (3.6.98.)}$$

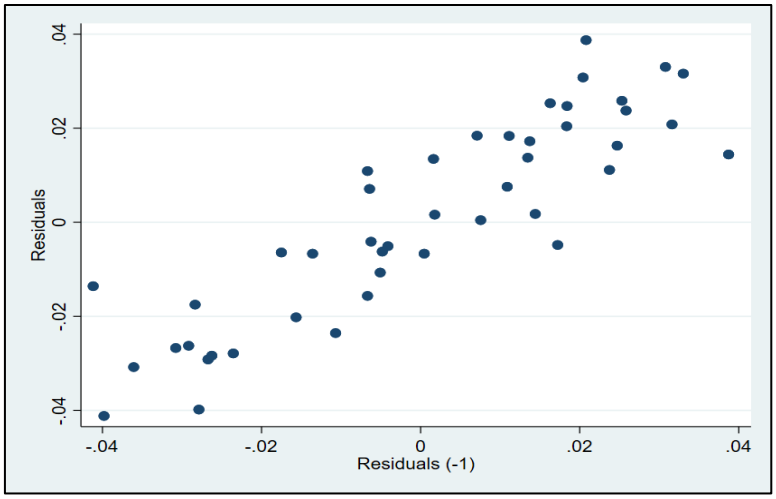
Veamos un ejemplo que (Gujarati & Porter, 2010) Muestran para ilustrar como se emplean estos gráficos.



Gráfica 3.23. Grafica de $\hat{\mu}_i$ y $\hat{\mu}_{est}$ respecto al tiempo.
Elaboración: (Gujarati & Porter, 2010)
Fuente: (Gujarati & Porter, 2010)

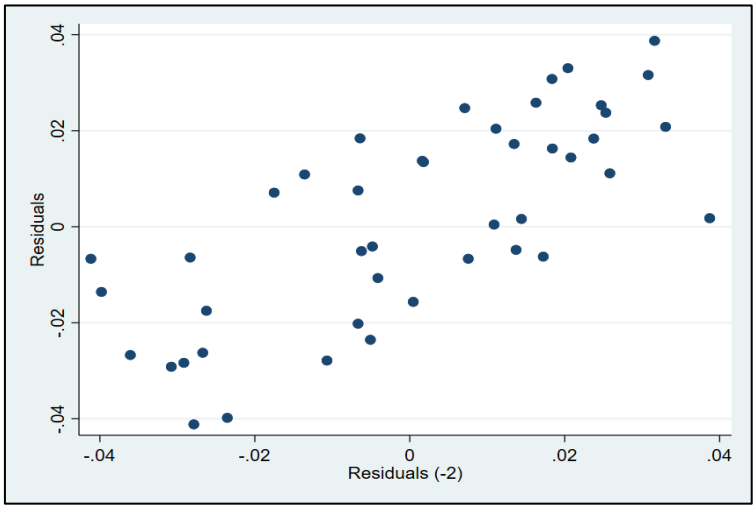
Para (Gujarati & Porter, 2010) Tanto $\hat{\mu}_i$ y $\hat{\mu}_{est}$ siguen un patrón similar por lo que no se puede asegurar que sean aleatorias y es probable que tengan autocorrelación en el modelo. (Pérez L., 2005) Recomienda usar los residuos estandarizados ya que estos pueden ser comparados con los residuos estandarizados de otros modelos econométricos y cumplen la condición de tener media igual a 0.

(Gujarati & Porter, 2010) También recomiendan realizar una gráfica de dispersión de $\widehat{\mu}_i$ versus $\widehat{\mu}_{est}$ el cual corresponde a una prueba empírica para *AR(1)*, a continuación se muestra la gráfica:

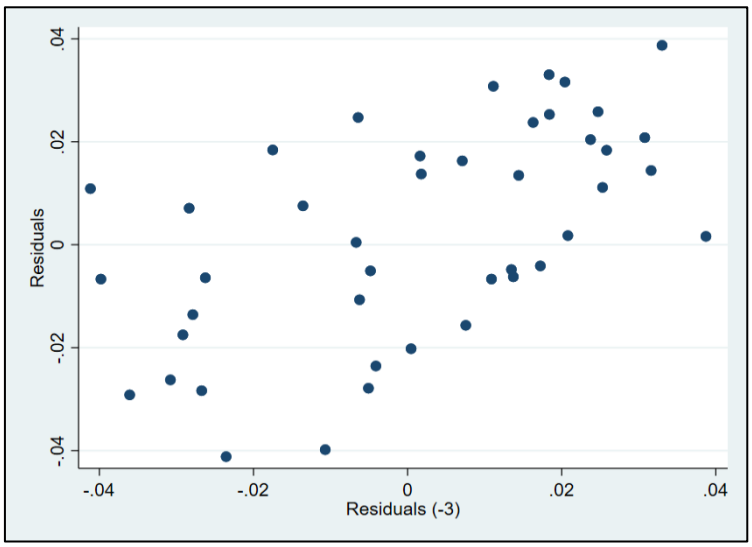


Gráfica 3.24. Grafica de dispersión $\widehat{\mu}_t$ versus $\widehat{\mu}_{t-1}$.
Elaboración propia
Fuente: (Gujarati & Porter, 2010)

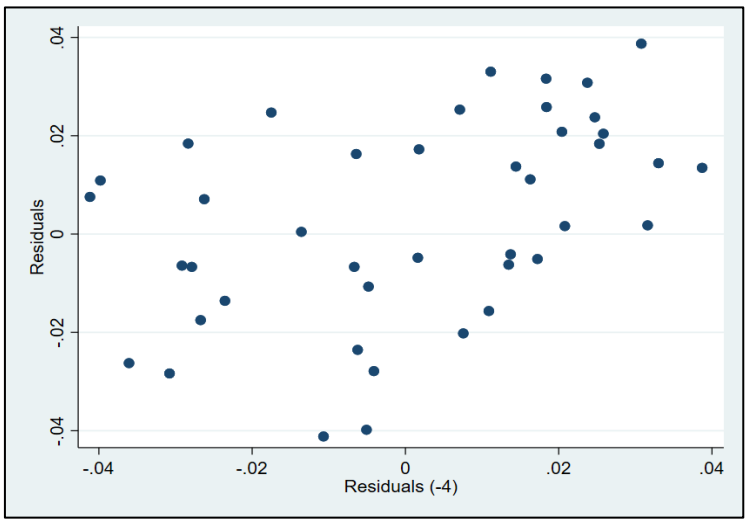
Podemos observar que en la gráfica 3.24. Se muestra un patrón muy evidente, por lo que los residuos del modelo no son aleatorios, de modo que podríamos asumir que existe autocorrelación en el modelo y ya que el patrón es creciente suponemos que se trata de la autocorrelación positiva. Sin embargo, al igual que las gráficas de la heterocedasticidad, estos métodos informales son subjetivos y deberían contrastarse con pruebas de hipótesis las cuales serán empleadas en los métodos formales para comprobar válidamente que existe autocorrelación en el modelo. Para acabar esta sección, veamos cómo se relacionan $\widehat{\mu}_t$ con $\widehat{\mu}_{t-2}$, $\widehat{\mu}_{t-3}$ y $\widehat{\mu}_{t-4}$ en los siguientes gráficos.



Gráfica 3.25. Grafica de dispersión $\widehat{\mu}_t$ versus $\widehat{\mu}_{t-2}$.
Elaboración propia
Fuente: (Gujarati & Porter, 2010)



Gráfica 3.26. Grafica de dispersión $\hat{\mu}_t$ versus $\hat{\mu}_{t-3}$.
Elaboración propia
Fuente: (Gujarati & Porter, 2010)



Gráfica 3.27. Grafica de dispersión $\hat{\mu}_t$ versus $\hat{\mu}_{t-4}$.
Elaboración propia
Fuente: (Gujarati & Porter, 2010)

Siguiendo la teoría propuesta de (Gujarati & Porter, 2010), las gráficas 3.25. 3.26. Y 3.27. Corresponden a los esquemas *AR(2), AR(3), AR(4)* respectivamente. Podemos observar cómo a medida que aumenta el número de rezagos, en las gráficas se ordenan los datos de tal forma que en la última gráfica no se aprecia un patrón ni una tendencia de manera tan evidente, por este motivo podemos argumentar que los residuos del modelo especificado por (Gujarati & Porter, 2010) pueden depender hasta 3 rezagos, entonces la autocorrelación puede aparecer hasta en el 3° rezago. No obstante, la interpretación de estas gráficas es subjetiva y debería ser contrastada con los métodos formales que veremos a continuación.

3.6.3.1.2. Métodos formales.

Los métodos formales para detectar autocorrelación en el modelo siguen un procedimiento parecido a los métodos formales que se usan para detectar la heterocedasticidad. En algunos de estos será necesario realizar una regresión auxiliar y

en otros no, pero en todos estos se usará una prueba de hipótesis para determinar la existencia de autocorrelación en el modelo.

- **Estadístico *d* de Durbin-Watson.**

El test de Durbin-Watson es el método más extendido y de mayor uso para detectar la autocorrelación en un modelo econométrico debido a que es muy práctico y de fácil desarrollo. Fue propuesto por James Durbin y Geoffrey Watson en 1951 y desde su publicación otros economistas han realizado algunas variaciones como el test **de Wallis** que será explicado posteriormente.

El estadístico *d* calculado de Durbin-Watson se halla mediante:

$$d = \frac{\sum(\hat{\mu}_t - \hat{\mu}_{t-1})^2}{\sum \hat{\mu}_t^2} \text{ (3.6.99.)}$$

El cual puede tomar valores desde $0 \leq d \leq 4$, (Pérez L., 2012) Explica lo que significa que el estadístico *d* se acerque a dichos valores en la siguiente cita.

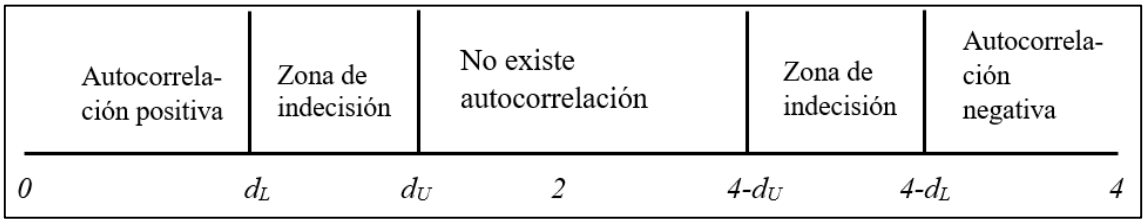
> *"Se puede adoptar la regla no demasiado rigurosa de que si d vale 0 hay autocorrelación perfecta positiva; si d se aproxima a 2 no hay autocorrelación y si d se aproxima a 4 hay autocorrelación perfecta negativa. No obstante, d se encuentra tabulado, y según la franja en la que caiga su valor, se acepta o rechaza la hipótesis de autocorrelación. En la tabla de d elegimos la columna relativa a k (número de regresores en el modelo) y en la fila relativa a n (tamaño muestral), lo que nos da valores d_L y d_U."* (Pérez L., 2012)

Con la cita anterior podemos construir la siguiente prueba de hipótesis:

$$H_0: No\ existe\ autocorrelación$$

$$H_1: Existe\ autocorrelación$$

El estadístico *d* calculado se distribuye de la siguiente forma:



Gráfica 3.28. Grafica de distribución del estadístico *d* calculado.
Elaboración propia
Fuente: (Pérez L., 2012)

El gráfico 3.28. Puede ser resumido en el siguiente esquema:

$$\begin{matrix} d \cong 0 \ \rightarrow p = 1 \\ d \cong 2 \ \rightarrow p = 0 \\ d \cong 4 \ \rightarrow p = -1 \end{matrix} \quad (3.6.100.)$$

Claro que (3.6.100.) puede darnos una sospecha sobre la existencia o no de autocorrelación en el modelo, pero no debe ser tomado como determinante, por el contrario, se debe usar la distribución expuesta en la gráfica 3.28. Donde a d_L y d_U se les denomina como **límite inferior y límite superior** respectivamente, los cuales se encuentran en la **tabla *d*** y se eligen acorde al número de regresores y al tamaño muestral.

Podemos decir que el test de Durbin-Watson es similar a las pruebas anteriores pero la diferencia radica en que no elegimos un estadístico tabulado sino dos estadísticos y además la regla de decisión concibe una tercera posibilidad aparte de aceptar o rechazar la hipótesis nula, la cual está presente en la **zona de indecisión**, como su nombre indica si el estadístico *d* calculado cae en esta zona entonces no podemos determinar la existencia o no de autocorrelación mediante este test. Aparte de la zona de indecisión, también podemos notar dos zonas donde si el estadístico *d* calculado cae en alguna de ellas podemos rechazar la hipótesis nula y concluir que existe autocorrelación en el modelo. Siendo:

- $d < d_L$ Se **rechaza la hipótesis nula** y se concluye que existe autocorrelación **positiva** en el modelo. Por lo que $d \cong 0 \ y \ p = 1$
- $4 - d_L < d \rightarrow -1 < p < 0$ Se **rechaza la hipótesis nula** y se concluye que existe autocorrelación **negativa** en el modelo. Por lo que $d \cong 4 \ y \ p = -1$
- $d_U < d < 4 - d_U$ Se **acepta la hipótesis nula** y se concluye que el modelo está libre de autocorrelación. Por lo que $d \cong 2 y \ p = 0$
- $d_L < d < d_U$ y $4 - d_U < d < 4 - d_L$ No se puede ni rechazar ni aceptar la hipótesis nula.

El test de Durbin-Watson guarda una relación con el coeficiente de autocorrelación donde a (3.6.99.) se puede escribir como.

$$d = \frac{\sum(\hat{\mu}_t - \hat{\mu}_{t-1})^2}{\sum \hat{\mu}_t^2} \cong 2(1 - p) \quad (3.6.101.)$$

(Novales, 1998) Explica cómo se encuentra la relación entre el estadístico *d* calculado y el coeficiente de **autocorrelación.**

$\frac{\sum \hat{\mu}_t^2 - 2\sum \hat{\mu}_t \hat{\mu}_{t-1} + \sum \hat{\mu}_{t-1}^2}{\sum \hat{\mu}_t^2} \cong 2\frac{\sum \hat{\mu}_t^2 - \sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2}$ (3.6.102.)

(Gujarati & Porter, 2010) Explican que $\sum \hat{\mu}_t^2$ y $\sum \hat{\mu}_{t-1}^2$ son aproximadamente iguales por lo que al reescribir (3.6.102.) tenemos:

$d \cong 2\left(1 - \frac{\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2}\right)$ (3.6.103.)

Por lo que el coeficiente de autocorrelación es igual a:

$p = \frac{\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_t^2}$ (3.6.104.)

Posteriormente se necesitará calcular el coeficiente de autocorrelación para poder ejecutar un método correctivo a la autocorrelación.

(Gujarati & Porter, 2010) Establecen algunos supuestos que debe cumplir este test para que tenga validez al momento de utilizarlo:

- El modelo econométrico debe tener el intercepto para que no afecte el cálculo de la Suma Cuadrática Residual, $\sum \mu_t^2$.
- La(s) variables(s) explicativa(s) no deben ser variables estocásticas.
- Solo se puede aplicar a modelos econométricos que utilizan el *AR(1)* para explicar la autocorrelación presente del término de error. Por lo que, no debe ser usado para determinar si existe una dependencia con periodos rezagados superiores. Una forma de saber cuál es el orden del rezago que sigue el término de error es consultando la **función de autocorrelación simple (FAC)** y la **función de autocorrelación parcial (FACP)**.
- El término de error debe tener distribución normal $\mu \sim N(0, \sigma^2)$, es decir debe cumplir el supuesto de normalidad de los errores el cual indica que tiene media igual a 0 y varianza constante. Esto puede ser comprobado fácilmente con un histograma y con el test de Shapiro-Wilk o el test de Shapiro-Francia.
- El modelo econométrico no debe incluir variables rezagadas, esta condición debe cumplirse para la variable dependiente como para la(s) variables(s) explicativa(s).
- La muestra empleada no debe tener observaciones faltantes.

Finalmente, aunque el test de Durbin-Watson se le considera mucho para un contraste de hipótesis para determinar o no la existencia de autocorrelación, algunos

autores plantean que este estadístico puede ser usado también para verificar si el modelo econométrico tiene un sesgo de especificación, ya sea por omisión de una regresora importante o por una incorrecta forma funcional ya que estos sesgos de especificación también hacen que el estadístico *d* sea significativo por lo cual podemos rechazar la hipótesis nula. (De Grange C., 2005) También señala que si la estructura de autocorrelación en los residuos es estacional entonces el test de Durbin-Watson pierde validez. Sin embargo, los residuos rara vez tienen un componente estacional bien definido, en la mayoría de casos, cuando consultamos con los gráficos con respecto al tiempo observamos patrones o tendencias y muy pocas veces un comportamiento repetitivo en periodos menores o iguales a un año. Un ejemplo de una variable con un componente estacional definido sería los ingresos de las empresas dedicadas al sector transporte, ya que, en los meses de marzo, julio y sobre todo diciembre sus ingresos tienden a ser demasiado elevados con respecto a los demás meses, ya que en esos meses las personas viajan con más frecuencia aprovechando los feriados y las fiestas de navidad y año nuevo.

- **Test alternativo de Durbin: la prueba h.**

Una desventaja del test de Durbin-Watson, es la imposibilidad de aplicarse a modelos econométricos autorregresivos, es decir, los modelos econométricos que tienen variables rezagadas de la variable dependiente como una variable explicativa no pueden ser contrastados mediante la prueba de Durbin-Watson. Un modelo econométrico autorregresivo se especifica de la siguiente forma.

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \cdots + \hat{\beta}_k X_{kt} + \hat{\gamma} Y_{t-1} + \hat{\mu}_t \quad (3.6.105.)$$

(Gujarati & Porter, 2010) Señalan que no se podría utilizar la prueba Durbin-Watson para determinar si existe autocorrelación en (3.6.105.) por lo que se debe utilizar **la prueba h**, la cual es un test alternativo planteado por Durbin en 1970. Se plantea la misma prueba de hipótesis.

$$H_0: No\ existe\ autocorrelación$$

$$H_1: Existe\ autocorrelación$$

Donde el estadístico *h* calculado se calcula con:

$$h = p\sqrt{\frac{n}{1-n*var(\hat{\gamma})}} = \frac{\sum \hat{\mu}_t \hat{\mu}_{t-1}}{\sum \hat{\mu}_{t-1}^2}\sqrt{\frac{n}{1-n*var(\hat{\gamma})}} \text{ (3.6.106.)}$$

El estadístico *h* sigue la siguiente distribución $h \sim N(0,1)$, por lo que si se utiliza una significancia de 5% como es lo habitual en la econometría, tenemos la siguiente regla de decisión explicada por (Pérez L., 2012):

- Si $|h| < 1.96$ entonces no se rechaza la hipótesis nula y no existe autocorrelación en el modelo autorregresivo.
- Si $|h| > 1.96$ entonces se rechaza la hipótesis nula y existe autocorrelación en el modelo autorregresivo.

Finalmente, cabe mencionar la diferencia entre los **modelos autorregresivos** y los **modelos de rezagos distribuidos**. Básicamente, un modelo autorregresivo es un modelo en el que la variable dependiente además de depender de las regresoras también depende de un número determinado de rezagos de la variable dependiente como se vio en (3.6.105.), (Gujarati & Porter, 2010) Llaman a los modelos autorregresivos como **modelos dinámicos**. Por otro lado, se denominan modelos de rezagos distribuidos a los modelos en el que su variable dependiente depende de las variables regresoras y además de los rezagos de las regresoras. (Gujarati & Porter, 2010) Especifican un modelo de rezagos distribuidos.

$$Y_t = \hat{\alpha}_1 + \hat{\beta}_1 X_t + \hat{\beta}_2 X_{t-1} + \hat{\beta}_3 X_{t-2} + \cdots + \hat{\beta}_k X_{t-p} + \hat{\mu}_t \text{ (3.6.107.)}$$

- **Test de Wallis.**

La prueba de Wallis es una variación del test de Durbin-Watson cuando se usa datos trimestrales. (Pérez L., 2012) Muestra el estadístico d_4 mediante la siguiente fórmula.

$$d_4 = \frac{\sum(\hat{\mu}_t - \hat{\mu}_{t-4})^2}{\sum \hat{\mu}_t^2} \text{ (3.6.108.)}$$

Donde al igual que el test de Durbin-Watson, debe cumplir los mismos supuestos anteriormente explicados, y los estadísticos d_{4L} y d_{4u} tabulados se obtienen de la **tabla $\boldsymbol{d_4}$**, tomando en cuenta el número de regresores y el tamaño de la muestra. Por último, también sigue las mismas reglas de decisión:

- $d_4 < d_{4L}$ Se **rechaza la hipótesis nula** y se concluye que existe autocorrelación **positiva** en el modelo. Por lo que $d_4 \cong 0\ y\ p = 1$
- $4 - d_{4L} < d_4 \rightarrow -1 < p < 0$ Se **rechaza la hipótesis nula** y se concluye existe autocorrelación **negativa** en el modelo. Por lo que $d_4 \cong 4\ y\ p = -1$
- $d_{4U} < d_4 < 4 - d_{4U}$ Se **acepta la hipótesis nula** y se concluye que el modelo está libre de autocorrelación. Por lo que $d_4 \cong 2y\ p = 0$
- $d_{4L} < d_4 < d_{4U}$ o $4 - d_{4U} < d_4 < 4 - d_{4L}$ No se puede ni rechazar ni aceptar la hipótesis nula.

- **Test de prueba general de Breusch-Godfrey.**

Debido a que el test de Durbin-Watson en ciertos modelos no puede ser válido su uso, en 1978 Trevor S. Breusch y Leslie G. Godfrey propusieron el test Breusch-Godfrey que en cierta medida puede resultar ser un contraste más eficiente que el Durbin y Watson.

(Gujarati & Porter, 2010) **Definen a este test como un test general ya que esta prueba de autocorrelación no solo permite contrastar a modelos con procesos autorregresivos de cualquier orden, sino también admiten el contraste sobre la existencia de autocorrelación en modelos con rezagos de las regresoras como variables explicativas y en los modelos con promedios móviles.**

Aunque los procesos de promedios móviles rara vez se utilizan, el contraste BG presenta resultados válidos para estos modelos econométricos. Brevemente, se explicará el concepto de proceso de promedios móviles. Según (Hanke & Wichern, 2006), un proceso o esquema de promedio móvil es un tipo de esquema utilizado en la teoría de econometría de series temporales y es muy parecido a los modelos autorregresivos *AR(p)*. Previamente a exponer su definición, (Hanke & Wichern, 2006) Muestran cómo se especifica un modelo de promedio móvil. Primero debemos especificar de forma general un proceso autorregresivo de *p* orden, es decir un *AR(p)*.

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \text{ (3.6.109.)}$$

Entonces especificamos el proceso de promedios móviles.

$$Y_t = \mu + e_t + \omega_1 e_{t-1} + \omega_2 e_{t-2} + \cdots + \omega_q e_{t-q} \text{ (3.6.110.)}$$

Con (3.6.110.) ya podemos vislumbrar un concepto sobre los esquemas de promedio móvil. Un proceso de promedio móvil es un tipo de proceso donde la variable dependiente depende del término de error y de periodos rezagados del término de error y

μ representa el término constante en el modelo (3.6.110.) Tanto en (3.6.109.) como en (3.6.110.) la variable e_t es una variable ruido blanco, es decir tiene media y varianza constante y además sus valores no están correlacionados. Por último los procesos de promedios móviles están representados como *MA(q).*

Hagamos un breve paréntesis en la explicación. Se puede reconocer el esquema *AR(1)* que hemos estado utilizando para explicar el comportamiento del término de error en presencia de autocorrelación en (3.6.109.), siendo $\hat{\mu}_t = p\hat{\mu}_{t-1} + e_t$ muy semejante a $Y_t = \phi_0 + \phi_1 Y_{t-1} + e_t$, la diferencia entre ambas expresiones es la falta del término constante en el esquema *AR(1)* usado para explicar la autocorrelación en (3.6.79.).

Al igual que con los esquemas *AR(1)*, se suele utilizar a los *MA(1)* cómo introducción al tema, siendo el siguiente modelo la especificación de un *MA(1)*.

$Y_t = \mu + e_t + \omega_1 e_{t-1}$ (3.6.111.)

(De Grange C., 2005) Brevemente explica que el proceso *MA(1)* es un modelo de memoria muy corta, es decir que toma en cuenta más los valores pasados cercanos al presente que los valores pasados más alejados.

Los procesos *AR(p)* y *MA(q)* se les conocen como modelos univariados y su uso no solo está limitado para explicar el comportamiento que presenta una variable usando el comportamiento del pasado, sino también para realizar pronósticos. Estos temas son muy importantes en la teoría de econometría de series temporales ya que su correcta estimación e interpretación nos puede brindar pronósticos cada vez más precisos los cuales son importantes para la toma de decisiones en las empresas o en las políticas de un organismo estatal, además que son los temas introductorios a la teoría de la econometría financiera.

Retomando el contraste BG, esta prueba se basa en el principio multiplicador de Lagrange, y tiene la siguiente prueba de hipótesis.

$$H_0: p_1 = p_2 = \cdots = p_p = 0$$

$$H_1: Algún\ p_p\ es\ diferente\ a\ 0$$

Al tener el siguiente modelo econométrico $Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \cdots + \hat{\beta}_k X_{kt} + \hat{\mu}_t$, se sigue el siguiente procedimiento.

Paso 1. Realizar la regresión mediante MCO.

Paso 2. Obtener los residuos del modelo.

Paso 3. Realizar la siguiente regresión auxiliar.

$$\hat{\mu}_t = \hat{\alpha}_1 + \hat{\alpha}_2 X_{2t} + \hat{\alpha}_3 X_{3t} + \cdots + \hat{\alpha}_k X_{kt} + p_1 \hat{\mu}_{t-1} + p_2 \hat{\mu}_{t-2} + \cdots + p_p \hat{\mu}_{t-p} + e_t$$ (3.6.112.)

Paso 4. Calcular el coeficiente de determinación de (3.6.112.).

Paso 5. Calcular el siguiente estadístico LM calculado.

$$LM = (n - p) R_i^2$$ (3.6.113.)

Donde *n* es el tamaño de la muestra, *p* es el número de rezagos en (3.6.112.) y R_i^2 es el coeficiente de determinación del modelo auxiliar.

Paso 6. El estadístico LM calculado sigue la siguiente distribución.

$$LM \sim X_p^2$$ (3.6.114.)

Donde el estadístico ji-cuadrado tabulado tiene *p* grados de libertad, siendo *p* es el número de rezagos del término de error introducidos en (3.6.112.)

Paso 7. Aplicar la regla de decisión:

- **Si estadístico *LM* calculado es mayor al estadístico X_p^2 entonces se rechaza la hipótesis nula y se concluye que existe autocorrelación en el modelo.**
- **Si estadístico *LM* calculado es menor al estadístico X_p^2 entonces se acepta la hipótesis nula y se concluye que el modelo está libre de autocorrelación.**

(Gujarati & Porter, 2010) Mencionan que el test BG puede mostrar resultados válidos al incluir rezagos de la variable dependiente como variables explicativas. Incluso está permitido aplicar este contraste a los esquemas de **promedios móviles**. (Gujarati & Porter, 2010) Representan a los esquemas ***MA(q)*** como:

$$\hat{\mu}_t = e_t + \omega_1 e_{t-1} + \omega_2 e_{t-2} + \cdots + \omega_q e_{t-q}$$ (3.6.115.)

Donde q es el número de rezagos del proceso de ruido blanco e_t extraído de (3.6.112.) que influyen en el término de error μ_t y además la variable e_t sigue siendo un proceso de ruido blanco.

La aplicación del contraste BG en (3.6.115.) sigue el mismo procedimiento, sin embargo rara vez se usa este tipo de esquema, ya que los errores que presentan autocorrelación casi siempre siguen un esquema *AR(1)*. De hecho cuando se usa un *AR(1)* en (3.6.112.) al test de BG se le conoce como **prueba m de Durbin** según (Gujarati & Porter, 2010).

Podemos intuir entonces que el test BG es conveniente cuando el test Durbin-Watson no puede determinar la presencia de autocorrelación, y tal como se puede observar en (3.6.112.) la principal ventaja del test BG es que permite utilizar esquemas autorregresivos de orden superior a 1, sin embargo esta también puede ser su principal desventaja, ya que esto implica conocer cuál es el orden *p* del esquema autorregresivo del que depende el término de error. Surge entonces la siguiente pregunta ¿Cómo podemos conocer cuál es el orden del esquema autorregresivo que sigue el término de error? Para ello (Pérez L., 2012) Recomienda el uso del **correlograma,** que en términos muy sencillos es un tipo de grafico que muestra la **función de autocorrelación simple (FAC) y la función de autocorrelación parcial (FACP)**, y a su vez la FAC y la FACP se utilizan para conocer el orden de *MA(q) y AR(p),* respectivamente. Sin embargo, estos conceptos son muy profundos y están relacionados a la especificación de modelos ARMA y ARIMA, por tanto, no serán tratados en esta guía de estudios, en vista que son propios de la teoría de la econometría de series temporales y su estudio y explicación requiere plantear otros saberes previos que no son objeto de análisis en este trabajo. Según (Brooks, 2008) La pregunta anteriormente formulada no tiene una respuesta clara y recomienda ***experimentar*** **con un determinado número de rezagos** y además de tomar en cuenta la **frecuencia de los datos**, supongamos que los datos son mensuales o trimestrales, entonces el número de rezagos con el que se puede *experimentar* seria 12 o 4 respectivamente. El punto es que se espera que los errores presenten correlación con los errores del año pasado y se escogería el número de retardos donde ya no haya autocorrelación, es decir cuando *p* ya no sea significativo; no obstante el problema de esto es que a medida que menor es la frecuencia mayor es el número de rezagos a probar, y realizar esto puede ser contraproducente. Por ejemplo si tuviéramos una frecuencia diaria entonces tendríamos que probar 30 rezagos para contrastar con el error del mes pasado o

365 rezagos para contrastar con el error del año pasado; por lo que esto sería recomendable en frecuencias altas y aun así según (Gujarati & Porter, 2010) Establecen que no se puede determinar de manera *a priori* el número de rezagos por lo que probar rezagos con órdenes exageradamente elevados podría ser ineficiente. También establecen utilizar los **criterios de información Akaike y Schwarz**, los criterios de información son empleados para la elección de modelos econométricos y parece ser una buena opción, posteriormente se ilustraran como calcularlos en el ejemplo que se realizara con STATA. En la elaboración de modelos econométricos se sigue un principio llamado el **principio de la parsimonia**, el cual establece que la respuesta correcta ante una situación complicada, suele ser la más sencilla. De esta manera podemos argumentar que seguir un esquema *AR(1)* no es incorrecta, de hecho la mayoría de modelos econométricos siguen este esquema y en la econometría básica suele ser muy recomendado para posteriormente investigar si se puede optar un esquema autorregresivo de orden superior. No obstante, STATA tiene una opción en un comando que permite determinar el número de rezagos, posteriormente será explicado.

- **Test de Box-Pierce.**

(Greene, 2012) Define al test de Box-Pierce como una prueba asintóticamente equivalente al test de BG, la cual tiene la siguiente prueba de hipótesis:

$$H_0: p = 0$$

$$H_1: p \neq 0$$

Donde la hipótesis nula indica que el modelo está libre de autocorrelación mientras que la hipótesis alternativa indica que el modelo presenta autocorrelación.

Cabe señalar que a diferencia del test BG, para poder ejecutar el test de Box-Pierce el modelo original no debe incluir rezagos de la(s) variable(s) explicativa(s). Para la ejecución del test de Box-Pierce, se siguen el siguiente procedimiento teniendo el modelo econométrico: $Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_{2t} + \hat{\beta}_3 X_{3t} + \cdots + \hat{\beta}_k X_{kt} + \hat{\mu}_t$.

Paso 1. Realizar la regresión del modelo mediante MCO.

Paso 2. Obtener los residuos del modelo econométrico.

Paso 3. Calcular el siguiente estadístico calculado:

$Q = n\sum_{j=1}^{p} r_j^2$ (3.6.116.)

Donde $r_j = \frac{\sum_{t=j+1}^{n}(\hat{\mu}_t\hat{\mu}_{t-j})}{\sum_{t=1}^{n}\hat{\mu}_t^2}$, el estadístico *Q* es igual al producto del tamaño muestral por la sumatoria de r_j^2 tomando en cuenta el número de *p* retardos que sigue el esquema autorregresivo en el modelo econométrico. Y r_j es igual a la división de la sumatoria del producto de $\hat{\mu}_t$ por $\hat{\mu}_{t-j}$ desde *j+1* hasta *n*, entre la sumatoria de $\hat{\mu}_t^2$.

Paso 4. El estadístico *Q* calculado sigue la siguiente distribución:

$Q \sim X_p^2$ (3.6.117.)

Donde *Q* se distribuye en ji-cuadrado con *p* grados de libertad, donde *p* es el número de retardos introducidos.

Paso 5. Aplicar la siguiente regla de decisión:

- **Si *Q* es mayor que X_p^2 entonces se rechaza la hipótesis nula y se concluye que existe autocorrelación en el modelo.**
- **Si *Q* es menor que X_p^2 entonces se acepta la hipótesis nula y se concluye que el modelo está libre de autocorrelación.**

(De Grange C., 2005) Explica que la principal diferencia entre la prueba de Box-Pierce con la prueba de BG, es que la primera hace uso de las correlaciones simples mientras que el segundo hace uso de las correlaciones parciales. (Greene, 2012) Complementa lo anterior afirmando que el uso de las correlaciones parciales en el test de BG sirve para el control de las variables explicativas. Además, bajo la hipótesis nula que el término de error no tiene autocorrelación y que las variables explicativas no están correlacionadas con el término de error, entonces ambas pruebas son equivalentes asintóticamente. Finalmente, también menciona que el estadístico *Q* calculado ha recibido una mejora, donde la fórmula para calcularlo es:

$Q' = n(n+2)\sum_{j=1}^{p}\frac{r_j^2}{n-j}$ (3.6.118.)

La fórmula (3.6.118.) fue propuesta por Ljung y Box en 1979, sin embargo (3.6.118.) se usa más para comprobar que el modelo esté libre de autocorrelaciones para que cumpla la condición de ruido blanco que para comprobar la existencia de

autocorrelación en un modelo, frecuentemente se usa como postestimación de modelos ARIMA o ARMA, por lo que no realizaremos este test en el ejemplo de autocorrelación con STATA que posteriormente se presentará.

3.6.3.2. Tratamiento para autocorrelación.

3.6.3.2.1. *Forma funcional correcta.*

Previamente se había expuesto que, si al usarse la prueba de Durbin-Watson se rechazaba la hipótesis nula, cabía la posibilidad que la autocorrelación era causada por un error en la forma funcional, por lo tanto, deberíamos contrastar si la forma funcional que se ha elegido es la correcta antes de contrastar si el modelo presenta autocorrelación. Para entenderlo se presenta un ejemplo recogido de (Pérez L., 2012). Se especifica el siguiente modelo y su tabla de datos.

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_t + \hat{\mu}_t \text{ (3.6.119.)}$$

Tabla 3.18. Base de datos para el modelo (3.6.119.)
Elaboración: (Pérez L., 2012)
Fuente: (Pérez L., 2012)

Al efectuarse la regresión mediante MCO en el modelo (3.6.119.) obtenemos:

$$\hat{Y}_t = 8.01 + 4.46 X_i + \hat{\mu}_t \text{ (3.6.120.)}$$

N	Y_t	X_t
1	6	-4
2	3	-3
3	1	-2
4	1	-1
5	1	1
6	4	2
7	6	3
8	16	4
9	25	5
10	36	6
11	49	7
12	64	8

$$ee = (4.06) \quad (0.92)$$

$$t = (1.97) \quad (4.85)$$

Para empezamos detectar la autocorrelación en el modelo, determinando la siguiente prueba de hipótesis.

$$H_0: No\ existe\ autocorrelación$$

$$H_1: Existe\ autocorrelación$$

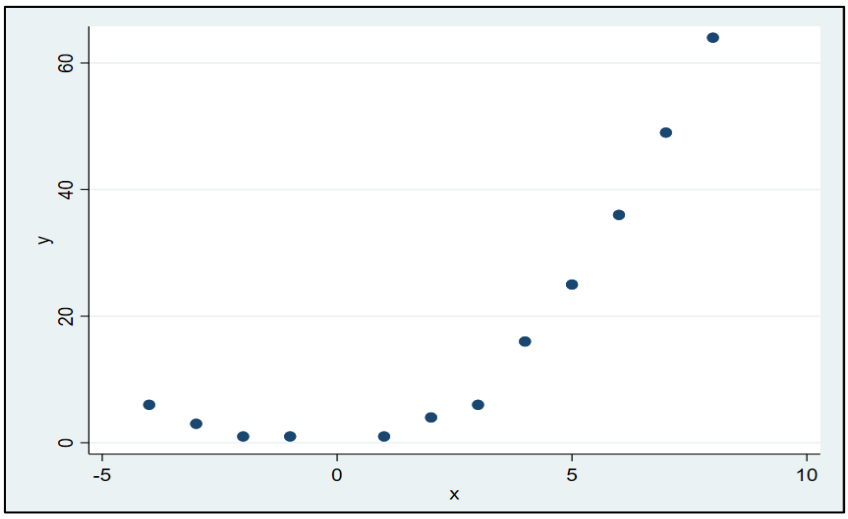
Donde el estadístico d calculado de la prueba Durbin-Watson:

$d = 0.32$ (3.6.121.)

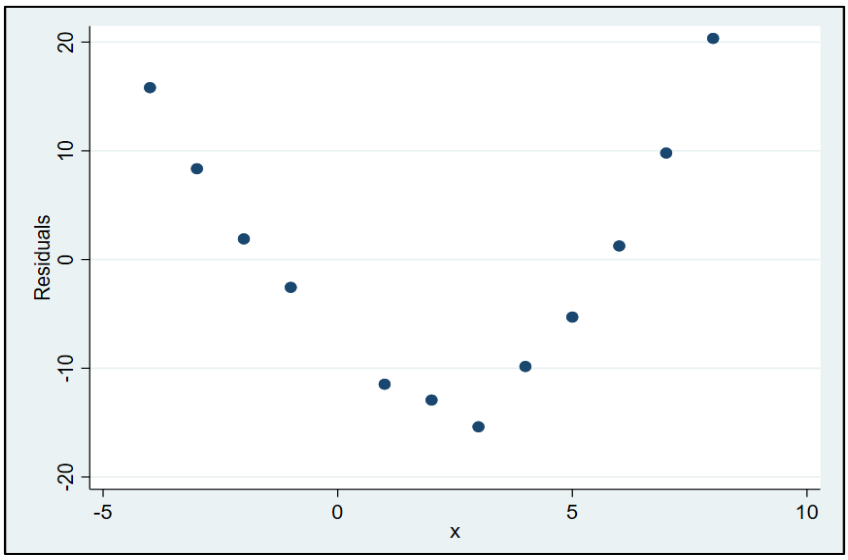
Y según la tabla d, los estadísticos tabulados usando el 5% de nivel de significancia son $d_L = 0.971$ y $d_U = 1.331$, al notar que $d < d_L$ entonces se rechaza la hipótesis nula y se acepta la hipótesis alternativa por lo que se asume que el modelo tiene autocorrelación. De hecho si ejecutamos el esquema *AR(1)* de (3.6.120.) obtenemos:

$$\hat{\mu}_t = 0.85\hat{\mu}_{t-1} + \hat{e}_t \text{ (3.6.122.)}$$

Podemos observar que $p = 0.85$ y se acerca a 1, por lo que podemos confirmar que existe autocorrelación positiva en el modelo. (Pérez L., 2012) Explica que si bien se ha demostrado la existencia de autocorrelación aún no se ha demostrado que la forma funcional es la correcta en (3.6.119.). Para ello se presentarán los siguientes gráficos.



Gráfica 3.29. Grafica de dispersión entre Y_t y X_i.
Elaboración propia
Fuente: (Pérez L., 2012)



Gráfica 3.30. Grafica de dispersión entre X_i y $\hat{\mu}_t$.
Elaboración propia
Fuente: (Pérez L., 2012)

En los dos gráficos anteriores podemos ver como los puntos no sugieren que las variables guardan una relación lineal sino una relación cuadrática, en consecuencia, concluimos que el test de Durbin-Watson está admitiendo que existe autocorrelación generada por una forma funcional incorrecta. Para corregir el problema debemos incluir la variable X^2 en (3.6.119.)

$$Y_t = \hat{\beta}_1 + \hat{\beta}_2 X_t + \hat{\beta}_3 X_t^2 + \hat{u}_t \text{ (3.6.123.)}$$

Y obtenemos los siguientes resultados al efectuar la regresión mediante MCO.

$$\hat{Y}_t = -1.78 + 1.03X_t + 0.88X_t^2 + \hat{u}_t$$

$$ee = \quad (0.74) \quad (0.04) \quad (0.21)$$

$$t = \quad -2.41 \quad 21.66 \quad 5.01$$

Al efectuar la regresión $\hat{u}_t = p\hat{u}_{t-1} + \hat{v}_t$ mediante MCO, tomando en cuenta que $\hat{u}_t$ es el término de error en la regresión (3.6.123.) obtenemos $\hat{u}_t = 0.30\hat{u}_{t-1} + \hat{v}_t$. Podemos notar que $p = 0.30$ en (3.6.123.) por lo que tenemos sospecha que el modelo está libre de autocorrelación, y para estar completamente seguro realizamos el test de Durbin-Watson.

El estadístico d calculado es $d = 1.21$. Con un nivel de significancia de 5% y tomando en cuenta que hay 2 regresores y 12 observaciones en la tabla d encontramos los siguientes estadísticos tabulados $d_L = 0.81$ y $d_U = 1.58$. Debido a que $d_L < d < d_U$ entonces el estadístico d caería en la zona de indecisión por lo cual no podemos

determinar la existencia de autocorrelación mediante el test de Durbin-Watson. Realicemos el test de BG para la comprobación de existencia de autocorrelación.

El estadístico *LM* calculado es 1.077 y el estadístico tabulado $X_1^2 = 3.84146$. Según el contraste de BG, el resultado obtenido fue $LM < X_1^2$, por esta razón no rechazamos la hipótesis nula y concluimos que el modelo (3.6.123.) está libre de autocorrelación, entonces habremos logrado corregir el modelo original. Cabe mencionar que en este ejemplo se ha utilizado un *AR(1)* para explicar la autocorrelación del término de error tanto en el modelo original como en el modelo corregido.

Finalmente, cuando la autocorrelación es originada por una mala especificación ya sea por un subajuste o por una forma funcional incorrecta como se ha visto en el ejemplo, entonces el modelo presenta **autocorrelación impura**. (Gujarati & Porter, 2010) Establecen que si al incluir variables relevantes o utilizar otra forma funcional, todavía existe autocorrelación entonces no estamos ante un modelo con autocorrelación impura sino **autocorrelación pura**, la cual no es causada por un sesgo de especificación sino por la naturaleza de las variables con datos de series temporales o de corte transversal.

3.6.3.2.2. Mínimos Cuadrados Generalizados Factibles.

Después de haber verificado que el modelo presenta autocorrelación pura entonces podemos optar por realizar un método correctivo que implica transformar el modelo original. Empecemos con el método correctivo por MCGF.

En primer lugar, cabe mencionar que para aplicar el método correctivo por MCGF debemos conocer cómo se correlacionan los errores entre sí, por ello asumimos que la autocorrelación sigue un *AR(1)*. (Novales, 1998) Explica que al tener el modelo $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$ cuyo término de error depende sus propios valores rezagados un periodo, es decir $\mu_t = p\mu_{t-1} + e_t$ entonces el método correctivo por MCGF empieza especificando el modelo original en forma de sus rezagos.

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + \mu_{t-1} \quad (3.6.124.)$$

(Brooks, 2008) Explica que esto es válido ya que se asume que el modelo original es correcto en el momento *t* entonces su primer rezago *t-1* también será válido. Después multiplicamos *p* a cada elemento de la ecuación (3.6.124.) entonces obtenemos.

$$pY_{t-1} = p\beta_1 + p\beta_2 X_{2t-1} + p\beta_3 X_{3t-1} + p\mu_{t-1} \quad (3.6.125.)$$

Y finalmente, restamos (3.6.125.) en el modelo original.

$$Y_t - pY_{t-1} = \beta_1 - p\beta_1 + \beta_2 X_{2t} - p\beta_2 X_{2t-1} + \beta_3 X_{3t} - p\beta_3 X_{3t-1} + \mu_t - p\mu_{t-1} \quad (3.6.126.)$$

(Gujarati & Porter, 2010) Denominan a (3.6.126.) como **regresión generalizada, cuasi generalizada o ecuación en diferencias**. De esta forma se habrá transformado el modelo original y a la transformación (3.6.126.) se le aplica un MCO. En realidad, el método correctivo **Mínimos Cuadrados Generalizados Factibles (MCGF)**, o simplemente **Mínimos Cuadrados Factibles (MCF)**, es a una extensión de los Mínimos Cuadrados Generalizados (MCG), cuya explicación de los MCG se ha detallado cuando se explicó sobre métodos correctivos para la heterocedasticidad. El procedimiento MCF como método correctivo de heterocedasticidad es muy parecido al MCG y MCP aplicados al tratamiento de la heterocedasticidad y se diferencia entre los MCP y los errores de White en no que intentamos acercarnos a la estructura de la varianza mediante las regresoras ni tampoco usamos los residuos sino realizamos una *estimación* con el logaritmo de los residuos, posteriormente se mostrará en STATA como realizar los MCF en modelos con presencia de heterocedasticidad con el fin de corregir la heterocedasticidad.

Los MCF aplicados a la corrección de modelos con presencia de autocorrelación también son una extensión de los MCG y su utilidad como medida correctiva de autocorrelación implica conocer la estructura de la matriz Ω de (3.6.96.). Para fines didácticos y ya que el proceso *AR(1)* es el más usado, asumimos que los errores siguen un esquema *AR(1).*

(Greene, 2012) Explica que este método se basa en **estimar los estimadores factibles**, los cuales son (3.6.126.), además establece que si asumimos que en el modelo original las variables explicativas y el término de error son procesos estacionarios y ergódicos entonces los estimadores de MCF también son estacionarios y ergódicos. (De Grange C., 2005) Define el concepto de ergódico como un proceso donde los promedios estadísticos se calculan a partir de una realización, lo que significa que los promedios estadísticos son los mismos que los promedios temporales, esto es algo deseable porque si aumentamos el número de retardos entonces *p* comienza a decrecer y recuerde que lo

deseable es que $p = 0$. Pero ¿Cómo podemos estar seguros que (3.6.126.) no tiene autocorrelación? Si factorizamos el modelo transformado tenemos:

$$Y_t - pY_{t-1} = (1-p)\beta_1 + \beta_2(X_{2t} - pX_{2t-1}) + \beta_3(X_{3t} - pX_{3t-1}) + \mu_t - p\mu_{t-1}$$
(3.6.127.)

En (3.6.127.) observamos que el término de error es $\mu_t - p\mu_{t-1} = e_t$ y al reemplazarlo (3.6.127.) es:

$$Y_t - pY_{t-1} = (1-p)\beta_1 + \beta_2(X_{2t} - pX_{2t-1}) + \beta_3(X_{3t} - pX_{3t-1}) + e_t$$
(3.6.128.)

Y si recordamos que e_t es un proceso de ruido blanco con media y varianza constante y valores independientes entre sí, entonces el modelo transformado (3.6.128.) está libre de autocorrelación. Se puede reescribir (3.6.128.) de la siguiente forma:

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + e_t \text{ (3.6.129.)}$$

Esta forma de estimar los estimadores factibles fue propuesta por los economistas Donald Cochrane y Guy Henderson Orcutt, por tanto, en honor a quienes lo plantearon **a este método se le conoce como estimación de MCGF mediante Cochrane-Orcutt o simplemente método Cochrane-Orcutt**.

Para entender de dónde proviene esta forma de estimación, veamos la breve explicación usando matrices que expone (Greene, 2012).

Para entender cómo funciona este método, conviene repasar ¿Por qué resulta la autocorrelación pura un problema para obtener estimadores MELI? Resumiendo lo anteriormente explicado, en presencia de autocorrelación la varianza del término de error ya no es insesgado y tampoco eficiente. La matriz de la varianza del error en condiciones que cumple los supuestos de MCO es: $E(\mu\mu') = \sigma^2 I$, sin embargo en presencia de autocorrelación la matriz se convierte en:

$$E(\mu\mu') = \sigma^2\Omega = \frac{\sigma_e^2}{1-p^2}\begin{bmatrix} 1 & p & p^2 & p^3 & \cdots & p^{n-1} \\ p & 1 & p & p^2 & \cdots & p^{n-2} \\ p^2 & p & 1 & p & \cdots & p^{n-3} \\ p^3 & p^2 & p & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & p \\ p^{n-1} & p^{n-2} & p^{n-3} & \cdots & p & 1 \end{bmatrix} \text{ (3.6.96.)}$$

(Greene, 2012) Explica que para obtener los estimadores de MCF mediante el método Cochrane-Orcutt primero tomemos la inversa de la matriz Ω.

$$\Omega^{-1} = \frac{1}{1-p^2}\begin{bmatrix} 1 & -p & 0 & \cdots & 0 & 0 \\ -p & 1+p^2 & -p & \cdots & 0 & 0 \\ 0 & -p & 1+p^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+p^2 & -p \\ 0 & 0 & 0 & \cdots & -p & 1 \end{bmatrix} \text{(3.6.130.)}$$

Podemos ver que los elementos de la diagonal de la matriz (3.6.130.) son iguales a excepción del primer y último elemento que son 1, los elementos por encima y por debajo de la diagonal son los mismos, específicamente $-p$ y 0. En los MCG, la inversa de la matriz Ω es igual a $\Omega^{-1} = P'P$, la matriz P será utilizada para transformar el modelo original y posteriormente estimar el modelo transformado por MCO. El método Cochrane-Orcutt utiliza la siguiente matriz P.

$$P = \begin{bmatrix} \sqrt{1-p^2} & 0 & 0 & \cdots & 0 & 0 \\ -p & 1 & 0 & \cdots & 0 & 0 \\ 0 & -p & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -p & 1 \end{bmatrix} \text{(3.6.131.)}$$

Para analizar mejor la matriz (3.6.131.) ignoremos por un momento la primera fila y generemos la submatriz P^* con los elementos restantes:

$$P^* = \begin{bmatrix} -p & 1 & 0 & \cdots & 0 & 0 \\ 0 & -p & 1 & \cdots & 0 & 0 \\ 0 & 0 & -p & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -p & 1 \end{bmatrix} \text{(3.6.132.)}$$

La matriz generada en (3.6.132.) corresponde propiamente a la transformación propuesta por C-O. En esta submatriz se puede apreciar vemos que la diagonal principal es 1, por lo cual se puede intuir que las varianzas serán homocedásticas Entonces con la submatriz P^* definida podemos transformar el modelo original. (Greene, 2012) Lo muestra con las siguientes matrices.

$$Y^* = P^*Y = \begin{bmatrix} Y_2 - pY_1 \\ Y_3 - pY_2 \\ \vdots \\ Y_t - pY_{t-1} \end{bmatrix}, X^* = P^*X = \begin{bmatrix} X_2 - pX_1 \\ X_3 - pX_2 \\ \vdots \\ X_t - pX_{t-1} \end{bmatrix}, \text{(3.6.133.)}$$

$$\begin{bmatrix} Y_2 - pY_1 \\ Y_3 - pY_2 \\ \vdots \\ Y_t - pY_{t-1} \end{bmatrix} = \begin{bmatrix} 1 & X_2 - pX_1 \\ 1 & X_3 - pX_2 \\ \vdots & \vdots \\ 1 & X_t - pX_{t-1} \end{bmatrix} \begin{bmatrix} \alpha(1-p) \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_3 \end{bmatrix} \text{ (3.6.134.)}$$

Estas matrices producen el siguiente modelo.

$$Y_t = \alpha(1-p) + X'_t\beta - X'_{t-1}\beta p + py_{t-1} + e_t \text{ (3.6.135.)}$$

Reorganizando (3.6.134.) obtenemos

$$Y_t - py_{t-1} = \alpha(1-p) + \beta(X'_t - X'_{t-1}p) + e_t \text{ (3.6.136.)}$$

Así es como se obtiene el modelo (3.6.128.). También podemos ver que en (3.6.135.) el término $\alpha(1-p)$ es el intercepto del modelo transformado. (3.6.136.) es la forma general del modelo (3.6.128.) anteriormente expresado.

El problema de utilizar el estimador del método Cochrane-Orcutt es la pérdida de la primera observación, lo que podría ocasionar problemas en la estimación sobre todo en modelos con muestras pequeñas. Por lo tanto, Sigbert Jon Prais y Christopher Blake Wisten propusieron el método **Prais-Wisten** como una mejora del método C-O. El método Prais-Wisten también utiliza un *p* conocido por lo tanto también asumimos que los errores siguen un esquema *AR(1)*.

La diferencia radica en que, en lugar de utilizar la matriz P^* utilizamos la matriz P para transformar las variables del modelo original. En consecuencia, las matrices (3.6.133.) y (3.6.134.) se convierten en:

$$Y^* = PY = \begin{bmatrix} \sqrt{1-p^2}Y_1 \\ Y_2 - pY_1 \\ Y_3 - pY_2 \\ \vdots \\ Y_t - pY_{t-1} \end{bmatrix}, X^* = PX = \begin{bmatrix} \sqrt{1-p^2}X_1 \\ X_2 - pX_1 \\ X_3 - pX_2 \\ \vdots \\ X_t - pX_{t-1} \end{bmatrix}, \text{ (3.6.137.)}$$

$$\begin{bmatrix} \sqrt{1-p^2}Y_1 \\ Y_2 - pY_1 \\ Y_3 - pY_2 \\ \vdots \\ Y_t - pY_{t-1} \end{bmatrix} = \begin{bmatrix} \sqrt{1-p^2} & \sqrt{1-p^2}X_1 \\ 1 & X_2 - pX_1 \\ 1 & X_3 - pX_2 \\ \vdots & \vdots \\ 1 & X_t - pX_{t-1} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_3 \end{bmatrix} \text{ (3.6.138.)}$$

(Greene, 2012) Comenta que las matrices (3.6.137.) son llamadas **diferencias parciales**, **cuasidiferencias** o **pseudodiferencias** y en esas matrices cada observación

esta transformada a excepción del primer dato, por lo tanto en muestras pequeñas el método Prais-Wisten los problemas de autocorrelación podrían volver a aparecer.

La matriz (3.6.138.) produce que la primera observación del modelo (3.6.136.) sea:

$$\sqrt{1-p^2}Y_1 = \alpha\sqrt{1-p^2} + \beta_2\sqrt{1-p^2}X_{21} + \cdots + \beta_k\sqrt{1-p^2}X_{k1} + e_1 \quad (3.6.139.)$$

Y para el resto de las observaciones del modelo se conserva el modelo (3.6.136.).

Podemos concluir entonces que ambas formas de estimar MCF **tienen como requisitos fundamentales que los residuos del modelo sigan un esquema *AR(1)* y conocer *p***, no obstante el cumplimiento de estos requisitos puede hacer que surja la siguiente pregunta ¿Cuál método es preferible de usar? En la teoría econométrica se puede encontrar que en muestras grandes la diferencia entre el método Cochrane-Orcutt y el método Prais-Wisten casi no se nota, sin embargo en muestras pequeñas es recomendable utilizar el método Prais-Wisten ya que puede mejorar la eficiencia de los estimadores, claramente teniendo cuidado que no aparezcan nuevamente los problemas ocasionados por la primera observación.

(Wooldrige, 2009) Compara las diferencias que existen entre los estimadores de MCO y los estimadores MCF, establece que los estimadores de MCF difícilmente pueden ser consistentes debido a que el supuesto de exogeneidad estricta se mantiene débilmente en estos métodos, por el contrario los estimadores de MCO son consistentes por la suposición de la ley de los grandes números. Además, las significancias individuales que producen los métodos de estimación mediante MCO y MCF podrían ser distintas, en ese caso se elegiría los estimadores de MCO. Por último, si los estimadores de MCF y MCO dan estimaciones parecidas entonces se opta por un estimador MCF si se demuestra que los estimadores de MCO tienen autocorrelación. (Wooldrige, 2009) Recomienda utilizar el método de Hausman para determinar si las diferencias entre ambos métodos de estimación son significativas. (Novales, 1998) Advierte que los estimadores del modelo transformado obtienen mejores propiedades que los de MCO, en consecuencia **solamente debemos sustituir los estimadores del modelo transformado en el modelo original, para obtener los residuos, la varianza del error y un coeficiente de determinación si se requiere**.

3.6.3.2.3. Métodos iterativos.

El cumplimiento del requisito que se conozca *p* puede tomarse más como una restricción debido a que en ciertos modelos no se conoce *p* entonces ¿Cómo utilizar estos métodos correctivos para en los modelos que no se conoce *p*? Esta clara desventaja de utilizar los MCF cuando no se conoce *p* ha sido solucionada por la teoría econométrica, la cual propone el uso del **método iterativo de C-O** para corregir la autocorrelación. El método iterativo de C-O tiene una variante denominada **método C-O en dos pasos**. La principal diferencia entre ambos métodos correctivos se centra en el número de veces que se repiten las regresiones recursivas. Para entenderlo, veamos primero los pasos que sigue el método iterativo de C-O.

Paso 1. **Teniendo el modelo $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$, cuyo término de error sigue un *AR(1)* especificado como $\mu_t = p_1 \mu_{t-1} + e_t$ que explica la autocorrelación en el modelo. Empezamos obteniendo los estimadores del modelo econométrico mediante MCO y cálculos los residuos.**

Paso 2. Con los residuos calculados estimamos el esquema *AR(1)* $\hat{\mu}_t = \hat{p}_1 \mu_{t-1} + e_t$.

Paso 3. Con $\hat{p}_1$ transformamos el modelo econométrico en:

$$(Y_t - \hat{p}_1 Y_{t-1)} = \beta_1(1 - \hat{p}_1) + \beta_2(X_{2t} - \hat{p}_1 X_{2t-1}) + \beta_3(X_{3t} - \hat{p}_1 X_{3t-1}) + (\mu_t - \hat{p}_1 \mu_{t-1}).$$

Paso 4. Estimar los estimadores del modelo transformado mediante MCO y calcular sus residuos.

Paso 5. En consecuencia que no estamos seguros si $\hat{p}_1$ realmente estima el verdadero valor de p_1, estimamos un esquema *AR(1)* usando los residuos del modelo transformado.

Paso 6. Volver a transformar el modelo transformado usando el nuevo valor de *p*. Repetir hasta que los residuos no presenten autocorrelación.

Este método engorroso puede resumirse en el **método iterativo de Cochrane-Orcutt en dos pasos**. En términos sencillos, (Pérez L., 2012) Describe los pasos a realizar estas regresiones sucesivas.

Paso 1. Considerando el modelo $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t$, asumimos que el término de error sigue un *AR(1)* entonces estimamos p_1 en el modelo $\mu_t = p_1 \mu_{t-1} + e_t$ y se transforma el modelo en:

$$Y_t - pY_{t-1} = (1-p)\beta_1 + \beta_2(X_{2t} - pX_{2t-1}) + \beta_3(X_{3t} - pX_{3t-1}) + e_t$$

(3.6.128.)

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + e_t \text{ (3.6.129.)}$$

Paso 2. Del modelo (3.6.128.) volvemos a asumir que el término de error sigue un esquema *AR(1)*, entonces estimamos p_2 del modelo $e_t = p_2 + e_{t-1} + v_t$ y transformamos el modelo (3.6.129.)

$$Y_t^* - p_2 Y_{t-1}^* = \beta_1^*(1-p_2) + \beta_2(X_{2t}^* - p_2 X_{2t-1}^*) + \beta_3(X_{3t}^* - p_2 X_{3t}^*) + v_t \text{ (3.6.140.)}$$

En la práctica no tenemos que realizar todas esas iteraciones ya que los programas estadísticos vienen equipados para realizar tantas iteraciones como el software lo considere necesario

A pesar que realicemos tantas iteraciones como creamos conveniente, la primera observación se ha omitido, no obstante, para recuperar la primera observación se calcula lo siguiente:

$$\sqrt{1-p^2}Y_1 = \sqrt{1-p^2}(\beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \beta_4 X_{41} + \cdots + \beta_k X_{k1} + \mu_1) \text{ (3.6.141.)}$$

Cuando recuperamos la primera observación con (3.6.141.), se está ejecutando el método iterativo de Prais-Wisten.

3.6.3.2.4. Método Newey-West.

Esta sección parte de la pregunta ¿Es posible que un modelo contenga heterocedasticidad y autocorrelación? Y de ser así ¿Qué método emplear para corregirlo? La respuesta a la primera pregunta es un rotundo sí, y para solucionar estos modelos emplearíamos los **errores estándar consistentes con heterocedasticidad y autocorrelación (CHA),** o por sus siglas en ingles **HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors** o simplemente **errores Newey-West**.

Este método resulta ser una extensión de los errores robustos de White y siguen un proceso parecido. El estimador de Newey-West es:

$$\widehat{Q_*} = S_0 + \frac{1}{t}\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l \mu_t \mu_{t-1}(X_t X'_{t-1} + X_{t-1} X'_t) \text{ (3.6.142.)}$$

$$w_l = 1 - \frac{l}{L+1} \text{ (3.6.143.)}$$

(Greene, 2012) Establece que la ventaja del estimador (3.6.142.) es consistente y robusto y sobre todo para perturbaciones autocorrelacionadas no especificadas, donde *L* es el retardo máximo que debe determinarse previamente. En la práctica, los programas estadísticos tienen comandos para obtener los errores CHA.

3.7. Ejemplo con STATA sobre Estimación con MCO y Verificación del Cumplimiento de los Supuestos y Medidas Correctivas

A continuación, se presentarán dos ejemplos de cómo realizar modelos econométricos en el programa estadístico STATA 15. Se realizarán dos modelos econométricos, uno con datos de corte transversal y el segundo con datos de series temporales.

3.7.1. Ejemplo con el uso de datos de corte transversal.

En este ejemplo se mostrará cómo construir un modelo econométrico mediante MCO usando datos de corte transversal, desde la especificación, estimación, evaluación e interpretación de un modelo que explique algunas características sobre el trabajador independiente, centrándose en sus niveles de ingresos. En este ejemplo seguiremos los pasos que (Gujarati & Porter, 2010) Han planteado.

No se hablará en profundidad del planteamiento del problema, tampoco de la recolección del marco teórico para explicar con mayor profundidad sobre la especificación, estimación, evaluación y la interpretación del modelo.

3.7.1.1. Problema de la investigación.

3.7.1.1.1. Planteamiento del problema.

Según (RPP, 2017) Los trabajadores independientes que se encuentran dentro del grupo de trabajadores informales y representan el 41% de la PEA. Existen dos motivos por los cuales sucede esto, la primera razón se debe a la amplia gama de actividades que producen los trabajadores independientes y la segunda razón es la alta concentración de trabajadores independientes como informales. Los trabajadores independientes por lo general presentan problemas muchas veces ligados por su propia condición de ser independientes, un problema muy común son los aportes que recibe del Sistema de

Pensiones al momento de su jubilación, debido a que el trabajador independiente no está obligado a aportar y no suele hacerse una cultura de ahorro.

Los trabajadores independientes suelen tener menores garantías que los asalariados, en palabras de (Flores C., 2020) Los trabajadores de las mypes podrán contar con protección de salud y vida desde el primer día de trabajo y ya no desde el cuarto año, así lo plantea el Decreto de Urgencia N°044-2019 publicado en El Peruano. Se trata aproximadamente de 300000 empresas que tendrán que contratar el seguro Vida Ley para sus trabajadores y cerca de 400000 trabajadores de las micro, pequeña y mediana empresas formales serán beneficiados, con lo que se espera que 3.7 millones de trabajadores cuenten con este seguro.

Debido a algunas disparidades entre los trabajadores independientes y los asalariados, la SUNAT ha decretado que los trabajadores independientes pueden estar exentos de pagar el Impuesto a la Renta. (Gestión, 2020) Explica que la SUNAT ha emitido una resolución en la cual, los trabajadores independientes que ganen hasta S/ 37,625 durante el año 2020 o S/ 3,135 al mes no pagarán Impuesto a la Renta, sin embargo en el año 2019, bastaba con ganar hasta S/ 36,750 al año o S/ 3,602 al mes para no pagar Impuesto a la Renta. La SUNAT asume que los trabajadores independientes ganan menos que los trabajadores en planilla, por ello es que además del aumento del tope, les brinda otro beneficio a los trabajadores independientes haciendo que cuenten con la deducción adicional del 20% de sus ingresos brutos anuales. Estos beneficios prometen ayudar a los trabajadores independientes en aliviar la carga tributaria que muchas veces es una causante de la promoción del empleo informal independiente.

(Costa A., 2018) Ha expuesto que en Lambayeque la población ocupada en condición de trabajador independiente fue del 42.2% durante el 2017, en Callao se obtuvo el menor porcentaje siendo 31.2% y en Loreto el mayor porcentaje con 51.9%. Solamente el 16.5% de la población ocupada en condición de trabajador independiente trabaja como persona Jurídica o Natural, de los cuales apenas el 7.6% es población con grado de instrucción superior y está registrado como Persona Jurídica, mientras que el 0.9% tiene secundaria y es Persona Jurídica. Estas cifras indican que a menudo los trabajadores independientes no están registrados, por lo tanto muchos de ellos trabajan como trabajadores independientes informales. Los trabajadores independientes dependen de muchos factores para iniciar un negocio o una actividad, entre ellos (Costa A., 2018) Identifica como los principales motivos a la necesidad económica, el deseo de ser

independiente, mayores ingresos, no encontrar trabajo como asalariado. Durante el año 2017, el 47.9% se dedicó a la prestación de servicios, el 32.2% a la compra y venta de mercadería, el 15% a la producción y extracción, el 0.4% a la producción y comercio y el 4.5% a otras actividades, así informo (Costa A., 2018).

3.7.1.1.2. Planteamiento de la pregunta.

¿Cuáles son los determinantes que influyeron sobre el trabajador independiente en el distrito de Chiclayo durante el año 2018?

3.7.1.1.3. Objetivo general y objetivos específicos.

- Objetivo general.
  - Determinar cuáles han sido los factores que han influido sobre el trabajador independiente en el distrito de Chiclayo durante el año 2018.
- Objetivos específicos.
  - Analizar el comportamiento de los trabajadores independientes en el distrito de Chiclayo durante el año 2018.
  - Medir cómo influyen los factores en los trabajadores independientes en el distrito de Chiclayo durante el año 2018.

3.7.1.2. Identificar el marco teórico.

Los trabajadores independientes y la seguridad social en el Perú por (Casalí & Pena, 2012)

(Casalí & Pena, 2012) Definen al trabajador independiente como un elemento perteneciente a un grupo muy heterogéneo, con una alta incidencia en la informalidad, como efecto de esa heterogeneidad los trabajos que laboran los trabajadores independientes están vinculados a una amplia variedad de actividades y de estas pueden recibir desde altos ingresos hasta ingresos muy pobres. El hecho que un trabajador sea independiente no necesariamente quiere decir que es informal, es difícil definir cuando cumple la condición de ser informal, ya que no existe un acuerdo claro. La definición más usada es la que caracteriza a un trabajador informal como una persona que labora al margen de las regulaciones laborales.

Una característica fundamental de los trabajadores independientes es que son componentes centrales de la economía informal, por ello es frecuente relacionarlo como un indicador clave de la informalidad laboral. La OIT propone las siguientes definiciones:

- **Empleo informal**, es el conjunto de puestos de trabajo informales desarrollados tanto en empresas formales como informales.
- **Economía informal**, son todas las actividades económicas desarrolladas por trabajadores y unidades productivas que insuficientemente están contempladas por los sistemas formales o no están en absoluto.
- **Informalidad**, Incluye las relaciones de producción como relaciones de empleo, ello implica incluir el término sector informal en la economía informal para considerar a todos los trabajadores que no están plenamente cubiertos por las leyes sobre el trabajo.

Algunas características de los trabajadores independientes se muestran a continuación:

Durante el 2010, el 65% era trabajador por cuenta propia, el 11% era trabajador familiar no remunerado y el 24% era empleador. Además el trabajador independiente no tiene un comportamiento homogéneo en cuento a su sexo se refiere, la mayoría de los hombres trabajan por cuenta propia y son empleadores, por otro lado la mayoría de las mujeres son trabajadoras familiares no remuneradas. Tantos hombres como mujeres tienen niveles de cobertura tanto en las aportaciones a pensiones como afiliaciones a EsSalud; no obstante, las mujeres tienen una mayor protección en materia de salud y una menor cobertura de aportaciones, respecto a los hombres. La edad también muestra una distribución notoria entre los trabajadores independientes, ya que la mayoría se encuentra entre los 25 años y 44 años y muy pocos logran ser aportantes a pensiones, la mayoría solo se encuentra afiliado a EsSalud, de hecho es más probable que un trabajador de 55 años sea afiliado a EsSalud que alguien de menor edad.

En este grupo también se ha encontrado una concentración de personas que no tienen grado de instrucción o solamente primaria que representa el 41.7%, mientras los trabajadores independientes con superior universitaria son muy pocos llegando a representar el 15.4%. La ubicación geográfica también parece ser otra característica importante en cuanto a los trabajadores independientes, ya que el 61.3% labora en zonas urbanas y el 38.7% en zonas rurales.

En cuanto a su nivel de ingresos, el 53% de los trabajadores independientes han recibido ingresos mensuales menores a los S/ 500 y solamente un 38% de los empleadores logran superar los S/ 1500 mensuales. La mayor parte de los trabajadores independientes

se ocupa como agricultor, ganadero o pescador, siendo el 38.5% de este grupo; en segundo lugar están los vendedores con un 24.4%; los encargados de brindar servicios solamente representan un 11.4%; mientras que un pequeño porcentaje realizan actividades dedicadas a profesionales, técnicos y afines, tal parece que los trabajadores que se dedican a las actividades profesionales tienen una alta probabilidad de tener afiliación a EsSalud y a ser aportantes a pensiones mientras que los trabajadores dedicados a las tres primeras actividades descritas tienen probabilidades altas a estar afiliados al SIS con subsidio.

Determinantes del desempeño del trabajador independiente y la microempresa familiar en el Perú por (Yamada, 2009)

Debido al poco marco teórico que se ha escrito sobre los empleos de trabajadores independientes resulta absolutamente difícil capturar datos precisos sobre el desempeño de los trabajadores independientes y sobre todo la microempresa familiar. En 1990 Smith y Stelcner determinaron que las empresas tienen mayor probabilidad de obtener ingreso si cuentan con local fijo y el tiempo en el mercado impacta de manera positiva sobre los ingresos percibidos. Han tomado como variable explicativa a los costos/gastos de la propia firma para explicar los ingresos de la misma, de igual forma los capitales y el número de horas también son un aporte positivo sobre los ingresos.

Otras variables como el nivel de instrucción y la experiencia también tienen un aporte significativo sobre el nivel de ingresos, ya que a mayor grado de instrucción de los empresarios muestra un impacto positivo sobre el nivel de ingresos que reciben, por otro lado la experiencia puede tener una influencia positiva cuando el coeficiente es lineal y negativo cuando es cuadrático, esto quiere decir que los empresarios con mayor experiencia tienen mayores ingresos pero los retornos crecen menos.

Un modelo de supervivencia realizado por López-García y Puente en 2006 sobre empresarios en España demostró que el tamaño inicial de las empresas aumenta las posibilidades que estas sobrevivan más tiempo, también el tipo de sector donde opera resulta un factor importante sobre la supervivencia de las empresas, ya que en los sectores concentrados las probabilidades de sobrevivir son mayores mientras que en sectores dinámicos las empresas tienen menores probabilidades de sobrevivir.

3.7.1.3. Especificación del modelo econométrico.

Con base al marco teórico anteriormente expuesto, podemos especificar el modelo econométrico que se utilizará para explicar los determinantes de los trabajadores independientes y también, cuáles han sido las características más sobresalientes durante el año 2018 en el distrito de Chiclayo. Siendo el modelo especificado el siguiente:

$$G_i = \beta_1 + \beta_2 I_i + \beta_3 C_i + \beta_4 N_i + \mu_i \text{ (3.7.1.)}$$

Donde:

G_i: Ganancia total neta mensual que perciben los trabajadores independientes.

I_i: Ingresos que percibe el trabajador independiente mensualmente.

C_i : Gastos que realiza el trabajador independiente de forma mensual. Está compuesto por los gastos en el establecimiento, gastos en mano de obra y gastos según el capítulo 50 de ENAHO sobre ingresos y gastos definidos por el INEI.

N_i: Número de personas que trabajan en el establecimiento a cargo del trabajador independiente. Pueden ser asalariados, familiares no remunerados o el mismo trabajador independiente.

μ_i es el error aleatorio del modelo, donde se incluyen variables relacionadas a otras características de los trabajadores independientes como la ubicación geográfica, el sexo, la afiliación a EsSalud o al SIS, entre otras. El subíndice *i* toma un valor distinto para cada trabajador independiente que han percibido ganancias totales netas, ingresos mensuales, gastos mensuales y tienen a trabajadores en su establecimiento, descartando a los trabajadores que no perciban alguna de estas variables.

Ya que es posible que haya diferencias entre los trabajadores debido a sus actividades que realizan, se ejecutará tres regresiones, una para cada tipo de actividad que han realizado los trabajadores independientes. Sin embargo es necesario mencionar que algunos trabajadores independientes pueden realizar más de un tipo de actividad.

Finalmente, en la siguiente lista se especificara cuáles son las variables para cada regresión.

- $G_i(e25t3), I_i(e14t), C_i(e16t + e25t1 + e25t2), N_i(e8a) \rightarrow$ Actv. Productiva /Extractiva.
- $G_i(e25t3), I_i(e17t), C_i(e19t + e25t1 + e25t2), N_i(e8a) \rightarrow$ Actv. Comercial.

- $G_i(e25t3), I_i(e20t), C_i(e21t + e25t1 + e25t2), N_i(e8a) \rightarrow$ Actv. Prestadora de servicios.

3.7.1.4. Acceso a la base de datos.

Para construir los datos que serán empleados en la estimación del modelo y en el análisis de los trabajadores independientes del distrito de Chiclayo durante el 2018, se usará el módulo 77 de la ENAHO, el cual trata sobre los ingresos de los trabajadores independientes. El módulo 77 se descarga ingresando al siguiente URL http://iinei.inei.gob.pe/microdatos/Consulta_por_Encuesta.asp, donde aparecerá la siguiente ventana.

Figura 3.1. Microdatos de INEI.

Haremos clic en "Consulta por Encuestas", seleccionaremos después "ENAHO Metodología Actualizada", "Condiciones de Vida y Pobreza-ENAHO", el año y el

Figura 3.2. Consulta por Encuesta de ENAHO.

periodo, que para este modelo serán 2018 y anual respectivamente. Nos aparecerá la siguiente ventana.

En la penúltima columna podemos encontrar un archivo PDF llamado “Ficha”, el cual contiene temas específicos sobre los cuestionarios de la ENAHO, técnicas de muestreo, tamaño de la muestra, entre otros. Buscamos el módulo 77 y hacemos clic en el icono a la derecha de SPSS para descargar el archivo del módulo en formato STATA de la ENAHO.

CodigoConglomerado_6_digitos	8/04/2019 19:09	Adobe Acrobat D...	175 KB
ConglomeA6digitos	8/04/2019 15:07	Archivo DO	1 KB
CUESTIONARIO.04 2018	8/04/2019 18:33	Adobe Acrobat D...	176 KB
Diccionario_2018	8/04/2019 10:31	Adobe Acrobat D...	10,443 KB
enaho04-2018-1-preg-1-a-13	11/04/2019 10:08	Archivo DTA	2,902 KB
enaho04-2018-2-preg-14-a-22	11/04/2019 10:09	Archivo DTA	4,230 KB
enaho04-2018-3-preg-23	11/04/2019 10:09	Archivo DTA	26,979 KB
enaho04-2018-4-preg-24	11/04/2019 10:10	Archivo DTA	523 KB
enaho04-2018-5-preg-25	11/04/2019 10:10	Archivo DTA	3,814 KB
Ficha Tecnica_2018	6/04/2019 19:41	Adobe Acrobat D...	778 KB

Figura 3.3. Datos descargados del módulo 77 de ENAHO.

Vemos que hay 5 archivos STATA en este módulo, para efecto de este ejemplo solamente se utilizara los archivos que se refieran al ENAHO como “enaho04-2018”, también se observa un archivo PDF llamado “CUESTIONARIO.04 2018” este archivo nos muestra las preguntas en el cuestionario y su importancia radica en que nos ayuda a guiarnos en el archivo STATA, en el cuestionario encontramos 25 preguntas cuyas respuestas están distribuidas en 4 archivos de STATA tal como se muestra en la figura 3.3. Al abrir el cuestionario y primer archivo de STATA vemos la siguiente figura:

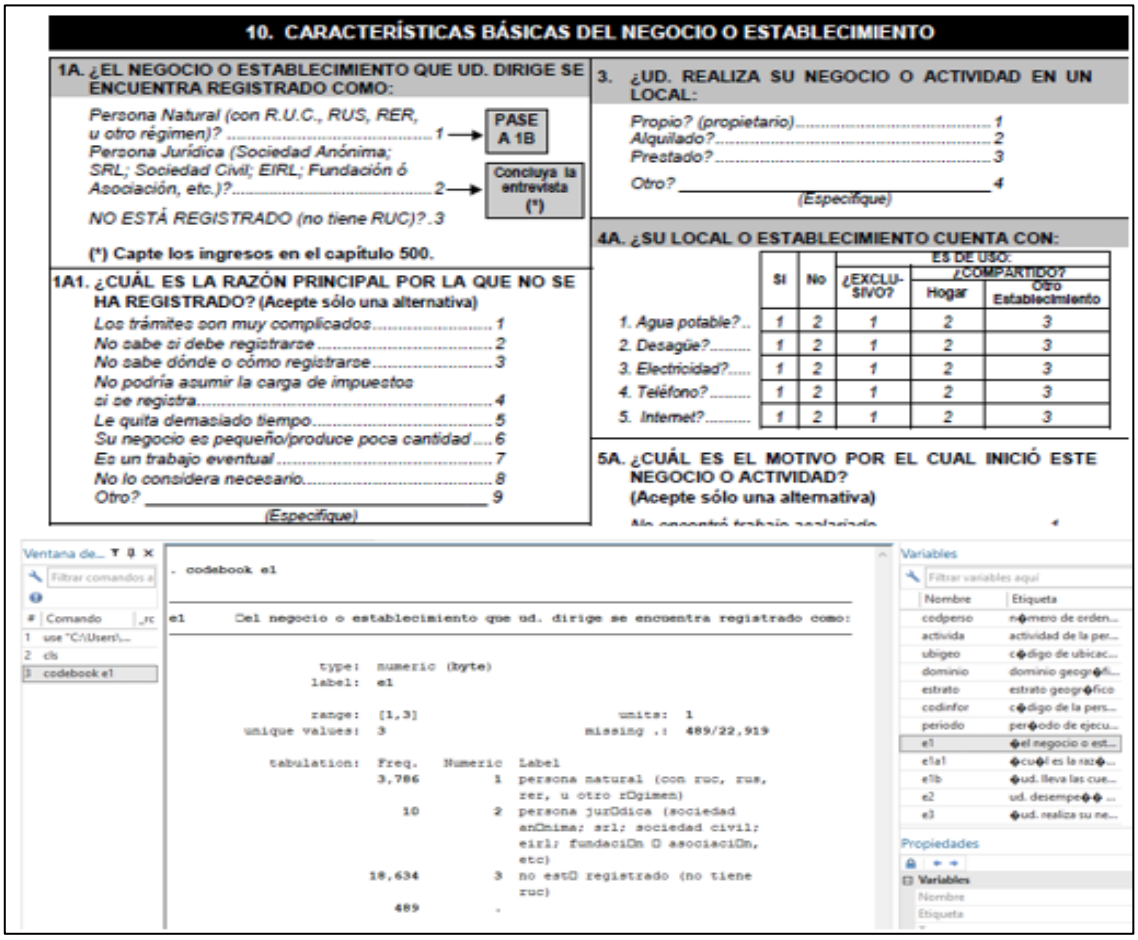


Figura 3.4. Cuestionario y archivo STATA del módulo 77.

La primera pregunta del cuestionario trata sobre la condición de registro del establecimiento donde la persona trabaja y tiene tres posibles respuestas: “Persona Natural”, “Persona Jurídica”, “No está registrado”. El archivo STATA con la ayuda del comando **codebook**, la variable ***e1*** muestra cómo están distribuidas las observaciones con respecto a la primera pregunta del cuestionario. El comando **codebook** es un comando muy útil que ordena la instrucción a STATA de darnos un resumen sobre una variable desde su tipo, etiqueta, valores, valores perdidos, frecuencia, etc. En STATA hay dos tipos de variables, numéricas y string, en el caso de la figura 3.4. La variable es numérica de tipo byte, este tipo de variables muestra valores numéricos que representan una característica o condición como se ve en el ejemplo, la base de datos puede ser de color azul o negro. Otros tipos de variable numérica son: “int”, “loung”, “doublé” y en la base de datos se muestran de color negro, por otro lado, las variables string son aquellas variables que contienen texto y se ven en la base de datos de color rojo. Tal como se ve en las siguientes figuras.

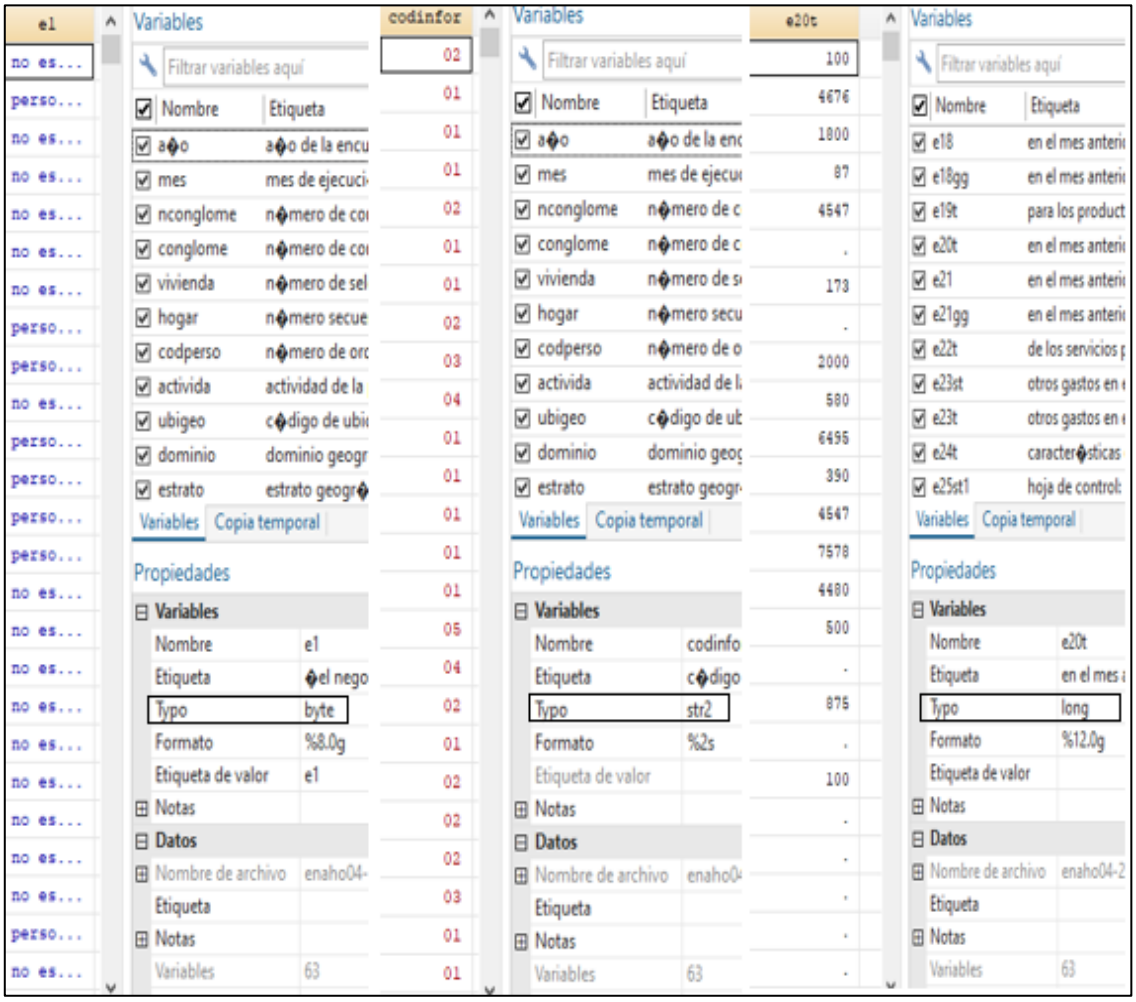


Figura 3.5. Tipos de variables en STATA.

En la figura 3.3. También podemos ver un archivo PDF llamado "Diccionario_2018" el cual contiene definiciones y conceptos sobre las variables, módulos, entre otros. Aunque para tener una definición más completa es mejor revisar la web de la ENAHO, la cual es la siguiente: https://webinei.inei.gob.pe/anda_inei/index.php/catalog/central/about. En la figura 3.6. Podemos ver la ventana del URL, hacemos clic en "Ver investigaciones" que está debajo del vínculo "Encuesta a Hogares".

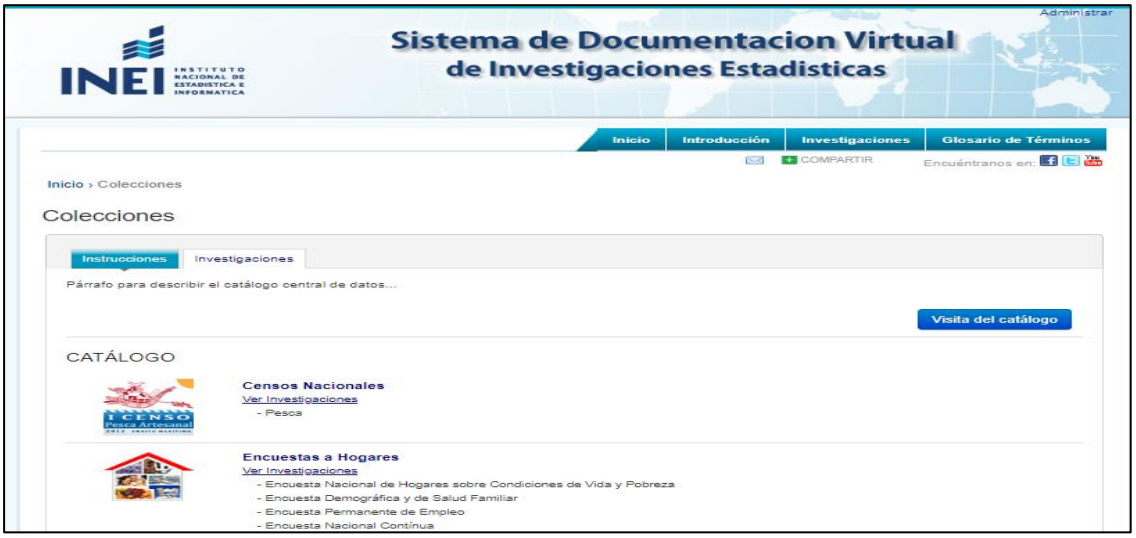


Figura 3.6. Sistema de Documentación Virtual de Investigaciones Estadísticas (1).

En la figura 3.7. Se muestra una ventana de todas las investigaciones de Encuesta a Hogares, filtremos el intervalo de años que se quiere investigar y posteriormente buscamos la investigación que deseamos, en este caso filtramos el año 2018 y elegimos "Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza 2018".

Figura 3.7. Sistema de Documentación Virtual de Investigaciones Estadísticas (2).

Esta ventana muestra detalles más específicos sobre la ENAHO 2018 y podemos encontrar definiciones y conceptos de las variables. Para ello hacemos clic en "Descripción de Variables" y elegimos el módulo con el cual estamos trabajando, en nuestro caso elegimos el módulo 77 que está representado en la pestaña "Enaho04-2018" y al igual que los archivos de STATA también está dividido en cuatro partes.

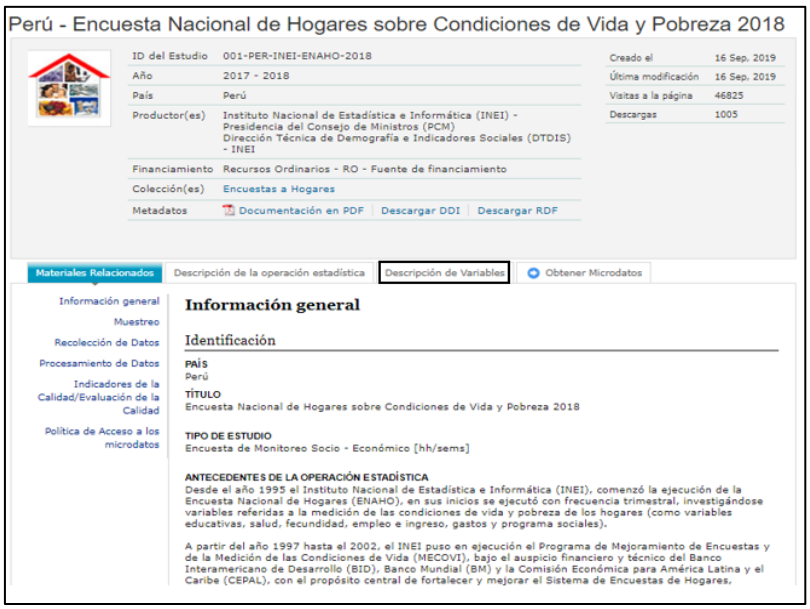


Figura 3.8. Encuesta Nacional de Hogares sobre Condiciones de vida y Pobreza 2018.

Es necesario tener cuidado en los módulos, ya que las observaciones usadas no siempre son las mismas en otros módulos. En la figura 3.9. Podemos inferir que este archivo STATA contiene información sobre el negocio que dirige el trabajador independiente, por ello podemos inferir que las observaciones serán los dueños de los negocios.

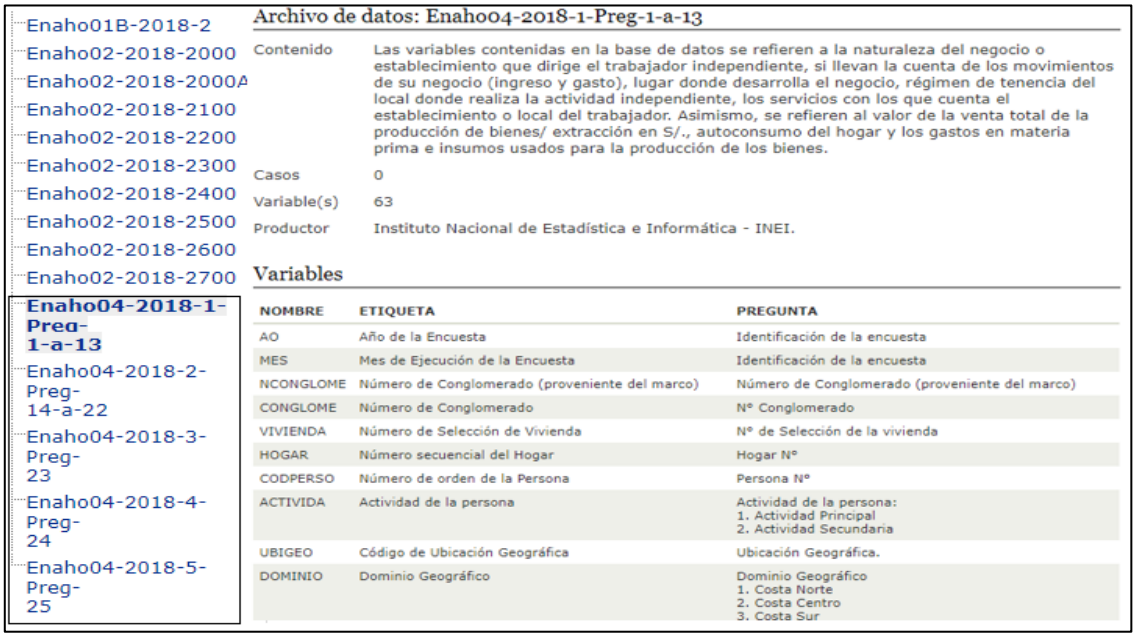
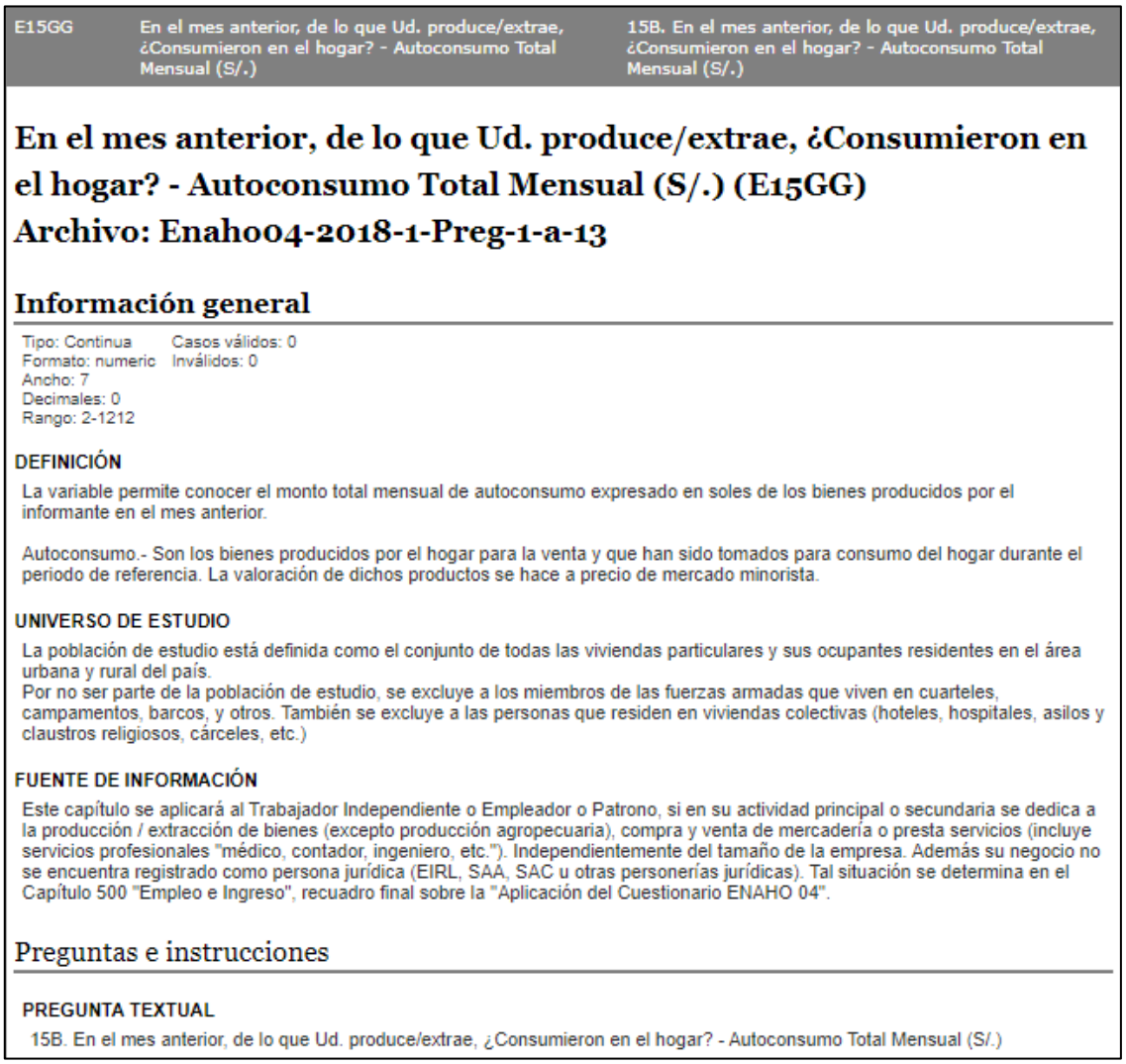


Figura 3.9. Enaho04-2018.

Para poder visualizar la definición de una variable basta con hacer clic en la variable de interés. Siguiendo con el ejemplo anterior vamos a ver cuál es la definición de la variable ***e15GG***, la cual trata sobre el monto de autoconsumo total en el mes pasado para los trabajadores dedicados a la producción o extracción.



E15GG	En el mes anterior, de lo que Ud. produce/extrae, ¿Consumieron en el hogar? - Autoconsumo Total Mensual (S/.)	15B. En el mes anterior, de lo que Ud. produce/extrae, ¿Consumieron en el hogar? - Autoconsumo Total Mensual (S/.)

En el mes anterior, de lo que Ud. produce/extrae, ¿Consumieron en el hogar? - Autoconsumo Total Mensual (S/.) (E15GG)
Archivo: Enaho04-2018-1-Preg-1-a-13

Información general

Tipo: Continua — Casos válidos: 0
Formato: numeric — Inválidos: 0
Ancho: 7
Decimales: 0
Rango: 2-1212

DEFINICIÓN

La variable permite conocer el monto total mensual de autoconsumo expresado en soles de los bienes producidos por el informante en el mes anterior.

Autoconsumo.- Son los bienes producidos por el hogar para la venta y que han sido tomados para consumo del hogar durante el periodo de referencia. La valoración de dichos productos se hace a precio de mercado minorista.

UNIVERSO DE ESTUDIO

La población de estudio está definida como el conjunto de todas las viviendas particulares y sus ocupantes residentes en el área urbana y rural del país.
Por no ser parte de la población de estudio, se excluye a los miembros de las fuerzas armadas que viven en cuarteles, campamentos, barcos, y otros. También se excluye a las personas que residen en viviendas colectivas (hoteles, hospitales, asilos y claustros religiosos, cárceles, etc.)

FUENTE DE INFORMACIÓN

Este capítulo se aplicará al Trabajador Independiente o Empleador o Patrono, si en su actividad principal o secundaria se dedica a la producción / extracción de bienes (excepto producción agropecuaria), compra y venta de mercadería o presta servicios (incluye servicios profesionales "médico, contador, ingeniero, etc."). Independientemente del tamaño de la empresa. Además su negocio no se encuentra registrado como persona jurídica (EIRL, SAA, SAC u otras personerías jurídicas). Tal situación se determina en el Capítulo 500 "Empleo e Ingreso", recuadro final sobre la "Aplicación del Cuestionario ENAHO 04".

Preguntas e instrucciones

PREGUNTA TEXTUAL

15B. En el mes anterior, de lo que Ud. produce/extrae, ¿Consumieron en el hogar? - Autoconsumo Total Mensual (S/.)

Figura 3.10. Definición de la variable *e15GG*.

Podemos observar una presentación más detallada y precisa del significado de la variable e15GG.

Ya que se ha definido a los trabajadores independientes del distrito de Chiclayo como la población objetivo del modelo anteriormente especificado, se procede a visualizar sus características y compararlas con otros distritos de Chiclayo. Para ello debemos seleccionar solamente las observaciones que están en la provincia de Chiclayo, para esto utilizaremos la variable ***ubigeo*** que se encuentra en cada uno de los archivos de STATA. La variable ***ubigeo*** muestra el **código UBIGEO** el cual es muy usado para

determinar la ubicación geográfica de una determinada observación, está compuesta por 6 dígitos, los dos primeros hacen referencia al departamento por esta razón su rango comprende desde el 01 hasta el 24; los dos siguientes representan a la provincia de cada departamento y los dos últimos dígitos son los distritos de cada provincia. El UBIGEO del departamento de Lambayeque es 14, el UBIGEO de sus provincias de Chiclayo, Ferreñafe y Lambayeque son 1401, 1402 y 1403 respectivamente y el UBIGEO del distrito de Chiclayo es 140101. Los códigos se pueden encontrar en el siguiente URL. http://webinei.inei.gob.pe:8080/sisconcode/proyecto/index.htm?proyectoTitulo=UBIGEO&proyectoId=3

Debido a que en los archivos de STATA la variable ***ubigeo*** es una variable string, no se puede utilizar para nuestro fin. Entonces, debemos transformar la variable string en una variable numérica con la ayuda del comando **destring** que tiene dos posibles opciones **gen** y **replace**. La opción **gen** convierte una variable string en una variable numérica creando una nueva variable la cual será el formato numérico de la variable string que deseamos convertir, mientras la opción **replace** reemplaza en la variable string seleccionada pero en formato numérico.

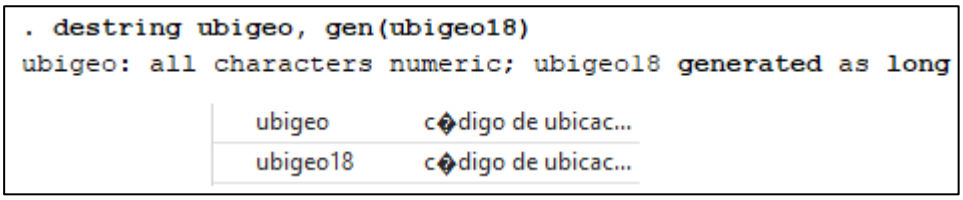


Figura 3.11. Transformación de una variable string en una variable numérica.

Ahora para seleccionar los datos que pertenecen a la provincia de Chiclayo utilizaremos el comando **keep**, el cual mantiene en la base de datos a las variables u observaciones que cumplan una característica que está ordenada por el condicional **if**.

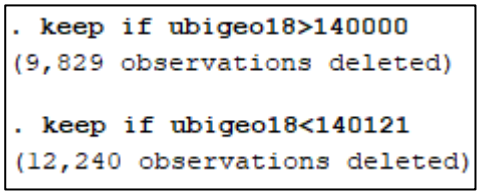


Figura 3.12. Selección de observaciones pertenecientes a la provincia de Chiclayo (1).

Observe como en la figura 3.12. Ha empleado la variable generada ***ubigeo18*** para la selección de las observaciones. Para comprobar que solamente tenemos observaciones

que pertenecen a la provincia de Chiclayo podemos utilizar otra vez el comando **codebook** con la variable **ubigeo**.

```
ubigeo                                                  c□digo de ubicaci□n geogr□fica
-------------------------------------------------------------------------------

                type:  string (str6)

        unique values:  20                       missing "":  0/850

             examples:  "140101"
                        "140105"
                        "140106"
                        "140112"
```

Figura 3.13. Selección de observaciones pertenecientes a la provincia de Chiclayo (2).

Podemos ver que tras el uso del comando **keep** hemos mantenido 850 observaciones y ahora la variable ***ubigeo*** cuenta con 20 valores en alusión a cada uno de los distritos registrados que conforman la provincia de Chiclayo en la variable ***ubigeo***. Para examinar con mayor detalle, podemos construir una tabla con el comando **tabulate** o su abreviatura **tab**; este comando muestra una tabla con frecuencia, porcentaje y porcentaje acumulado de cada uno de los valores que conforman la variable, en el caso de la variable ***ubigeo*** mostrará los estadísticos anteriormente nombrados de cada uno de los distritos que han sido registrados.

```
. tab ubigeo
```

c□digo de ubicaci□n geogr□fica	Freq.	Percent	Cum.
140101	246	28.94	28.94
140102	15	1.76	30.71
140103	16	1.88	32.59
140104	6	0.71	33.29
140105	167	19.65	52.94
140106	119	14.00	66.94
140107	4	0.47	67.41
140108	54	6.35	73.76
140109	4	0.47	74.24
140110	6	0.71	74.94
140111	5	0.59	75.53
140112	53	6.24	81.76
140113	16	1.88	83.65
140114	6	0.71	84.35
140115	17	2.00	86.35
140116	15	1.76	88.12
140117	15	1.76	89.88
140118	33	3.88	93.76
140119	21	2.47	96.24
140120	32	3.76	100.00
Total	850	100.00	

Figura 3.14. Selección de observaciones pertenecientes a la provincia de Chiclayo (3).

La figura 3.14. Muestra que hay una alta concentración de la muestra en torno a los distritos Chiclayo, JLO y La Victoria, ya que solo estos tres distritos representan el 62.6% de la muestra; por otro lado, los distritos Lagunas, Nueva Arica, Oyotun, Picsi, Puerto Eten y Santa Rosa conforman al 3.66% de la muestra. Con estos datos se puede notar la existencia de una brecha muy profunda en cuanto a distribución de la población se refiere. Una forma de saber la condición de formalidad o informalidad es revisando si el establecimiento está registrado como Persona Natural, Persona Jurídica o no; para ello el comando **tab** mostrará cómo están distribuidos los trabajadores independientes en cuanto a su condición de registro o no en cada distrito. Las variables ***ubigeo*** y ***e1*** serán requeridas; la primera es requerida para ordenar a STATA que muestre los distritos y la segunda es necesaria para indicar a STATA que queremos que nos informe sobre la condición de registro del establecimiento. Recuerde que el cuestionario nos brinda información sobre las variables.

```
. tab ubigeo e1
```

código de ubicación geográfica	establecimiento que ud. dirige se encuentra registrado como: persona n	no está r	Total
140101	55	188	243
140102	1	14	15
140103	1	15	16
140104	1	5	6
140105	23	144	167
140106	14	105	119
140107	0	4	4
140108	1	53	54
140109	0	3	3
140110	0	6	6
140111	1	3	4
140112	8	44	52
140113	0	16	16
140114	0	6	6
140115	3	14	17
140116	3	12	15
140117	1	14	15
140118	6	27	33
140119	0	21	21
140120	2	29	31
Total	120	723	843

Figura 3.15. El establecimiento está registrado o no.

Podemos notar según la figura 3.15. Que el 85.77% de los trabajadores independientes trabajan en un establecimiento que no está registrado ni como Persona Natural ni tampoco como Persona Jurídica, es decir se consideran como empleos informales. Con estos datos podemos concluir que la informalidad puede estar generalizada entre los trabajadores independientes de los distintos distritos que

conforman la provincia de Chiclayo. En cuanto al distrito de Chiclayo, el 77.36% no está registrado y puede operar en el sector informal. Con el comando **tab** también podemos indicar al programa STATA que genere un cuadro resumen sobre algunos estadísticos descriptivos, esta función es muy usada en variables continuas como ingresos, gastos, costos, pesos, etc. Veamos el siguiente ejemplo que hace uso del comando **tab** y la opción **sum()** para crear una tabla que contenga información sobre el nivel de ventas en soles mensuales en cuanto a los trabajadores que se dedican a producir/extraer en cada distrito. Las variables usadas serán: ***ubigeo*** para usar a los distritos registrados como categorías, y la variable ***e14t*** que representa a los ingresos por ventas de los trabajadores dedicados a producir y/o extraer.

```
. tab ubigeo, sum( e14t)
```

c□digo de ubicaci□n geogr□fica	Summary of en el mes anterior, de lo que ud. produce/extrae □a cu□nto ascendieron sus venta		
	Mean	Std. Dev.	Freq.
140101	2046.8824	2793.474	34
140102	1011.6667	1309.736	3
140103	84	10.839742	5
140104	700	0	2
140105	6003.9091	16611.728	11
140106	2841.75	3441.2665	8
140108	201.3	180.1734	10
140112	3487.25	3743.7448	4
140113	900	0	1
140114	20	0	1
140115	554	440.71533	5
140116	646.66667	999.2664	3
140118	383	263.86107	4
140119	2500	2165.6408	3
140120	187.5	130.81475	2
Total	2023.1771	6021.7307	96

Figura 3.16. Nivel de ventas en cada distrito de los trabajadores independientes dedicados a producir/extraer.

En la figura 3.16. Se nota una tabla que muestra tres estadísticos descriptivos, los cuales son: promedio, desviación estándar y la frecuencia, respectivamente en cada columna, para cada distrito registrado de la provincia de Chiclayo. En el distrito de Chiclayo, en promedio cada trabajador independiente que se dedica a producir o extraer obtiene S/ 2046.90 en ventas mensuales. Por otro lado, en el distrito JLO los trabajadores independientes que se dedican a producir o extraer obtienen S/ 6003.90 en promedio, siendo este el mayor de todos los distritos.

(Escobar M., Fernández M., & Bernardi, 2012) Recomiendan usar la preinstrucción **bysort**, con el fin de obtener estadísticos descriptivos en función de las categorías o valores de dos o más variables cuantitativas y la opción **sum** permite mostrar un resumen sobre los estadísticos descriptivos de la variable continua que se encuentra dentro del paréntesis. Veamos el siguiente ejemplo que muestra estadísticos descriptivos sobre el nivel de ventas de los trabajadores independientes que se dedican a la producción/extracción tomando en cuenta a las condiciones: sobre el acceso a agua o no en el establecimiento, sobre la importancia de la actividad productiva o extractiva siendo la actividad principal o secundaria para el trabajador independiente y sobre la condición de registro del establecimiento (está registrado o no), siendo las variables ***e14t***, ***actividа***, ***e4a1*** y ***e1*** usadas para el ejemplo, según el cuestionario.

```
. bysort actividа:tabulate e1 e4a1, sum( e14t)

-> actividа = 1

                Means, Standard Deviations and Frequencies
of en el mes anterior, de lo que ud. produce/extrae □a cu□nto ascendieron sus venta

       □el
 negocio o
establecim
 iento que
ud. dirige      □su local o
        se     establecimiento
 encuentra    cuenta con: agua
registrado         potable?
     como:           si           no         Total
-----------------------------------------------------
 persona n    14120.667    2391.6667    8256.1667
              21024.804    1247.5643    15464.629
                      6            6           12
-----------------------------------------------------
 no est□ r       1914.5    547.65116    762.05882
              1573.1068     757.623    1039.5388
                      8           43           51
-----------------------------------------------------
     Total    7145.7143    773.44898    2189.5079
              14513.546    1018.5149    7218.1501
                     14           49           63

-> actividа = 2

                Means, Standard Deviations and Frequencies
of en el mes anterior, de lo que ud. produce/extrae □a cu□nto ascendieron sus venta

       □el
 negocio o
establecim
 iento que
ud. dirige      □su local o
        se     establecimiento
 encuentra    cuenta con: agua
registrado         potable?
     como:           si           no         Total
-----------------------------------------------------
 persona n            .          500           500
                      .            0             0
                      0            1             1
-----------------------------------------------------
 no est□ r    683.33333        249.1    349.30769
              419.32485    273.77462    348.97168
                      3           10           13
-----------------------------------------------------
     Total    683.33333    271.90909    360.07143
              419.32485    270.51819    337.69136
                      3           11           14
```

Figura 3.17. Nivel de ventas según la actividad, el registro del establecimiento y la condición de existencia de agua o no en el establecimiento.

En la figura 3.17. Se visualizan dos tablas, una tabla para cada valor de la variable ***actividа*** la cual toma los valores "1" cuando la actividad realizada es la principal y "2" cuando la actividad realizada es secundaria. A su vez, en cada celda de las tablas hay tres números que corresponden al promedio, desviación estándar y la frecuencia de arriba hacia abajo. En las filas de las tablas se encuentran las categorías correspondientes a la variable ***e1*** y en las columnas se presentan a las etiquetas de la variable ***e4a1*.**

Según los resultados de las tablas, los establecimientos que no cuentan con agua obtienen niveles de ingresos por ventas bajos, en comparación de los establecimientos que si tienen si la actividad es la principal. De hecho, la brecha entre ambos grupos según el acceso a agua en su establecimiento es muy evidente, siendo S/ 7145.70 en promedio que recibe cada trabajador cuando tiene acceso a agua contra S/773.50 en promedio que recibe cada trabajador sin acceso a agua, por lo que podemos inferir que el acceso de agua puede maximizar el nivel de ingresos de los trabajadores dedicados a actividades productivas o extractivas. También muestra que los trabajadores que son independientes como actividad principal, es decir si ***actividа*=1**, reciben en promedio S/ 8256.10 cuando el establecimiento está registrado como "Persona Natural" mientras que los establecimientos que aquellos que no están registrados reciben en promedio S/ 762.60. Cuando ***actividа*=2** podemos ver que no se registran trabajadores con ingresos cuando en el local hay servicio de agua y está registrado como "Persona Natural", por otro lado cuando el local no cuenta con servicio de agua y está registrado como "Persona Natural" solo se registra una observación que recibe S/ 500. De esta figura podemos ver la enorme brecha que existe cuando los trabajadores se dedican a empleos independientes como actividad principal y secundaria, se puede interpretar que en algunos casos la actividad laboral independiente permite obtener niveles de ingresos altos y en otros casos sus niveles de ingresos son bajos. Este resultado puede ser atribuido a la inherente heterogeneidad del sector informal.

La opción **sum** es una abreviatura de **summary**, esta opción es confundida en ocasiones con el comando **summarize** debido a que ambas instrucciones tienen usos similares. Usemos el comando **summarize** que sirve para generar un cuadro con estadísticos descriptivos sobre una variable continua, por ejemplo la variable ***e14t*** informa sobre el nivel de ingresos por ventas de los trabajadores independientes dedicados al rubro de producción/extracción. A continuación, se muestran algunos estadísticos descriptivos con el comando **summarize**.

```
. summarize e14t

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        e14t |         96    2023.177    6021.731        20      56000
```

Figura 3.18. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción (1).

En la figura 3.18. Se muestra que el comando **summarize** detalla estadísticos descriptivos, los cuales son el número de observaciones, promedio, desviación estándar, valor mínimo y valor máximo de izquierda a derecha. Si se desea un resumen más detallado conviene utilizar la opción **detail**.

```
. summarize e14t, detail

      en el mes anterior, de lo que ud. produce/extrae
                □a cu□nto ascendieron sus venta
-------------------------------------------------------------
      Percentiles      Smallest
 1%           20             20
 5%           50             30
10%           75             35       Obs                  96
25%          120             40       Sum of Wgt.          96

50%          500                      Mean           2023.177
                      Largest         Std. Dev.      6021.731
75%         1950          8580
90%         4500         10000       Variance       3.63e+07
95%         8200         12124       Skewness       7.721156
99%        56000         56000       Kurtosis       68.83269
```

Figura 3.19. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción (2).

Ahora podemos ver la pregunta completa que representa la variable **e14t** y otros estadísticos descriptivos como los percentiles al 1%, 5%, 10%, 25%, 50%, 75%, 90%, 95% y 99%, la varianza, la kurtosis y la asimetría estadística (Skewness).

Supongamos que ahora deseamos visualizar los descriptivos pero solamente a los trabajadores independientes dedicados al rubro producción/extracción en el distrito de Chiclayo, para lograrlo utilizaremos el componente condicional **if** y la variable **ubigeo**.

```
. summarize e14t if ubigeo=="140101", detail

      en el mes anterior, de lo que ud. produce/extrae
                □a cu□nto ascendieron sus venta
-------------------------------------------------------------
      Percentiles      Smallest
 1%           50            50
 5%           91            91
10%          120           100       Obs                  34
25%          300           120       Sum of Wgt.          34

50%          650                     Mean           2046.882
                        Largest       Std. Dev.      2793.474
75%         3200          4500
90%         4500          6000       Variance        7803497
95%        10000         10000       Skewness       2.166749
99%        12124         12124       Kurtosis       7.591547
```

Figura 3.20. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción en Chiclayo (1).

Ahora, si por algún motivo se desea solamente utilizar las primeras 100 observaciones, se utiliza el componente **in**.

```
. summarize e14t in 1/100, detail

      en el mes anterior, de lo que ud. produce/extrae
                □a cu□nto ascendieron sus venta
-------------------------------------------------------------
      Percentiles      Smallest
 1%           30            30
 5%           30           100
10%           30           120       Obs                   8
25%          110           350       Sum of Wgt.           8

50%          400                     Mean            2096.75
                        Largest       Std. Dev.       4117.052
75%         1800           450
90%        12124          1800       Variance        1.70e+07
95%        12124          1800       Skewness        2.131204
99%        12124         12124       Kurtosis        5.782371
```

Figura 3.21. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción en Chiclayo (2).

En la figura 3.21. la instrucción **in 1/100** ha tomado las observaciones desde la primera hasta la observación número 100, sin embargo STATA solamente ha tomado 8 observaciones debido a que en este intervalo solamente se han registrado 8 observaciones que se dedican al rubro producción/extracción.

STATA tiene otros comandos para la generación de tablas siendo el comando **tabstat** uno de los más importantes y ampliamente usados. El comando **tabstat**, permite cruzar información entre dos o más variables cuantitativas con una variable cualitativa cuando se incorpora la opción **by()**. A demás, con la opción **s()** ordena que se muestre en la tabla algunos estadísticos descriptivos, una instrucción parecida a la opción **sum** cómo se mostró en la figura 3.17.

La facilidad de este comando radica en que no es necesario ejecutar procedimientos previos ni acomodar las observaciones. En el siguiente ejemplo veremos cómo los ingresos de los trabajadores dedicados en los rubros producción/extracción y comercial se distribuyen con respecto a la condición de estar registrados o no. Las variables ***e14t*** y ***e17t*** representan a los ingresos de los trabajadores independientes dedicados a los rubros señalados respectivamente, mientras la variable ***e1*** informa sobre la condición de registro del establecimiento.

```
. tabstat e14t e17t, by(e1)

Summary statistics: mean
  by categories of: e1 (El negocio o establecimiento que ud. dirige se encuentra registrado como:)

                e1 |      e14t      e17t
-------------------+--------------------
persona natural    |  7415.875  5763.459
no est registra    |  944.6375  2593.725
-------------------+--------------------
             Total |  2023.177  3213.449
-------------------------------------------
```

Figura 3.22. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción y comercial.

Por encima de la tabla generada con el comando **tabsat**, se puede apreciar la expresión “Summary statistics”, esta expresión es muy útil ya que está informando sobre los estadísticos descriptivos que se ven en cada celda de la tabla. En las columnas de la tabla se observan las variables ***e14t*** y ***e17t*** y en las filas de la tabla se aprecian las etiquetas de la variable ***e1*** por efecto de la opción **by**, por tanto, es imprescindible que se incluya esta opción cuando se requiera ordenar a la tabla en torno a una variable cualitativa, en este ejemplo la variable ***e1*** cumple ese rol. Podemos interpretar que, cada trabajador independiente recibe en promedio S/ 7415.90 en el rubro productivo/extractivo, mientras que los trabajadores independientes en el rubro comercial reciben ingresos en promedio S/ 5763.50, ambos resultados cuando sus establecimientos están registrados como “Persona Natural”. Por defecto, el promedio se muestra en el ejemplo, si se requiere ver otros estadísticos se debe utilizar la opción **statistics** o su abreviatura **s**.

```
. tabstat e14t e17t, by(e1) s(n sum mean sd)

Summary statistics: N, sum, mean, sd
  by categories of: e1 (□el negocio o establecimiento que ud. dirige se encuentra registrado como:)

             e1 |      e14t      e17t
----------------+--------------------
persona natural |        16        61
                |    118654    351571
                |  7415.875  5763.459
                |  13448.34  8982.571
----------------+--------------------
no est□ registra|        80       251
                |     75571    651025
                |  944.6375  2593.725
                |  1482.351  9887.141
----------------+--------------------
          Total |        96       312
                |    194225   1002596
                |  2023.177  3213.449
                |  6021.731  9784.353
----------------------------------------
```

Figura 3.23. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción y comercial distribuido según la condición de registro del establecimiento.

Ahora podemos ver en cada celda de la tabla 4 resultados, los cuales corresponden a cada estadístico que se ha especificado en la instrucción siguiendo el orden establecido en la opción **s()**, siendo precisamente los más utilizados: frecuencia, suma, promedio y la desviación estándar, representados con **n, sum, mean, sd**, respectivamente. Esta no es la única ventaja del comando **tabstat** frente al comando **table**, otra ventaja es la inclusión de más variables numéricas como se ve en el siguiente ejemplo.

```
. tabstat e14t e16t e17t e19t e20t e22t, by(e1) s(n sum mean sd)

Summary statistics: N, sum, mean, sd
  by categories of: e1 (□el negocio o establecimiento que ud. dirige se encuentra registrado como:)

             e1 |      e14t      e16t      e17t      e19t      e20t      e22t
----------------+------------------------------------------------------------
persona natural |        16        14        61        60        53        29
                |    118654     60748    351571    252938    153856     36542
                |  7415.875  4339.143  5763.459  4215.633  2902.943  1260.069
                |  13448.34  10527.06  8982.571  7625.772  2621.802   1848.91
----------------+------------------------------------------------------------
no est□ registra|        80        58       251       247       407       280
                |     75571     18457    651025    507500    502022    190487
                |  944.6375  318.2241  2593.725  2054.656  1233.469  680.3107
                |  1482.351  582.6211  9887.141  9635.105  1345.887  892.1567
----------------+------------------------------------------------------------
          Total |        96        72       312       307       460       309
                |    194225     79205   1002596    760438    655878    227029
                |  2023.177  1100.069  3213.449  2476.997  1425.822  734.7217
-------------------------------------------------------------------------------
```

Figura 3.24. Nivel de ventas de los trabajadores independientes en el rubro producción/extracción, comercial y servicios.

Las variables ***e14t***, ***e16t***, ***e17t*** representan los ingresos de los trabajadores independientes dedicados a los rubros producción/extracción, comercio y servicios respectivamente, y las variables ***e19t***, ***e20t*** y ***e21t*** representan a los gastos de los negocios de los trabajadores independientes dedicados a los rubros producción/extracción, comercio y servicios.

Para terminar con los comandos sobre la generación de tablas se presenta al comando **table.** Este comando permite mostrar una tabla que cruza información de variables cualitativas, por ejemplo, en la siguiente tabla se visualiza las características sobre la condición de acceso a servicio básico del agua siguiendo una distribución acorde a la condición de estar registrado o no. Las variables que representan las características mencionadas son ***e1*** y ***e4a1***, respectivamente.

```
. table e1 e4a1

                                                  |su local o
                                                  |establecimi
                                                  |ento cuenta
                                                  | con: agua
el negocio o establecimiento que ud.              |  potable?
dirige se encuentra registrado como:              |   si    no
--------------------------------------------------+----------
persona natural (con ruc, rus, rer, u ot          |   26    70
       no est registrado (no tiene ruc)           |   55   193
```

Figura 3.25. El servicio de agua distribuido según la condición de registro del establecimiento.

También se puede cruzar información de dos variables cualitativas y ordenarla según las categorías de otra variable cualitativa, por ejemplo, en la siguiente figura se observan cuántos trabajadores se encuentran en cada distrito de la provincia de Chiclayo y se reparten acorde a su condición de estar registrados y también a la condición de pertenencia del local donde desarrollan sus actividades laborales. Se puede ver que se cruzan ambas condiciones, representados en las variables ***e1*** y ***e3***, en cada distrito en la figura 3.26.

```
. table ubigeo e1 e3
```

código de ubicación geográfica	¿Ud. realiza su negocio o actividad en un local: and ¿El negocio o establecimiento que ud. dirige se encuentra registrado como: ¿propio? (propietario) persona natural (con	no est¿ registrado (	¿alquilado? persona natural (con	no est¿ registrado (
140101	25	39	17	10
140102		5		1
140103	1	8		
140104	1			
140105	10	31	8	10
140106	5	14	1	5
140107		1		
140108	1	9		5
140109		2		
140110				1
140111	1	1		
140112	2	12	3	5
140113		1		1
140114		5		
140115	2	2	1	2
140116	3	5		2
140117		4		1
140118	3	7	1	
140119		8		
140120		8		

código de ubicación geográfica	¿Ud. realiza su negocio o actividad en un local: and ¿El negocio o establecimiento que ud. dirige se encuentra registrado como: ¿prestado? persona natural (con	no est¿ registrado (	¿otro? persona natural (con	no est¿ registrado (
140101	1	10		
140102		3		1
140103				
140104				2
140105	4	5		1
140106	2	1		
140107				
140108		6		
140109				
140110				
140111				
140112	1	3		
140113		3		
140114		1		
140115				3
140116				
140117				
140118	1	1		1
140119		1		
140120	2	1		

Figura 3.26. Número de trabajadores en cada distrito de la provincia de Chiclayo distribuido según la condición de pertenencia del local y el registro de su establecimiento.

En la figura 3.26. Notamos que con el comando **table** se pueden utilizar más de dos variables, también notamos que la segunda variable numérica divide a la primera variable numérica, en el ejemplo serían las variables ***e3*** y ***e1*** respectivamente. Podríamos interpretar que en el distrito de Chiclayo hay 25 trabajadores que tienen local propio y además están registrados como “Persona Natural”, mientras que hay 39 trabajadores cuyo local es propio, pero no se encuentra registrado.

Con estas tablas podemos empezar a analizar la variable dependiente, la cual es el nivel de ingreso de los trabajadores independientes en el distrito de Chiclayo. Empecemos seleccionando a las observaciones que se encuentran en este distrito. Para ello usaremos el comando **keep**.

```
. keep if ubigeo=="140101"
(604 observations deleted)
```

Figura 3.27. Trabajadores independientes de Chiclayo.

El siguiente paso es crear la variable ***rubro***, la cual contendrá información sobre el tipo de actividad que realizan los trabajadores independientes de Chiclayo. Acorde al INEI la variable ***rubro*** tendrá 3 valores que representarán a cada categoría, siendo "1" si el trabajador independiente se dedica a la actividad productiva/extractiva, "2" si el trabajador independiente se dedica a la actividad comercial y "3" el trabajador independiente se dedica a la actividad prestadora de servicios. Las categorías mencionadas se encuentran contenidas en la información de tres variables ***e13a***, ***e13b*** y ***e13c***, respectivamente. Para la creación de tal variable con sus correspondientes categorías se usarán los comandos **gen** y **replace**. El primero permite generar una nueva variable según la instrucción que se le ordene, y el segundo reemplaza los valores de una variable según la instrucción determinada. En ambos comandos se usará la condicional **if** para indicar que use los valores que se encuentran en las variables ***e13a***, ***e13b*** y ***e13c***. Posteriormente, crearemos una tabla que muestre información sobre la variable ***rubro***.

```
. gen rubro=1 if e13a==1
(209 missing values generated)

. replace rubro=2 if e13b==2
(87 real changes made)

. replace rubro=3 if e13c==3
(129 real changes made)
```

Figura 3.28. Comando gen y replace.

```
. tab rubro

      rubro |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         31       12.76       12.76
          2 |         83       34.16       46.91
          3 |        129       53.09      100.00
------------+-----------------------------------
      Total |        243      100.00
```

Figura 3.29. Distribución de los trabajadores independientes según el rubro en el que se dedican (1).

La variable ***rubro*** tiene tres valores, el valor "1" representa el sector producción/extracción, el valor "2" representa el sector comercial y el valor "3" el sector servicios. En STATA es posible otorgar las etiquetas a los valores de una variable, para este fin se utiliza el comando **label define** y posteriormente **label values**.

```
. label define rubro 1 "prod/extrac" 2 "comercial" 3 "servicios"

. label values rubro rubro

. tab rubro
```

rubro	Freq.	Percent	Cum.
prod/extrac	31	12.76	12.76
comercial	83	34.16	46.91
servicios	129	53.09	100.00
Total	243	100.00	

Figura 3.30. Distribución de los trabajadores independientes según el rubro en el que se dedican (2).

El comando **label define** otorga etiquetas a una lista de valores y almacena las etiquetas otorgadas bajo un nombre. En la figura anterior el comando está definiendo las etiquetas "prod/extrac", "comercial" y "servicios" a los valores "1", "2" y 3 respectivamente, y guarda estas etiquetas bajo una lista con nombre **rubro**. Posteriormente al comando **label define**, se usa el comando **label values** para utilizar la lista creada con nombre **rubro** para otorgarle una etiqueta a cada valor de la variable ***rubro***. Ahora, si deseamos darle una etiqueta a una variable entonces el comando **label variable** es la solución para este requerimiento. Veamos la siguiente figura.

```
. label variable rubro "Rubro en el que se dedica el negocio"
```

Nombre	Etiqueta
e23t	otros gastos en el ...
e24t	caracter◆sticas d...
e25st1	hoja de control: s...
e25st2	hoja de control: s...
e25st3	hoja de control: s...
e25st4	hoja de control: s...
e25t1	gastos en mano d...
e25t2	total gasto mensu...
e25t3	total ganancia neta
ticuest04	origen de cuestio...
factora07	factor de expansi...
rubro	Rubro en el que se...

Figura 3.31. Etiqueta de la variable rubro.

En la figura 3.30. Se observan los sectores a los que se dedican los trabajadores independientes en el año 2018 y podemos ver que más de la mitad se dedica a actividades prestadoras de servicios, mientras que el 34.79% se dedica a las actividades comerciales y solamente el 12.76% a las actividades productivas/extractivas. En la siguiente figura se detallan cómo se distribuyen los establecimientos según su condición de estar registradas o no.

```
. tab e1

 □el negocio o establecimiento que ud. |
  dirige se encuentra registrado como: |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
persona natural (con ruc, rus, rer, u o |         55       22.63       22.63
     no est□ registrado (no tiene ruc)  |        188       77.37      100.00
---------------------------------------+-----------------------------------
                                 Total |        243      100.00
```

Figura 3.32. Condición de estar registrados de los establecimientos.

Casi el 80% de los establecimientos no están registrados, veamos cuales son los motivos.

```
. tab e1a1

 □cu□l es la raz□n principal por la que |
                  no se ha registrado? |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
           no sabe si debe registrarse |          1        0.53        0.53
no podr□a asumir la carga de impuestos |          2        1.06        1.60
             le quita demasiado tiempo |          1        0.53        2.13
su negocio es peque□o/produce poca cant |         49       26.06       28.19
                es un trabajo eventual |         14        7.45       35.64
             no lo considera necesario |        121       64.36      100.00
---------------------------------------+-----------------------------------
                                 Total |        188      100.00
```

Figura 3.33. Motivos por el cual no se registra el establecimiento.

Según la figura 3.33. El 64.36% considera que no es necesario estar registrado para establecerse en su negocio, este motivo mayoritario puede ser un indicador de la carente educación financiera que pueden tener algunos trabajadores en el distrito de Chiclayo, por otro lado el 26.06% indica que su negocio produce pocos ingresos para estar registrados, por lo tanto podemos concluir que la poca educación tanto en aspectos financieros como en aspectos de cómo llevar una empresa pueden determinar si un negocio está registrado o no. Para estar más seguros de esta conclusión, podríamos utilizar la información de la variable *e1b*, la cual registra qué tipos de libros son usados en los negocios para llevar las cuentas contables.

```
. tab e1b

 □ud. lleva las cuentas de su negocio o
                              actividad: |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
por medio de libros de ingresos y gasto |          6        2.47        2.47
por medio de apuntes, registros o anota |         61       25.10       27.57
                      no lleva cuentas  |        176       72.43      100.00
-----------------------------------------+-----------------------------------
                                 Total   |        243      100.00
```

Figura 3.34. Libros usados por los independientes para llevar cuentas.

Según la figura 3.34. Solamente el 2.47% de la muestra registra sus cuentas contables mediante libros de ingresos y gastos exigidos por SUNAT, mientras que el 25.10% registra sus cuentas en apuntes personales y el 72.43% no lleva cuentas.

Estas dos últimas figuras muestran indicios de la existencia de una carente educación financiera por parte de los trabajadores independientes, algunos autores consideran que esta puede ser la causa de la alta informalidad en la que se encuentran estos trabajadores en el distrito de Chiclayo. Veamos en la siguiente figura cuáles han sido las motivaciones de los trabajadores independientes para iniciar un negocio.

```
. tab e5

   □cu□l es el motivo por el cual
  inici□ este negocio o actividad? |      Freq.     Percent        Cum.
-----------------------------------+-----------------------------------
    no encontr□ trabajo asalariado |          5        2.06        2.06
obtiene ingresos / mayores ingresos |         28       11.52       13.58
          quiere ser independiente |         64       26.34       39.92
            por tradici□n familiar |         12        4.94       44.86
           por necesidad econ□mica |        118       48.56       93.42
                             otro  |         16        6.58      100.00
-----------------------------------+-----------------------------------
                             Total |        243      100.00
```

Figura 3.35. Motivaciones para iniciar actividades laborales independientes.

La motivación más frecuente entre los trabajadores independientes es la necesidad económica, ya que el 48.56% ha declarado ser esta la motivación por la cual son trabajadores independientes.

3.7.1.5. Estimación de los coeficientes de regresión.

- **Actividades productivas/extractivas.**

En esta sección se explicará cómo realizar una regresión múltiple mediante MCO para obtener los estimadores cuando el modelo econométrico (3.7.1.) utiliza datos

pertenecientes a trabajadores que han realizado actividades relacionadas al sector productivo/extractivo.

Empezamos seleccionando los datos, para ello comenzamos ejecutando el comando **preserve.** Este comando permite guardar la base de datos en la memoria de STATA y esta es su importancia, ya que se puede manipular los datos y posteriormente recuperar la base de datos original con el comando **restore**. Después de haber ejecutado el comando **preserve**, se utilizará el comando **keep** con la condicional **if** y la variable ***e13a*** para ordenar a STATA que solo mantenga en la base de datos a las observaciones si la variable ***e13a*** es igual a "1". Se utilizará a esta variable, debido a que registra dos posibles valores, "0" si el trabajador no se ha dedicado a actividades productivas/extractivas y "1" si los trabajadores se han dedicado a actividades productivas/extractivas. En consecuencia a la heterogeneidad de los trabajadores independientes, es posible que en algunas observaciones se hayan registrado a trabajadores independientes dedicándose a otras actividades.

```
. preserve

. keep if e13a==1
(209 observations deleted)
```

Figura 3.36. Comandos preserve y keep.

Posteriormente a la introducción de ambos comandos, construiremos la variable ***gastos*** con el comando **gen**. Esta variable se compone con tres variables ***e16t***, ***e25t1*** y ***e25t2*** que representan gastos del trabajador en su negocio, gastos en mano de obra y gastos provenientes del capítulo 50 respectivamente. Sin embargo, si abrimos la base de datos observamos que en la variable ***e16t*** existen ".", lo cual significan datos faltantes o valores vacíos. Por ello, solucionaremos el problema con el comando **replace** reemplazando los datos vacíos con 0. Este procedimiento se ejecutara porque no es posible sumar valores vacíos con valores numéricos.

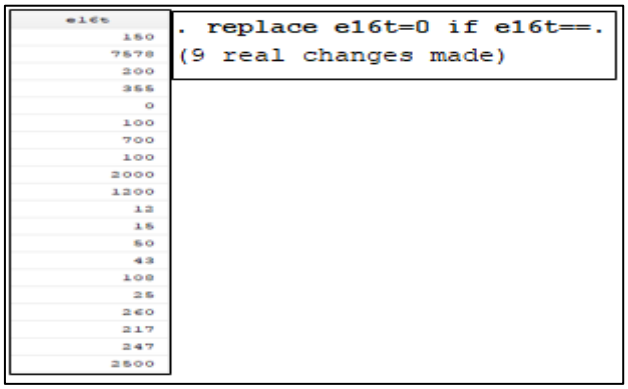


Figura 3.37. Comando replace.

```
. gen gastos=e16t+ e25t1+ e25t2
```

Figura 3.38. Creación de la variable gastos.

En la figura 3.38. Hemos construido la variable ***gastos*** como la suma de las variables ***e16t***, ***e25t1*** y ***e25t2*** usando el comando **gen**. En la sección sobre la especificación del modelo econométrico se ha detallado que solo se tomarán a los trabajadores independientes que perciban ganancias, ingresos, gastos y tengan trabajadores en su establecimiento. No obstante, al abrir la base de datos y examinar la variable ***gastos*** podemos notar que existen datos faltantes, por ello se debe descartar esos datos inexistentes con el comando **drop** y su condicional **if**. Sin embargo, cabe aclarar que este procedimiento no es recomendable para realizar regresiones con datos faltantes, ya que borra otras observaciones de otras variables. Este procedimiento se llevará a cabo en este ejemplo sólo para fines didácticos y con motivo de no extender la explicación.

```
. drop if gastos==0
(9 observations deleted)
```

Figura 3.39. Comando drop.

Una vez que ya se tienen todas las variables necesarias para realizar la regresión, se ejecuta el comando **regress** para estimar los estimadores del modelo especificado. STATA presenta la siguiente sintaxis del comando **regress**.

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

Figura 3.40. Sintaxis del comando reg.

En los manuales de STATA podemos ver que muchos comandos tienen una línea por debajo de algunas letras como se puede ver en la figura 3.40. Estas líneas indican las abreviaturas que algunos comandos pueden tener, en el caso del comando **regress** su abreviatura es **reg** según la figura 3.40.

Al lado del comando **reg** está el componente ***depvar***, el cual sirve para indicarle a STATA cuál es la variable dependiente, al lado derecho del componente ***depvars*** se encuentra el componente ***indepvars*** que indica a STATA cuáles son las variables explicativas, posteriormente están los componentes ***if*** y ***in*** que ya han sido explicados anteriormente y sirven como condicionales. El componente ***weight*** tiene la función de indicar a STATA que realice la regresión tomando en cuenta los pesos o ponderaciones de las variables, este componente es muy útil para aplicar MCGP como método correctivo

para tratar la heterocedasticidad cuando el esquema de la varianza del término de error es conocido. Finalmente, los componentes que se encuentran al lado derecho de la coma son las opciones que se le puede agregar al comando **reg**, entre las más conocidas son:

- ***vce*(*vcetype*),** Indica al programa STATA cuál es el procedimiento que se desea para hallar la varianza de los estimadores, basta con poner entre los paréntesis el tipo de método para ejecutar la instrucción. Por ejemplo, ***vce*(*ols*)** calcula la varianza siguiendo el método de MCO, no es necesario expresarlo en la sintaxis porque está por defecto en el comando **reg**, mientras que ***vce*(*robust*)** o simplemente ***robust*** indica a STATA que se requiere el método de los errores robustos para halla la varianza de los estimadores, ***robust*** corrige la heterocedasticidad mediante el método correctivo de White.
- ***level*(#)**, este comando es usado para la inferencia de los estimadores, por defecto el nivel de confianza que usa STATA en las regresiones es del 95%, sin embargo con la opción ***level*(#)** se puede escoger el nivel de confianza. Por ejemplo, si se desea utilizar un nivel de confianza del 90%, basta con ***level*(90)** para indicar a STATA que realiza la regresión al 90% del nivel de confianza.
- ***noconstant***, esta opción indica a STATA que no calcule el intercepto.

Para ejecutar la regresión en STATA en el ejemplo, introducimos la siguiente instrucción.

```
. reg e25t3 e14t gastos e8a

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(3, 23)        =     19.15
       Model |  23465252.3     3  7821750.76   Prob > F        =    0.0000
    Residual |  9394003.34    23  408434.928   R-squared       =    0.7141
-------------+------------------------------   Adj R-squared   =    0.6768
       Total |  32859255.6    26  1263817.52   Root MSE        =    639.09

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e14t |   .5242571   .1713889     3.06   0.006     .1697121    .8788021
    gastos |  -.5867962   .1818076    -3.23   0.004    -.9628938   -.2106986
       e8a |   387.0503    175.662     2.20   0.038      23.6657    750.4349
     _cons |  -91.22355   228.4641    -0.40   0.693    -563.8376    381.3905
------------------------------------------------------------------------------
```

Figura 3.41. Regresión para los trabajadores independientes que se han dedicado a actividades productivas/extractivas.

En la figura 3.41. Podemos ver 2 cuadros, uno en el lado superior y el otro en el lado inferior, y al lado derecho del cuadro superior se encuentra una lista de detalles con

algunos estadísticos propios de la regresión. De arriba hacia abajo, estos son los elementos de esa lista:

- Número de observaciones.
- Estadístico *F* calculado con sus grados de libertad entre paréntesis para determinar la relevancia global.
- Probabilidad del estadístico *F* calculado.
- El coeficiente de determinación.
- El coeficiente de determinación ajustado.
- Error estándar de regresión.

El cuadro inferior muestra la información con respecto a los estimadores de la regresión, de izquierda a derecha las columnas son:

- Estimadores.
- Error estándar de los estimadores.
- Estadístico *t* calculado para determinar la relevancia individual.
- Probabilidad del estadístico *t* calculado.
- Las dos últimas columnas muestran los intervalos de confianza de los estimadores, el primero es el intervalo inferior mientras el segundo es el intervalo superior al 95%.

Source	SS	df	MS
Model	23465252.3	3	7821750.76
Residual	9394003.34	23	408434.928
Total	32859255.6	26	1263817.52

Figura 3.42. Tabla ANOVA del comando reg.

Por último, El cuadro superior es el cuadro ANOVA de la regresión, donde cada columna de la tabla indica los siguientes valores de la suma cuadrática, grados de libertad y la media cuadrática. Por otro lado, las filas $Model$ y $Residual$ muestran lo relacionado a la información proveniente del modelo especificado y la información sobre la parte que no se puede explicar.

Con la figura 3.41. Obtenemos el siguiente modelo.

$$\hat{G}_i = -91.22 + 0.52I_i - 0.58C_i + 387.05M_i + \hat{\mu}_i \quad (3.7.2.)$$

$$ee = (228.46) \quad (0.17) \quad (0.18) \quad (175.66)$$

$$t = -0.40 \quad 3.06 \quad -3.23 \quad 2.20$$

Del modelo (3.7.2.) podemos interpretar los siguientes estimadores:

Manteniendo el supuesto de ceteris paribus, se interpretan los estimadores.

Ingresos. Si los ingresos de los trabajadores independientes aumentan en una unidad monetaria, sus ganancias totales netas aumentan en 0.52 unidades monetarias.

Gastos. Si los gastos de los trabajadores independientes aumentan en una unidad monetaria, sus ganancias totales netas disminuyen en 0.58 unidades monetarias.

Número de trabajadores. Si el número de trabajadores aumenta en una unidad, las ganancias totales netas de los trabajadores independientes aumentan 387.05 unidades monetarias.

Ahora veamos si los estimadores tienen relevancia individual con las siguientes pruebas de hipótesis:

- **Ingresos**

$H_0: \beta_2 = 0$ (3.7.3.)

$$H_1: \beta_2 \neq 0$$

El estadístico t calculado según la figura 3.41. Es: $tc = 3.06$. Para hallar el estadístico t tabulado utilizamos el comando **scalar**.

```
. scalar tt=invt(23,1-0.05/2)
```

Figura 3.43. Comando scalar.

El comando **scalar** en la figura 3.43. Calcula el estadístico *t* tabulado usando los grados de libertad, $n - k = 27 - 4 = 23$, y usando un nivel de significancia del 0.05 dividido entre 2 debido a que este estadístico usa dos colas según la hipótesis alternativa en (3.7.3.)

Para visualizar el resultado de 3.44. Ejecutamos el comando **display** o su abreviatura **disp.** Seguido del nombre de la scalar, en este caso hemos colocado el nombre **tt**.

```
. disp tt
2.0686576
```

Figura 3.44. Comando disp.

Por lo tanto podemos inferir lo siguiente:

$|tc| > tt_{23,\frac{0.05}{2}}$ (3.7.4.)

Entonces con (3.7.4.) rechazamos la hipótesis nula y podemos inferir que el estimador de ***e14t*** es diferente de 0, por este motivo la variable ***e14t*** tiene relevancia individual y no debe ser descartado del modelo.

- **Gastos.**

$H_0: \beta_3 = 0$ (3.7.5.)

$$H_1: \beta_3 \neq 0$$

El estadístico *t* calculado es -3.23 y el estadístico *t* tabulado es 2.07 según la figura 3.44. Por lo tanto se infiere:

$|tc| > tt_{23,\frac{0.05}{2}}$ (3.7.6.)

Al ser mayor el estadístico *t* calculado mayor al estadístico *t* tabulado, rechazamos la hipótesis nula y concluimos que el estimador de ***gastos*** es diferente de 0 y la variable ***gastos*** tiene relevancia individual en el modelo.

- **Número de trabajadores.**

$H_0: \beta_4 = 0$ (3.7.7.)

$$H_1: \beta_4 \neq 0$$

El estadístico *t* calculado es 2.20 y el estadístico *t* tabulado es 2.07 según la figura 3.44. Por lo tanto se infiere:

$|tc| > tt_{23,\frac{0.05}{2}}$ (3.7.8.)

En consecuencia a (3.7.8.), rechazamos la hipótesis nula y concluimos que el estimador de ***e8a*** es diferente de 0 y la variable ***e8a*** tiene relevancia individual en el modelo.

Los estimadores de las variables explicativas cumplen los signos esperados y además tienen relevancia individual con una significancia del 5% como se ha visto en sus pruebas de hipótesis. Otra forma de comprobarlo es observando el valor-p de sus respectivos estadísticos t calculados, $(P > |t|)$, debido a que en los tres estimadores, los valores-p son menores a la significancia del 5% y además el 0 no se encuentra en sus intervalos de confianza, entonces concluimos que los estimadores tienen significancia individual. Esta regla decisión se aplica no solo a las pruebas de hipótesis sobre relevancia individual, sino también a todas las pruebas de hipótesis que se plateen.

Para verificar si el modelo tiene relevancia global se establece el siguiente test de hipótesis.

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \text{ (3.7.9.)}$$

$$H_1: Ningún\ \beta_k\ es\ igual\ a\ 0$$

El estadístico *F* calculado es 19.15, y para hallar el estadístico *F* tabulado utilizamos una vez más el comando **scalar** y posteriormente el comando **disp.**

```
. scalar ft=invF(3,23,1-0.05)

. disp ft
3.0279984
```

Figura 3.45. Calculando el estadístico *F* tabulado.

En el comando **scalar** debemos colocar los grados de libertad siendo: $(k - 1 = 4 - 1 = \mathbf{3}, n - k = 27 - 4 = \mathbf{23})$ y un nivel de significancia del 5%. A diferencia de la anterior prueba de hipótesis no se debe dividir entre 2 debido a que la prueba *F* para la relevancia global utiliza una cola.

En consecuencia, el estadístico *F* tabulado es 3.03, por lo que podemos concluir:

$$Fc > Ft^{3}_{23,0.05} \text{ (3.7.10.)}$$

En vista a que el estadístico *F* calculado es mayor al estadístico *F* tabulado, podemos rechazar la hipótesis nula y concluir que el modelo presenta significancia global y es útil para explicar a la variable endógena. Similar a la prueba de hipótesis sobre relevancia individual, la significancia global también se puede determinar observando el valor-p del estadístico *F* calculado; podemos llegar a la misma conclusión: el modelo tiene relevancia global ya que el valor-p del estadístico *F* calculado es menor al 5% de la

significancia. Por último, el coeficiente de determinación es 71.41%, con lo cual el modelo explica el 71.41% de la variabilidad de la variable dependiente. Con estas interpretaciones, podemos concluir que el modelo está correctamente especificado y estimado ya que los estimadores tienen significancia individual y cumplen con el signo esperado, el modelo tiene significancia global y el coeficiente de determinación, pese a no ser tan elevado, muestra una buena bondad de ajuste. No obstante, estos resultados no son determinantes para estar completamente seguros si el modelo cumple con los supuestos de MCO. En la próxima sección veremos cómo comprobar si cumplen los supuestos de MCO.

- **Actividades comerciales.**

Debido a que la base de datos ha sido manipulada por la eliminación de algunas observaciones debemos ejecutar el comando **restore** cuya función es restaurar la base de datos original. Cabe recalcar, que el comando **restore** solo funciona si previamente a las modificaciones que hemos realizado a la base de datos, hemos ejecutado el comando **preserve**.

```
. restore
```

Figura 3.46. Comando restore (1).

Después del comando **restore**, volvemos a ejecutar el comando **preserve** para volver a guardar la base de datos original en la memoria de STATA. Es posible combinar ambos comandos con el fin de agilizar el uso en STATA; si utilizamos al comando **preserve** como opción del comando **restore**, entonces no será necesario ordenar el comando **preserve** después del comando **restore**.

```
. restore,preserve
```

Figura 3.47. Comando restore (2).

Continuamos realizando el mismo procedimiento previo a ejecutar la regresión, del mismo modo como se ha realizado en el punto anterior. Utilizaremos la variable ***e13b*** para seleccionar a los trabajadores que han realizado actividades comerciales y también a la variables ***e19t***, ***e25t1*** y ***e25t2*** para la construcción de la variable ***gastosc***, la cual muestra información sobre los gastos de los trabajadores independientes dedicados a actividades comerciales.

```
. keep if e13b==2
(156 observations deleted)

. replace e19t=0 if e19t==.
(1 real change made)

. gen gastosc=e19t+ e25t1+ e25t2

. drop if gastosc==0
(1 observation deleted)
```

Figura 3.48. Manteniendo las observaciones sobre los trabajadores dedicados a actividades comerciales y construyendo la variable *gastosc*.

La variable ***gastosc*** representa los gastos de los trabajadores independientes que han realizado actividades comerciales, y se compone con los gastos sobre el establecimiento de los negocios, la mano de obra y los gastos mensuales según el capítulo 50. Se ha descartado una observación, ya que un trabajador ha reportado no haber percibido gastos, no obstante, es necesario volver a mencionar que este procedimiento solo se ha efectuado con fines didácticos y para no hacer engorrosa la explicación. Continúa realizar la regresión con el comando **reg**.

```
. reg e25t3 e17t gastosc e8a

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(3, 82)        =     67.64
       Model |  58918846.4     3  19639615.5   Prob > F        =    0.0000
    Residual |  23809986.8    82  290365.692   R-squared       =    0.7122
-------------+------------------------------   Adj R-squared   =    0.7017
       Total |  82728833.2    85  973280.391   Root MSE        =    538.86

------------------------------------------------------------------------------
       e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e17t |   .5539907   .0898667     6.16   0.000     .3752171    .7327642
     gastosc |   -.508537   .1025509    -4.96   0.000    -.7125434   -.3045306
        e8a |   779.6868   76.10672     10.24   0.000     628.2863    931.0873
       _cons |    -635.51   122.5403    -5.19   0.000    -879.2817   -391.7383
------------------------------------------------------------------------------
```

Figura 3.49. Regresión para los trabajadores independientes que se han dedicado a actividades comerciales.

Con la figura 3.49. Obtenemos el siguiente modelo estimado.

$$\hat{G}_i = -635.51 + 0.55I_i - 0.51C_i + 779.69M_i + \hat{\mu}_i \text{ (3.7.11.)}$$

$$ee = (122.54) \quad (0.89) \quad (0.10) \quad (76.11)$$

$$t = -5.19 \qquad 6.16 \qquad -4.96 \qquad 10.24$$

Manteniendo el supuesto de ceteris paribus, el modelo (3.7.11.) indica los siguientes resultados.

Ingresos. Si los ingresos de los trabajadores independientes que se han dedicado a actividades comerciales aumentan en una unidad monetaria, sus ganancias netas totales aumentan en 0.55 unidades monetarias.

Gastos. Si los gastos de los trabajadores independientes que se han dedicado a actividades comerciales aumentan en una unidad monetaria, sus ganancias netas totales disminuyen en 0.51 unidades monetarias.

Número de trabajadores. Si el número de trabajadores a cargo de los trabajadores independientes aumentan en una unidad, las ganancias totales netas aumentan en 779.69 unidades monetarias.

Veamos si los estimadores tienen la relevancia individual.

- **Ingresos.**

$H_0: \beta_2 = 0$ (3.7.12.)

$$H_1: \beta_2 \neq 0$$

El estadístico *t* calculado según la figura 3.49. Es: $tc = 6.16$. Para hallar el estadístico *t* tabulado utilizamos el comando **scalar** y el comando **disp**.

```
. scalar tt=invt(82,1-0.05)

. disp tt
1.6636492
```

Figura 3.50. Comando scalar y disp.

Por lo tanto podemos inferir lo siguiente:

$|tc| > tt_{82,\frac{0.05}{2}}$ (3.7.13.)

Con (3.7.13.) rechazamos la hipótesis nula y podemos inferir que el estimador de ***e17t*** es diferente de 0 y la variable ***e17t*** tiene relevancia individual en el modelo (3.7.11.).

- **Gastos.**

$H_0: \beta_3 = 0$ (3.7.14.)

$$H_1: \beta_3 \neq 0$$

El estadístico *t* calculado es -4.96 y el estadístico *t* tabulado es 1.66 según la figura 3.50. Entonces inferimos:

$|tc| > tt_{82,\frac{0.05}{2}}$ (3.7.15.)

Por tal motivo, rechazamos la hipótesis nula y concluimos que el estimador de ***gastosc*** es diferente de 0 y por ello la variable ***gastosc*** tiene relevancia individual en el modelo (3.7.11.).

- **Número de trabajadores.**

$H_0: \beta_4 = 0$ (3.7.16.)

$$H_1: \beta_4 \neq 0$$

El estadístico *t* calculado es 10.24 y el estadístico *t* tabulado es 1.66 según la figura 3.50. Por consiguiente:

$|tc| > tt_{82,\frac{0.05}{2}}$ (3.7.17.)

Se rechaza la hipótesis nula y concluimos que el estimador de ***e8a*** es diferente de 0 y la variable ***e8a*** tiene relevancia individual en el modelo.

De igual forma que en la anterior regresión, los valores-p de los estimadores son menores a una significancia del 5% entonces podemos rechazar sus respectivas hipótesis nulas y concluir que tienen relevancia individual.

En cuanto a su relevancia global se plantea el siguiente test de hipótesis.

$H_0: \beta_2 = \beta_3 = \beta_4 = 0$ (3.7.18.)

$$H_1: Ningún\ \beta_k\ es\ igual\ a\ 0$$

El estadístico *F* calculado es 67.64 y para hallar el estadístico *F* tabulado utilizamos el comando **scalar** y posteriormente el comando **disp.**

```
. scalar ft2=invF(3,82,1-0.05)

. disp ft2
2.7159366
```

Figura 3.51. Calculando el estadístico *F* tabulado.

En consecuencia a que el estadístico *F* tabulado es 2.71. Y hemos obtenido:

$Fc > Ft^{3}_{82,0.05}$ (3.7.19.)

Según (3.7.20.) podemos rechazar la hipótesis nula y concluir que el modelo tiene significancia estadística global. Por tal motivo, el modelo sirve para explicar a la variable endógena. Si revisamos el valor-p del estadístico *F* calculado con respecto al nivel de significancia, siendo.

$Prob > F = 0.0000$ (3.7.20.)

Llegamos a la misma conclusión, ya que al ser el primero menor a una significancia del 5% podemos rechazar la hipótesis nula en (3.7.18.) y asumir que el modelo tiene relevancia global. Finalmente, el coeficiente de determinación del modelo (3.7.11.) es 71.22%, y se interpreta que el modelo explica el 71.22% de la variabilidad de la variable endógena.

Las significancias individuales y globales, el cumplimiento de los signos esperados de los estimadores y la buena bondad de ajuste nos hacen inferir que realmente el modelo está correctamente especificado y estimado, sin embargo aún hace falta revisar si cumple los supuestos de MCO. En la siguiente sección veremos, si el modelo planteado para los trabajadores independientes dedicados a la actividad comercial, cumple los supuestos de MCO.

- **Actividad prestadora de servicios.**

Al igual que en las anteriores actividades, empezamos restaurando la base de datos con el comando **restore** y a guardarla con el comando **preserve**. Proseguimos eligiendo las observaciones sobre los trabajadores independientes dedicados a la actividad prestadora de servicios con el comando **keep** y la condicional **if** usando la variable ***e13c***.

```
. restore

. preserve

. keep if e13c==3
(114 observations deleted)
```

Figura 3.52. Seleccionando a los trabajadores dedicados a la actividad prestadora de servicios.

Después de la selección de los trabajadores independientes dedicados a actividades prestadoras de servicios, construimos la variable ***gastoss*** que recoge

información sobre los gastos en los negocios, en la mano de obra y los gastos mensuales del capítulo 50 representados en ***e22t***, ***e25t1*** y ***e25t2*** respectivamente. En el caso de la existencia de trabajadores que no han registrado gastos, debemos descartar a los datos concernientes a aquellos trabajadores que no han percibido gastos. Recuerde; el descarte de datos faltantes conlleva a eliminar datos de otras variables con lo cual puede afectar a la estimación del modelo. El motivo por el cual se ha descartado los datos faltantes ha sido para que no la explicación no sea cansada.

```
. replace e22t=0 if e22t==.
(41 real changes made)

. gen gastoss=e22t+ e25t1+ e25t2

. drop if e22t==0
(42 observations deleted)
```

Figura 3.53. Construyendo la variable *gastoss.*

```
. reg e25t3 e20t gastoss e8a
```

Source	SS	df	MS	Number of obs	=	87
				F(3, 83)	=	651.25
Model	98242834.8	3	32747611.6	Prob > F	=	0.0000
Residual	4173593.14	83	50284.2547	R-squared	=	0.9592
				Adj R-squared	=	0.9578
Total	102416428	86	1190888.7	Root MSE	=	224.24

e25t3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e20t	.9820805	.0247914	39.61	0.000	.9327713	1.03139
gastoss	-.9630771	.0404672	-23.80	0.000	-1.043565	-.8825894
e8a	29.12427	45.43964	0.64	0.523	-61.25334	119.5019
_cons	17.79308	59.2564	0.30	0.765	-100.0655	135.6517

Figura 3.54. Regresión para los trabajadores independientes que se han dedicado a actividades prestadoras de servicios.

De la figura 3.54. Podemos ver el siguiente modelo econométrico.

$$\hat{G}_i = 17.80 + 0.98I_i - 0.96C_i + 29.12M_i + \hat{\mu}_i \text{ (3.7.21.)}$$

$$ee = (59.26) \quad (0.02) \quad (0.04) \quad (45.44)$$

$$t = 0.30 \quad 39.61 \quad -23.80 \quad 0.64$$

Interpretamos los resultados del modelo (3.7.21.).

Ingresos. Si los ingresos de los trabajadores independientes que se han dedicado a las actividades prestadoras de servicios aumentan en una unidad monetaria, las ganancias totales netas aumentan en 0.98 unidades monetarias.

Gastos. Si los gastos de los trabajadores independientes que se han dedicado a las actividades prestadoras de servicios aumentan en una unidad monetaria, las ganancias totales netas disminuyen en 0.96 unidades monetarias.

Número de trabajadores. Si el número de trabajadores a cargo de los trabajadores independientes, las ganancias netas totales aumentan en 29.12 unidades monetarias.

Veamos ahora si los estimadores tienen significancia individual en el modelo.

- **Ingresos.**

$$H_0: \beta_2 = 0 \text{ (3.7.22.)}$$

$$H_1: \beta_2 \neq 0$$

El estadístico *t* calculado según la figura 3.54. Es: $tc = 39.61$. Para hallar el estadístico *t* tabulado utilizamos el comando **scalar** y el comando **disp**.

```
. scalar tt=invt(83,1-0.05/2)

. disp tt
1.9889598
```

Figura 3.55. Calculando el estadístico *t* tabulado.

En la figura 3.55. Notamos que el estadístico *t* tabulado es: $tt_{83,\frac{0.05}{2}} = 1.98$ por lo tanto podemos rechazar la hipótesis nula, en consecuencia concluimos que el estimador de la variable ***e20t*** es distinto a 0 y la variable ***e20t*** tiene significancia individual.

- **Gastos**.

$$H_0: \beta_3 = 0 \text{ (3.7.23.)}$$

$$H_1: \beta_3 \neq 0$$

El estadístico *t* calculado es -23.80 y el estadístico *t* tabulado es 1.98 según la figura 3.55.

Por lo tanto se infiere:

$$|tc| > tt_{83,\frac{0.05}{2}} \text{ (3.7.24.)}$$

Entonces rechazamos la hipótesis nula y concluir que el estimador de ***gastoss*** es diferente de 0 y por tal motivo, la variable ***gastoss*** tiene relevancia individual en el modelo (3.7.21.).

- **Número de trabajadores.**

$$H_0: \beta_4 = 0 \text{ (3.7.25.)}$$

$$H_1: \beta_4 \neq 0$$

El estadístico *t* calculado es 0.64 y el estadístico *t* tabulado es 1.98 según la figura 3.55. Por consecuencia:

$$|tc| < tt_{82,\frac{0.05}{2}} \text{ (3.7.26.)}$$

Aceptamos la hipótesis nula y se concluye que el estimador de ***e8a*** es igual a 0 y podemos considerar retirar la variable ***e8a*** del modelo especificado (3.7.21.) debido a que no tiene relevancia individual.

De igual manera que en las anteriores regresiones, al revisar los respectivos valores-p de cada estimador, podemos notar que son menores a una significancia del 5%, por lo que se llega a las mismas conclusiones obtenidas en las pruebas de hipótesis sobre la relevancia individual de los estimadores de cada variable usada en el modelo econométrica (3.7.21.). De hecho, al revisar el valor-p del estimador de la variable ***e8a*** y observar que es mayor a una significancia del 5%, concluimos que se acepta la hipótesis nula y esta variable no es significativa.

En cuanto a la relevancia global del modelo utilizaremos la prueba F.

Planteamos la siguiente prueba de hipótesis.

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \text{ (3.7.27.)}$$

$$H_1: Ningún\ \beta_k\ es\ igual\ a\ 0$$

El estadístico *F* calculado es 651.25 y para hallar el estadístico *F* tabulado digitamos el comando **scalar** y posteriormente el comando **disp.**

```
. scalar ft=invF(3,83,1-0.05)

. disp ft
2.7145651
```

Figura 3.56. Calculando el estadístico *F* tabulado.

En consecuencia, el estadístico *F* tabulado es 2.71, entonces obtenemos:

$Fc > Ft^{3}_{82,0.05}$ (3.7.28.)

Se rechaza la hipótesis nula y se concluye que el modelo tiene significancia estadística global, por lo tanto el modelo sirve para explicar a la variable endógena. Si revisamos el valor-p del estadístico *F* calculado, llegamos a la misma conclusión, ya que este es 0.0000 y es menor a una significancia del 5%.

Finalmente, el coeficiente de determinación de esta regresión es 95.92% lo que significa que el modelo especificado explica el 95.92% de la variabilidad de la variable endógena.

Todas estas características, nos permiten concluir que el modelo econométrico está correctamente especificado y estimado. Sin embargo, aún tenemos que verificar si se satisface los supuestos de MCO. En el siguiente apartado se revisará si el modelo cumple los supuestos de MCO.

3.7.1.6. Evaluación del cumplimiento de los supuestos.

En esta sección se explicará cómo comprobar si los modelos econométricos anteriores cumplen los supuestos de MCO. En caso que el modelo no cumpla los supuestos de MCO, se explicará cómo corregir la violación del supuesto.

- **Actividades productivas/extractivas.**

A continuación, se muestra una réplica de la regresión anteriormente explicada con el fin de comparar con los resultados de un modelo corregido en caso sea necesario corregir el modelo especificado.

```
. reg e25t3 e14t gastos e8a

      Source |       SS           df       MS      Number of obs   =        27
-------------+----------------------------------   F(3, 23)        =     19.15
       Model |  23465252.3         3  7821750.76   Prob > F        =    0.0000
    Residual |  9394003.34        23  408434.928   R-squared       =    0.7141
-------------+----------------------------------   Adj R-squared   =    0.6768
       Total |  32859255.6        26  1263817.52   Root MSE        =    639.09

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e14t |   .5242571   .1713889     3.06   0.006     .1697121    .8788021
     gastos |  -.5867962   .1818076    -3.23   0.004    -.9628938   -.2106986
       e8a |   387.0503    175.662     2.20   0.038      23.6657    750.4349
      _cons |  -91.22355   228.4641    -0.40   0.693    -563.8376    381.3905
------------------------------------------------------------------------------
```

Figura 3.41. Regresión para los trabajadores independientes que se han dedicado a actividades productivas/extractivas.

- **No multicolinealidad**.

Ya que se ha podido estimar el modelo sin tener la necesidad de descartar alguna variable podemos inferir que no existe multicolinealidad perfecta. Sin embargo, debemos verificar si existe multicolinealidad imperfecta en el modelo. El comando **correlate** ejecuta una matriz de correlaciones de las variables que hemos seleccionado. En el caso de esta regresión múltiple, las variables son ***e25t3***, ***e14t***, ***gastos*** y ***e8a***.

```
. correlate e25t3 e14t gastos e8a
(obs=27)

             |   e25t3    e14t  gastos     e8a
-------------+------------------------------------
      e25t3  |  1.0000
       e14t  |  0.6900  1.0000
     gastos  |  0.5245  0.9503  1.0000
       e8a   |  0.7624  0.8671  0.7911  1.0000
```

Figura 3.57. Matriz de correlación de las variables utilizadas en la regresión para los trabajadores independientes que se han dedicado a actividades productivas/extractivas.

En esta matriz podemos notar que la correlación entre las variables **ingresos** (***e14t***) y **gastos** (***gastos***) tienen el coeficiente de correlación más alto siendo de 0.9503, mientras que las variables **ingresos** (***e14t***) y **número de trabajadores** (***e8a***) tienen el segundo coeficiente de correlación más alto 0.8671, y por último la correlación entre las variables **gastos** (***gastos***) y **número de trabajadores** (***e8a***) tienen el tercer

coeficiente más alto y es de 0.7911. Observando que las explicativas tienen coeficientes de correlación altísimos, podemos inferir que la multicolinealidad imperfecta está presente y se tienen a las variables ***e14t*** y ***gastos*** como las posibles variables causantes de la multicolinealidad.

Otra forma de verificar la existencia de multicolinealidad es usando el factor de inflación de varianza (VIF) y el índice de tolerancia (TOL). STATA permite obtener ambos índices empleando el comando **vif** como comando postestimación, es decir este comando debe ser ejecutado después del comando **reg**.

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
        e14t |     16.29    0.061377
      gastos |     10.81    0.092540
         e8a |      4.22    0.236990
-------------+----------------------
    Mean VIF |     10.44
```

Figura 3.58. VIF y TOL de la regresión.

La segunda columna de la tabla que se muestra en la figura anterior indica el factor de inflación de varianza (VIF) y la tercera columna señala el índice de tolerancia (TOL). Acorde a (Hanke & Wichern, 2006) Los estimadores de las variables que tienen un VIF cercano a 1 tienden a tener resultados más estables a comparación de los estimadores de aquellas variables cuyos VIF se acercan a 10 o en su defecto lo superan. La variable $\boldsymbol{e8a}$ tiene un VIF cercano a 1, entonces esta variable tiene un estimador y estadístico *t* calculado más estable. En contraposición, $\boldsymbol{e14t}$ y $\boldsymbol{gastos}$ tienen un VIF superior a 10, por lo tanto sus estimadores no son tan estables y podrían ser las causantes de multicolinealidad imperfecta, sobre todo la variable $\boldsymbol{e14t}$.

Algunos autores como (Escobar M., Fernández M., & Bernardi, 2012) Aconsejan no introducir en el modelo variables con VIF superior a 10, mientras que otros autores aconsejan descartar VIF superiores a 30. Entones, deberíamos plantearnos la existencia de multicolinealidad por parte de $\boldsymbol{e14t}$.

Por otro lado, el índice de tolerancia (TOL), el cual es el inverso del VIF, nos permite llegar a la misma conclusión; en el caso de la variable $\boldsymbol{e14t}$ es la variable cuyo índice de tolerancia se acerca más a 0, por lo que se concluye lo mismo que se pudo inferir con el VIF.

Ahora veamos la siguiente gráfica matricial sobre las variables explicativas para descubrir un patrón entre ellas y tener más indicios que verdaderamente existe multicolinealidad. El comando **graph matrix** y la opción **half name (G3)** es la instrucción usada para generar la gráfica matricial**.**

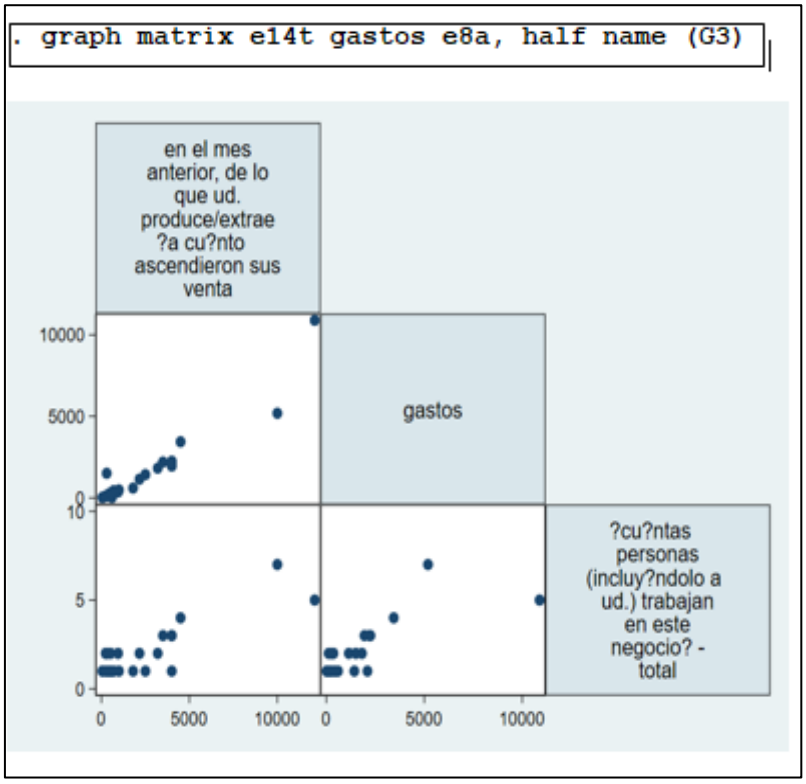


Figura 3.59. Gráfica matricial entre las variables explicativas.

En el gráfico que se puede ver en la figura 3.59. Podemos observar la existencia de una correlación positiva muy notoria entre las variables $\boldsymbol{e14t}$ y $\boldsymbol{gastos}$. Incluso, en las demás gráficas también se observan patrones de correlación positiva, sin embargo, en el caso de $\boldsymbol{e14t}$ y $\boldsymbol{gastos}$ es más notorio. Sin embargo, (Gujarati & Porter, 2010) Indican que esta gráfica no es determinante para concluir la existencia de multicolinealidad en el modelo, y se debería complementar con la información obtenida en la matriz de correlaciones, y el índice de VIF y TOL. Con estos indicios podemos concluir que la variable $\boldsymbol{e14t}$ es posiblemente la causante de multicolinealidad imperfecta en el modelo.

Una vez seleccionada la variable que asumimos es la culpable de generar multicolinealidad en el modelo, aplicaremos la regla de Klein. El modelo auxiliar especificado para aplicar la regla de Klein será:

$$I_i = \alpha_1 + \alpha_2 C_i + \alpha_3 N_i + e_i \text{ (3.7.29.)}$$

Al realizar la regresión del modelo auxiliar (3.7.29.) obtenemos el siguiente resultado.

```
. reg e14t gastos e8a

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(2, 24)        =    183.51
       Model |  212637921      2  106318960   Prob > F        =    0.0000
    Residual |   13904564     24  579356.834   R-squared       =    0.9386
-------------+------------------------------   Adj R-squared   =    0.9335
       Total |  226542485     26  8713172.49   Root MSE        =    761.15

------------------------------------------------------------------------------
        e14t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gastos |   .9203056   .1076862     8.55   0.000     .6980522    1.142559
        e8a |    620.5742    166.505     3.73   0.001     276.9248    964.2235
       _cons |   -280.6165   266.0033    -1.05   0.302    -829.6202    268.3873
------------------------------------------------------------------------------
```

Figura 3.60. Resultados de la regresión auxiliar (3.7.29.).

En la figura 3.60. Podemos notar que el coeficiente de determinación del modelo auxiliar es 0.9386 mientras el coeficiente del modelo original es 0.7141. Según la **regla de Klein**, al ser $R_i^2 > R^2$, entonces concluimos que la multicolinealidad está presente en esta variable explicativa. Veamos ahora cómo se verifica la multicolinealidad mediante el $R^2\ de\ Thiel$, se especifican los siguientes modelos auxiliares en (3.7.30.), (3.7.31.) y (3.7.32.).

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 C_i + v_i \text{ (3.7.30.)}$$

$$G_i = \theta_1 + \theta_2 C_i + \theta_3 N_i + v_i \text{ (3.7.31.)}$$

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 N_i + v_i \text{ (3.7.32.)}$$

Posteriormente, tomaremos sus respectivos coeficientes de determinación y los utilizaremos para calcular el $R^2\ de\ Thiel$ con la siguiente fórmula.

$$R^2 Theil = R^2 - [\sum R^2 - R_i^2] \text{ (3.6.47.)}$$

Veamos los resultados de las regresiones con el comando **reg**.

```
. reg e25t3 e14t gastos

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(2, 24)        =     22.66
       Model |  21482347.4     2  10741173.7   Prob > F        =    0.0000
    Residual |  11376908.3    24  474037.844   R-squared       =    0.6538
-------------+------------------------------   Adj R-squared   =    0.6249
       Total |  32859255.6    26  1263817.52   Root MSE        =     688.5

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e14t |   .7529069   .1469486     5.12   0.000     .4496198    1.056194
     gastos |  -.6717919   .1914055    -3.51   0.002    -1.066833   -.2767503
      _cons |   270.7832   171.0284     1.58   0.126    -82.20207    623.7686
------------------------------------------------------------------------------
```

Figura 3.61. Resultados de la regresión auxiliar (3.7.30.).

```
. reg e25t3 gastos e8a

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(2, 24)        =     17.84
       Model |  19643645.7     2  9821822.87   Prob > F        =    0.0000
    Residual |  13215609.9    24  550650.412   R-squared       =    0.5978
-------------+------------------------------   Adj R-squared   =    0.5643
       Total |  32859255.6    26  1263817.52   Root MSE        =    742.06

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gastos |  -.1043195   .1049845    -0.99   0.330    -.3209967    .1123578
       e8a |   712.3907   162.3275     4.39   0.000     377.3632    1047.418
      _cons |  -238.3387   259.3295    -0.92   0.367    -773.5684    296.891
------------------------------------------------------------------------------
```

Figura 3.62. Resultados de la regresión auxiliar (3.7.31.).

```
. reg e25t3 e14t e8a

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(2, 24)        =     16.89
       Model |  19210506.6     2  9605253.3   Prob > F        =    0.0000
    Residual |    13648749    24  568697.876   R-squared       =    0.5846
-------------+------------------------------   Adj R-squared   =    0.5500
       Total |  32859255.6    26  1263817.52   Root MSE        =    754.12

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e14t |   .0443457    .100577     0.44   0.663     -.163235    .2519265
       e8a |   507.3456   202.5605     2.50   0.019     89.28134    925.4099
      _cons |  -85.67793   269.5784    -0.32   0.753    -642.0605    470.7046
------------------------------------------------------------------------------
```

Figura 3.63. Resultados de la regresión auxiliar (3.7.32.).

$$R^2\ de\ Theil = 0.7141 - (0.7141 - 0.6538) - (0.7141 - 0.5978) - (0.7141 - 0.5500) = 0.9343 \quad (3.7.33.)$$

El método del $R^2\ de\ Theil$ nos indica que existe multicolinealidad, ya que según (3.7.33.), $R^2\ de\ Theil > R^2$.

Finalmente veamos el contraste de hipótesis, la cual plantea la siguiente prueba de hipótesis.

H_0: $No\ existe\ multicolinealidad$ (3.7.34.)

$$H_1: Existe\ multicolinealidad$$

Para este contraste se utiliza la siguiente regresión auxiliar $I_i = \alpha_1 + \alpha_2 C_i + \alpha_3 N_i + e_i$ cuyos resultados se pueden ver en la figura 3.60. Los estadísticos *F* de esta regresión serán usados para el contraste de hipótesis. El estadístico *F* calculado es hallado mediante.

$$Fc = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)} = \frac{0.9386/(3-2)}{(1-0.9386)/(27-3+1)} = 382.16$$

Por otro lado el estadístico *F* tabulado se halla en STATA mediante las instrucciones **scalar** y **disp**, usando sus respectivos grados de libertad y una significancia del 5%.

```
. scalar ft=invF(1,25,1-0.05)

. disp ft
4.2416991
```

Figura 3.64. Prueba *F* para la multicolinealidad.

Concluimos que $|Fc| > Ft^1_{25,0.05}$, por lo que rechazamos la hipótesis nula y en consecuencia el modelo original presenta multicolinealidad imperfecta causada por la variable $\boldsymbol{e14t}$ tal como se muestra en la regresión auxiliar (3.7.29.). Es necesario mencionar que esta prueba de hipótesis se utiliza sobre las regresiones auxiliares, donde cada variable explicativa pasa a ser la variable dependiente y es explicada por las demás variables explicativas. En este ejemplo debieron ser 3 modelos auxiliares, sin embargo como en el modelo auxiliar (3.7.29.) se consiguió el coeficiente de determinación más alto, las demás regresiones auxiliares no se han tomado en cuenta.

Estos indicios nos indican que realmente existe multicolinealidad en el modelo y ha sido necesario utilizar todos estos métodos, porque es difícil determinar si el problema de la multicolinealidad en el modelo realmente afecta de manera considerablemente negativa en sus resultados, según la teoría expuesta.

Podríamos intentar corregir el modelo retirando la variable $\boldsymbol{e14t}$, sin embargo, si estimamos un modelo sin esta variable, la variable $\boldsymbol{gastos}$ no es significativa y peor aún no tiene el signo esperado, pese a que tiene relevancia global, según la figura 3.62. En consecuencia, este método correctivo no podría ser aplicado.

Otra forma de intentar corregir el modelo, sería utilizando **restricciones**, sin embargo el marco teórico no muestra que restricción se pueda utilizar al modelo, por lo que no podríamos realmente usar este método correctivo. Se podría intentar aumentar la muestra ya que el modelo usa una muestra pequeña en comparación a los otros dos modelos, no obstante, esta muestra es toda la muestra disponible en ENAHO entonces no habría más datos disponibles para usar.

Analicemos si, transformando el modelo con respecto a la variable que genera multicolinealidad, se soluciona este problema.

$$\frac{G_i}{I_i} = \beta_1 \left(\frac{1}{I_i}\right) + \beta_3 \left(\frac{C_i}{I_i}\right) + \beta_4 \left(\frac{N_i}{I_i}\right) + \left(\frac{\mu_i}{I_i}\right) \text{ (3.7.35.)}$$

En STATA crearemos las variables del modelo transformado (3.7.35.) con los comandos **gen** de la siguiente forma.

```
. gen yx2= e25t3/e14t

. gen x3x2=gastos/e14t

. gen x4x2=e8a/e14t
```

Figura 3.65. Variables explicativas del modelo transformado.

Y con estas variables procedemos a realizar la regresión transformada.

```
. reg yx2 x3x2 x4x2

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(2, 24)      =  123.39
       Model |  49.0456234     2  24.5228117            Prob > F      =  0.0000
    Residual |  4.76970077    24  .198737532           R-squared     =  0.9114
-------------+------------------------------           Adj R-squared =  0.9040
       Total |  53.8153241    26  2.06982016           Root MSE      =   .4458

------------------------------------------------------------------------------
         yx2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        x3x2 |   1.540512   .0996832    15.45   0.000     1.334776    1.746248
        x4x2 |     8.0028   20.79975     0.38   0.704    -34.92578    50.93138
       _cons |  -.2249454   .1196378    -1.88   0.072    -.4718656    .0219747
------------------------------------------------------------------------------
```

Figura 3.66. Regresión del modelo transformado.

Los aspectos positivos de la regresión del modelo transformado son: la obtención de un coeficiente de determinación superior al del modelo original, la conservación de la relevancia global del modelo y el estimador de la variable $\left(\frac{C_i}{I_i}\right)$ tiene significancia individual. No obstante, también tienen aspectos negativos en comparación con el modelo original, entre ellos el estimador no cumple con los signos esperados y el estimador de la

variable $\left(\frac{N_i}{I_i}\right)$ no es significativo, por lo que transformar el modelo no parece ser una solución eficaz contra la multicolinealidad, ya que algunos estimadores pierden la capacidad de cumplir los signos esperados y su significancia individual. Además es posible que el modelo contenga un término de perturbación heterocedástico, por lo que puede estar afectando a la varianza de los estimadores y concluir falsos resultados.

Para despejar las dudas, veremos si el modelo transformado tiene varianza heterocedástica mediante las pruebas de White y Bresuch-Pagan (BP) con los comandos **estat imtest** y **estar hettest** respectivamente, para ambos test utilizaremos la misma prueba de hipótesis.

$$H_0: No\ existe\ heterocedasticidad$$

$$H_1: Existe\ heterocedasticidad$$

Empecemos con la prueba de White.

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     19.19
         Prob > chi2  =    0.0018

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      19.19      5    0.0018
            Skewness |       2.37      2    0.3065
            Kurtosis |       5.91      1    0.0150
---------------------+-----------------------------
               Total |      27.47      8    0.0006
---------------------------------------------------
```

Figura 3.67. Prueba de White para la regresión del modelo transformado.

El comando **estat imtest** con la opción **white** indica a STATA que muestre la prueba de hipótesis de White junto a una tabla donde se ven otros estadísticos como Skewness y Kurtosis, de momento estos estadísticos no interesan en el análisis.

Según la figura 3.67., se puede apreciar que la probabilidad es menor a la significancia del 5%, ya que $Prob > chi2 = 0.0018$, por lo tanto rechazamos la

hipótesis nula y comprobamos que en el modelo transformado existe heterocedasticidad. Sin embargo, recordemos que al tomar productos cruzados de las variables explicativas para realizar esta prueba, no solo puede existir heterocedasticidad, sino posiblemente también un sesgo de especificación, según lo explicado. En consecuencia, para tener una mejor resultado debemos retirar el producto cruzado tendremos que realizar la prueba manualmente en STATA porque el comando **estat imtest** si toma en cuenta el producto cruzado. Primero debemos obtener los residuos del modelo transformado, con el comando **predict** y su opción **residuals**.

```
. predict uw, resid
```

Figura 3.68. Comando predict.

Seguido del comando **predict** debemos colocar el nombre de la nueva variable que representarán los residuos, en la figura anterior podemos ver que se le ha nombrado como "***uw***". Posteriormente debemos elevar al cuadrado los residuos y a las variables explicativas con el comando **gen**.

```
. gen uw2=uw^2

. gen x3x2c= x3x2^2

. gen x4x2c= x4x2^2
```

Figura 3.69. Creando las variables para la prueba de White de heterocedasticidad pura.

```
. reg uw2 x3x2 x4x2 x3x2c x4x2c

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(4, 22)        =      0.75
       Model |  .343798989     4  .085949747   Prob > F        =    0.5663
    Residual |  2.50906147    22  .114048249   R-squared       =    0.1205
-------------+------------------------------   Adj R-squared   =   -0.0394
       Total |  2.85286046    26  .109725402   Root MSE        =    .33771

------------------------------------------------------------------------------
         uw2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        x3x2 |  -.6430882   .4899323    -1.31   0.203    -1.659146    .3729692
        x4x2 |  -62.11597   52.99613    -1.17   0.254    -172.0232    47.79127
      x3x2c |   .1233462   .0917832     1.34   0.193    -.0670004    .3136928
      x4x2c |    2567.86   2801.795     0.92   0.369    -3242.708    8378.427
       _cons |   .5728265   .2576966     2.22   0.037     .0383964    1.107257
------------------------------------------------------------------------------
```

Figura 3.70. Prueba de heterocedasticidad de White pura.

Construimos el siguiente estadístico calculado con la figura 3.70.

$27 * 0.5663 = 15.29$ (3.7.36.)

Y el estadístico tabulado es $X^2_{0.05,4} = 0.711$, entonces al ser el estadístico calculado mayor al estadístico tabulado podemos rechazar la hipótesis nula y concluir que el modelo transformado efectivamente tiene una varianza heterocedástica.

Apliquemos ahora el comando **estat hettest** para probar la heterocedasticidad mediante la prueba BP.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of yx2

         chi2(1)      =     0.14
         Prob > chi2  =   0.7075
```

Figura 3.71. Prueba de heterocedasticidad de BP.

Debido a que la probabilidad es mayor al 5%, aceptamos la hipótesis nula y concluimos que mediante esta prueba el modelo transformado no presenta heterocedasticidad. Esta prueba contradice la conclusión que obtuvimos de la prueba de White, entonces surge la pregunta ¿Cuál es la prueba debemos seguir para verificar la existencia o no de heterocedasticidad? En este tipo de situaciones conviene revisar las gráficas de dispersión entre los residuos y las variables.

El comando **rvfplot** indica a STATA que muestre una gráfica de dispersión entre los residuos y la variable dependiente estimada y el comando **rvpplot** ordena a STATA que genere gráficas de dispersión entre los residuos y los valores de la variable independiente.

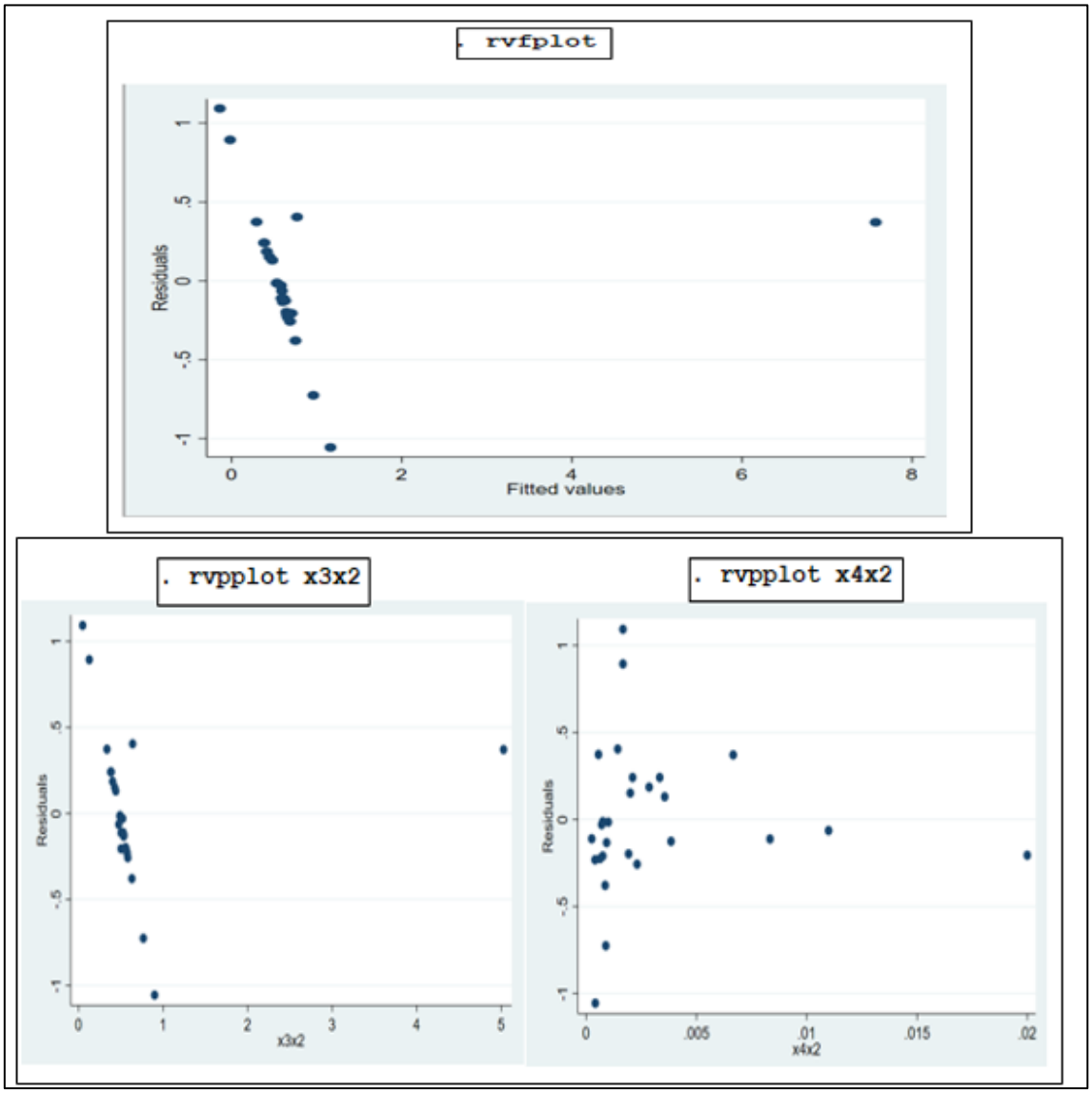


Figura 3.72. Gráfica de dispersión entre los residuos y la variable dependiente estimada del modelo transformado y gráficas de dispersión entre los residuos y las variables explicativas.

En la figura 3.72., se han generado tres gráficas, en la gráfica superior se observa la dispersión entre los residuos y los valores estimadores de la variable dependiente. A simple vista, se puede notar la existencia de un patrón descendente y la existencia de un dato atípico, por lo que esta gráfica sugiere que efectivamente hay heterocedasticidad.

En las dos gráficas inferiores se aprecian gráficas de dispersión entre los residuos y las variables independientes. En ambas gráficas observamos valores atípicos y la

existencia de un patrón entre las variables, por lo que complementando la información obtenida de las gráficas con la prueba de hipótesis de White, tanto de heterocedasticidad pura como la prueba de White de heterocedasticidad con sesgo de especificación, se concluye que el modelo transformado puede tener heterocedasticidad. Si bien se puede ejecutar un tratamiento a la multicolinealidad en el modelo original, también puede ocasionar estimadores ineficientes. Entones, este método correctivo queda descartado.

Se podría utilizar el **análisis de componentes principales** sin embargo este método, al igual que el **método de regresión en cadena,** pueden ser contraproducentes para la correcta estimación del modelo, por ello no realizaremos estos métodos correctivos.

Finalmente, en vista que los estimadores del modelo original tienen significancias individuales y un coeficiente de determinación relativamente alto, se puede confiar en que la multicolinealidad no afecta demasiado al modelo original. Por lo que siguiendo la teoría que proponen (Gujarati & Porter, 2010) Se elegirá no realizar ningún método correctivo.

- **Homocedasticidad.**

Recordando que estos son los valores del modelo original para los trabajadores dedicados a las actividades productivas/extractivas.

```
. reg e25t3 e14t gastos e8a
```

Source	SS	df	MS	Number of obs	=	27
				F(3, 23)	=	19.15
Model	23465252.3	3	7821750.76	Prob > F	=	0.0000
Residual	9394003.34	23	408434.928	R-squared	=	0.7141
				Adj R-squared	=	0.6768
Total	32859255.6	26	1263817.52	Root MSE	=	639.09

e25t3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e14t	.5242571	.1713889	3.06	0.006	.1697121	.8788021
gastos	-.5867962	.1818076	-3.23	0.004	-.9628938	-.2106986
e8a	387.0503	175.662	2.20	0.038	23.6657	750.4349
_cons	-91.22355	228.4641	-0.40	0.693	-563.8376	381.3905

Figura 3.41. Regresión para los trabajadores independientes que se han dedicado a actividades productivas/extractivas.

Veremos ahora si el modelo original presenta problemas de heterocedasticidad y en caso de ser así, se analizará cómo ejecutar un método que sirva de tratamiento para corregir el modelo.

Primero veremos las gráficas de dispersión entre los residuos y los valores estimados de la variable dependiente y con las variables explicativas. El comando **rvfplot** muestra una gráfica de dispersión entre los residuos del modelo y los valores estimados de la variable dependiente y su opción **yline(0)** traza una línea horizontal cuando $\hat{Y}_i = 0$.

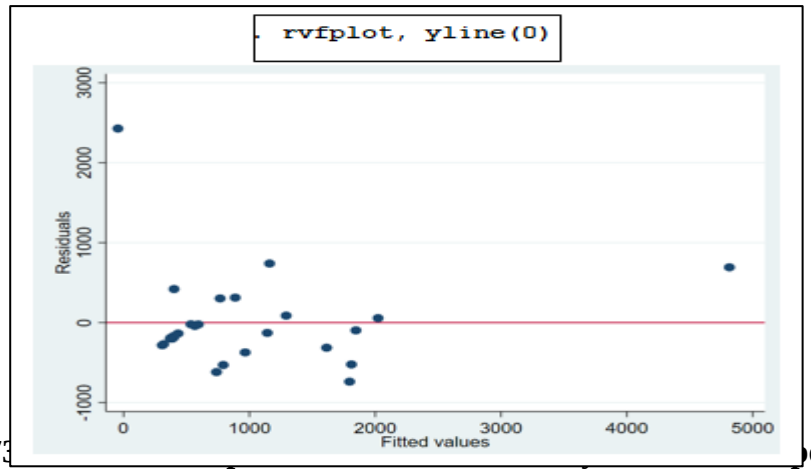


Figura 3.73 [illegible] pendiente estimada del modelo original.

Este gráfico ya nos indica que podemos tener sospechas que el modelo original tiene heterocedasticidad debido a que hay dos valores atípicos mientras que existe una nube de dispersión que se concentran en la esquina inferior izquierda.

Comprobemos ahora la gráfica de dispersión entre los residuos y las variables explicativas.

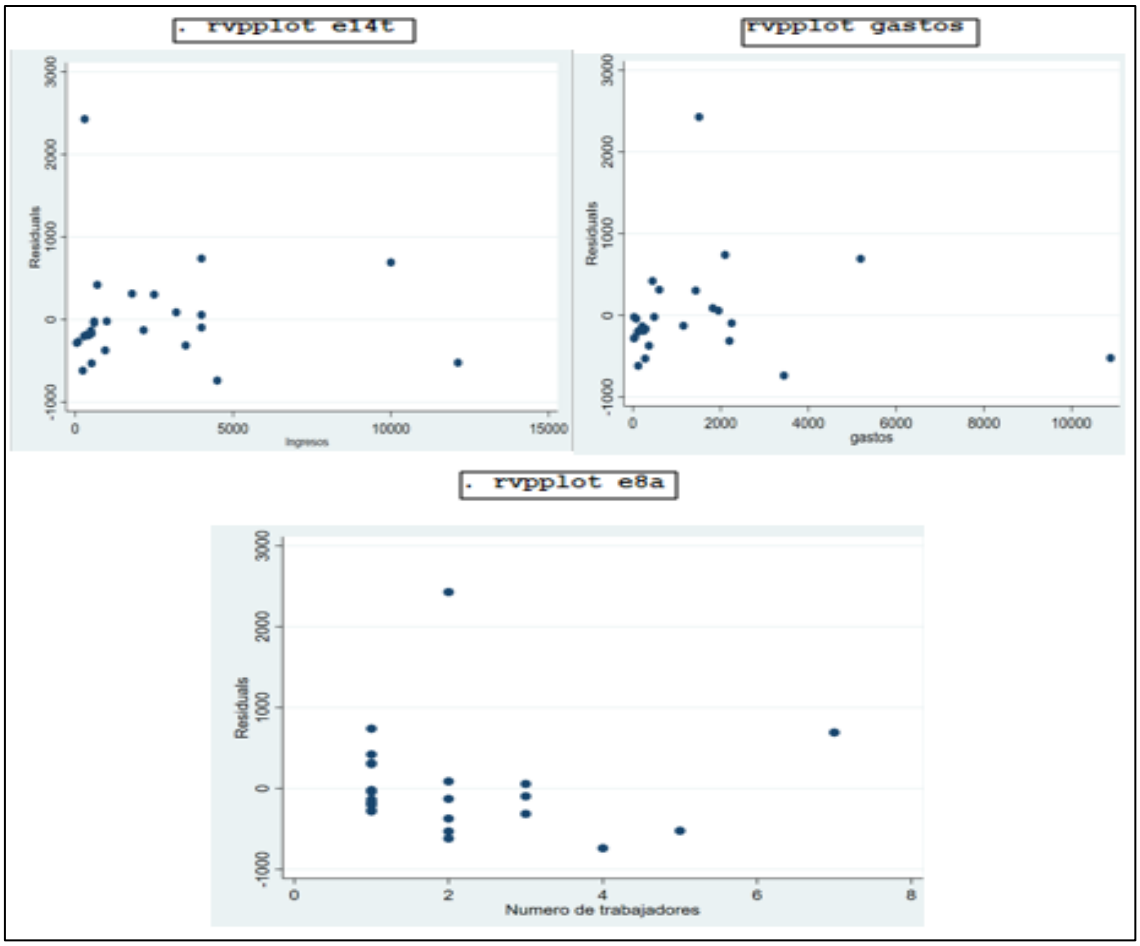


Figura 3.74. Grafica de dispersión entre los residuos y las variables explicativas del modelo original.

Dado que en los tres gráficos de la figura 3.74., se pueden ver valores atípicos, se puede concluir que el modelo original puede tener heterocedasticidad pero no se puede tener una idea clara sobre cuál es la variable que la causa.

Veamos los resultados que se pueden obtener de las pruebas formales para determinar si existe heterocedasticidad en el modelo. Siguiendo en ambas la siguiente prueba de hipótesis.

$$H_0: No\ existe\ heterocedasticidad$$

$$H_1: Existe\ heterocedasticidad$$

Con el comando **estat hettest** se le indica a STATA que ejecute la prueba BP.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of e25t3

         chi2(1)      =     2.72
         Prob > chi2  =   0.0989
```

Figura 3.75. Prueba BP del modelo original.

En esta figura podemos ver que la prueba BP indica que la probabilidad, que es $Prob > chi2 = 0.0989$, es mayor a una significancia del 5% por lo tanto no se rechaza la prueba de hipótesis y se concluye que no existe heterocedasticidad en el modelo. El comando **estat hettest** utiliza a los valores estimados de la variable dependiente del modelo, sin embargo, al introducir una variable explicativa al lado del comando podemos comprobar si una variable explicativa causa heterocedasticidad.

```
. estat hettest e14t

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e14t

         chi2(1)      =     0.39
         Prob > chi2  =   0.5298

. estat hettest gastos

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: gastos

         chi2(1)      =     0.80
         Prob > chi2  =   0.3714

. estat hettest e8a

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e8a

         chi2(1)      =     1.09
         Prob > chi2  =   0.2956
```

Figura 3.76. Prueba BP para las variables *e14t, gastos* y *e8a* del modelo original (1).

Según las pruebas de BP aplicadas a cada una de las variables explicativas del modelo, en ninguna prueba de hipótesis se pueden rechazar la hipótesis nula ya que sus

valores-p son mayores a 5%, por ello según la prueba BP, ninguna variable explicativa podría generar heterocedasticidad en el modelo. Estas pruebas de hipótesis pueden contenerse en una sola tabla aplicando la opción **mtest** y digitando las variables explicativas.

```
. estat hettest e14t gastos e8a, mtest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

-------------------------------------------------------
    Variable |      chi2    df      p
-------------+-----------------------------------------
        e14t |      0.39     1    0.5298 #
      gastos |      0.80     1    0.3714 #
        e8a |      1.09     1    0.2956 #
-------------+-----------------------------------------
simultaneous |     41.89     3    0.0000
-------------------------------------------------------
                  # unadjusted p-values
```

Figura 3.77. Prueba BP para las variables *e14t, gastos* y *e8a* del modelo original (2).

La novedad de la opción **mtest**, es que muestra si las variables explicativas del modelo causan heterocedasticidad en el modelo original utilizando la prueba de BP, la segunda columna muestra los estadísticos calculados, la siguiente muestra los grados de libertad y la última columna muestra sus respectivos valores-p.

Su característica más importante es que muestra si las variables explicativas simultáneamente generan heterocedasticidad y al observar que su valor-p es 0.0000 entonces podemos rechazar la hipótesis nula y concluir que existe heterocedasticidad en el modelo. Por otro lado, al no poder rechazar las pruebas de hipótesis de las variables explicativas no se puede determinar cuál es la estructura de la varianza heterocedástica del término de perturbación, es decir, la varianza heterocedástica es desconocida.

Veamos ahora si mediante la prueba de White se puede concluir la existencia de heterocedasticidad en el modelo original con el comando **estat imtest** y su opción **white**.

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     26.87
         Prob > chi2  =    0.0015

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	26.87	9	0.0015
Skewness	11.79	3	0.0082
Kurtosis	1.31	1	0.2523
Total	39.96	13	0.0001

Figura 3.78. Prueba de White general para la heterocedasticidad.

En la prueba de White general indica que el valor-p es 0.0015 y ya que es menor a una significancia del 5% se rechaza la hipótesis nula y se concluye que existe heterocedasticidad en el modelo.

El rechazo de la hipótesis nula no necesariamente se debe a presencia de heterocedasticidad, sino también puede estar ocasionado por un sesgo de especificación, por lo tanto, se recomienda aplicar la prueba de White pura retirando los productos cruzados de la prueba de la regresión auxiliar, según la teoría de (Gujarati & Porter, 2010). Para ello, con el comando **predict** y la opción **resid** obtendremos los residuos del modelo original y después con el comando **gen** crearemos una variable que represente los residuos al cuadrado.

```
. predict u,resid

. gen uc=u^2
```

Figura 3.79. Generando los residuos al cuadrado del modelo original.

Ahora crearemos los cuadrados de las variables explicativas para usarlas en la regresión auxiliar.

```
. gen x2c=e14t^2

. gen x3c=gastos^2

. gen x4c=e8a^2
```

Figura 3.80. Generando los cuadrados de la variable explicativa del modelo original.

Finalmente podremos realizar la regresión auxiliar para verificar si existe heterocedasticidad en el modelo mediante la prueba de White pura.

```
. reg uc e14t gastos e8a x2c x3c x4c

      Source |       SS       df       MS      Number of obs   =        27
-------------+------------------------------   F(6, 20)        =     15.22
       Model |  2.6817e+13     6  4.4695e+12   Prob > F        =    0.0000
    Residual |  5.8748e+12    20  2.9374e+11   R-squared       =    0.8203
-------------+------------------------------   Adj R-squared   =    0.7664
       Total |  3.2692e+13    26  1.2574e+12   Root MSE        =    5.4e+05

------------------------------------------------------------------------------
          uc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e14t |  -2730.289   332.5385    -8.21   0.000    -3423.952   -2036.626
      gastos |    3878.34   440.2744     8.81   0.000     2959.944    4796.737
         e8a |    1067581   562357.2     1.90   0.072    -105475.5    2240637
         x2c |    .272186    .064741     4.20   0.000     .1371387    .4072333
         x3c |   -.3849661   .0652432    -5.90   0.000     -.521061   -.2488711
         x4c |   -339339.9   129716.3    -2.62   0.017    -609923.5   -68756.36
       _cons |   -94361.47   460337.2    -0.20   0.840     -1054608      865885
------------------------------------------------------------------------------
```

Figura 3.81. Regresión auxiliar para la prueba de White de heterocedasticidad pura.

El estadístico calculado es: $27 * 0.8203 = 22.1481$, mientras el estadístico tabulado es $X^2_{0.05,6} = 12.6$, siendo el estadístico calculado mayor al estadístico tabulado. Por lo tanto, se rechaza la hipótesis nula y se concluye que efectivamente existe heterocedasticidad en el modelo mediante la prueba de White pura.

Estas pruebas formales e informales, indican que existe heterocedasticidad en el modelo. No obstante, la varianza del término de perturbación no es conocida por lo que es recomendable utilizar el método correctivo de los errores de White, para ello simplemente agregamos la opción **robust** en la regresión original.

```
. reg e25t3 e14t gastos e8a, robust

Linear regression                               Number of obs     =         27
                                                F(3, 23)          =      17.10
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7141
                                                Root MSE          =     639.09

------------------------------------------------------------------------------
             |               Robust
      e25t3  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e14t |   .5242571   .4133925     1.27   0.217    -.3309105    1.379425
      gastos |  -.5867962   .3896446    -1.51   0.146    -1.392837     .219245
         e8a |   387.0503    296.036     1.31   0.204    -225.3469    999.4475
       _cons |  -91.22355   187.1858    -0.49   0.631    -478.4469    295.9998
------------------------------------------------------------------------------
```

Figura 3.82. Modelo corregido mediante los errores robustos de White.

Con los resultados de la figura 3.82. Se muestran los siguientes resultados del modelo original con errores robustos.

$$\hat{G}_i = -91.22 + 0.52I_i - 0.58C_i + 387.05M_i + \hat{\mu}_i \text{ (3.7.37.)}$$

$$ee = (187.19)\ (0.41)\quad (0.38)\quad (296.03)$$

$$t = -0.49\quad 1.27\quad -1.51\quad 1.31$$

Y estos son los resultados del modelo original.

$$\hat{G}_i = -91.22 + 0.52I_i - 0.58C_i + 387.05M_i + \hat{\mu}_i \text{ (3.7.2.)}$$

$$ee = (228.46)\quad (0.17)\quad (0.18)\quad (175.66)$$

$$t = -0.40\quad 3.06\quad -3.23\quad 2.20$$

Lo primero que se observa es que, los errores estándares de los estimadores que acompañan a las variables (calculados mediante errores de White), son mayores que los errores estándares de los estimadores hallados mediante MCO. En consecuencia, los estadísticos *t* calculados son menores, y con estos también han cambiado sus respectivos valores-p. Por tanto, las conclusiones de las pruebas de hipótesis sobre las significancias individuales indican que ningún estimador es significativo. Sin embargo, este método permite conservar los signos esperados de los estimadores y además, el modelo con errores robustos conserva una buena bondad de ajuste. De esta forma el modelo original ha sido corregido por el método de errores robustos.

- **Actividad comercial.**

En esta sección comprobaremos si el modelo estimado para los trabajadores independientes que se han dedicado a actividades comerciales, cumple con los supuestos de independencia entre los regresores y homocedasticidad.

Veamos los resultados de su regresión, previamente a realizar los procedimientos necesarios para verificar si realmente cumplen los supuestos de MCO.

```
. reg e25t3 e17t gastosc e8a

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(3, 82)        =     67.64
       Model |  58918846.4     3  19639615.5   Prob > F        =    0.0000
    Residual |  23809986.8    82  290365.692   R-squared       =    0.7122
-------------+------------------------------   Adj R-squared   =    0.7017
       Total |  82728833.2    85  973280.391   Root MSE        =    538.86

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |   .5539907   .0898667     6.16   0.000     .3752171    .7327642
    gastosc |   -.508537   .1025509    -4.96   0.000    -.7125434   -.3045306
        e8a |   779.6868   76.10672    10.24   0.000     628.2863    931.0873
      _cons |    -635.51   122.5403    -5.19   0.000    -879.2817   -391.7383
------------------------------------------------------------------------------
```

Figura 3.49. Regresión para los trabajadores independientes que se han dedicado a actividades comerciales.

- No multicolinealidad.

En vista que STATA ha logrado realizar la regresión sin mostrar ningún error en el modelo, se puede intuir que no existe multicolinealidad perfecta, entonces se verificará si existe multicolinealidad imperfecta.

Veamos en primer lugar los principales indicios para determinar la existencia de multicolinealidad: **matriz de correlación** e **índices VIF y TOL** de las variables empleadas en el modelo.

```
. corr e25t3 e17t gastosc e8a
(obs=86)

             |   e25t3    e17t  gastosc     e8a
-------------+------------------------------------
      e25t3  |  1.0000
       e17t  |  0.5653  1.0000
    gastosc  |  0.5350  0.9883  1.0000
        e8a  |  0.6136  0.1126  0.1468  1.0000
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
     gastosc |     45.63    0.021916
        e17t |     45.22    0.022115
         e8a |      1.07    0.932957
-------------+----------------------
    Mean VIF |     30.64
```

Figura 3.83. Matriz de correlación e índices VIF y TOL de las variables explicativas de la regresión para los trabajadores independientes que se han dedicado a actividades comerciales.

En la parte superior de la anterior figura se aprecia la matriz de correlación, la cual indica el coeficiente de correlación entre las variables **ingresos** (***e17t***) y **gastos** (***gastosc***) es demasiado alto a comparación de los otros coeficientes de correlación. Entonces, ya podemos tener un indicio que existe multicolinealidad imperfecta y que los estimadores de las variables **ingresos** (***e17t***) y **gastos** (***gastosc***)**,** pueden estar influenciados**.**

En la parte inferior de la figura, se observa la tabla sobre los índices de factor de inflación de la varianza y tolerancia. Cuyas interpretaciones señalan a las variables **ingresos** (***e17t***) y **gastos** (***gastosc***) como las causantes de multicolinealidad, y ya que sus índices VIF son mayores a 30, se concluye que los estimadores pueden estar influenciados y ser inestables por la multicolinealidad.

Al revisar la gráfica matricial de las variables explicativa nos puede dar una mejor idea sobre cómo se correlacionan las variables.

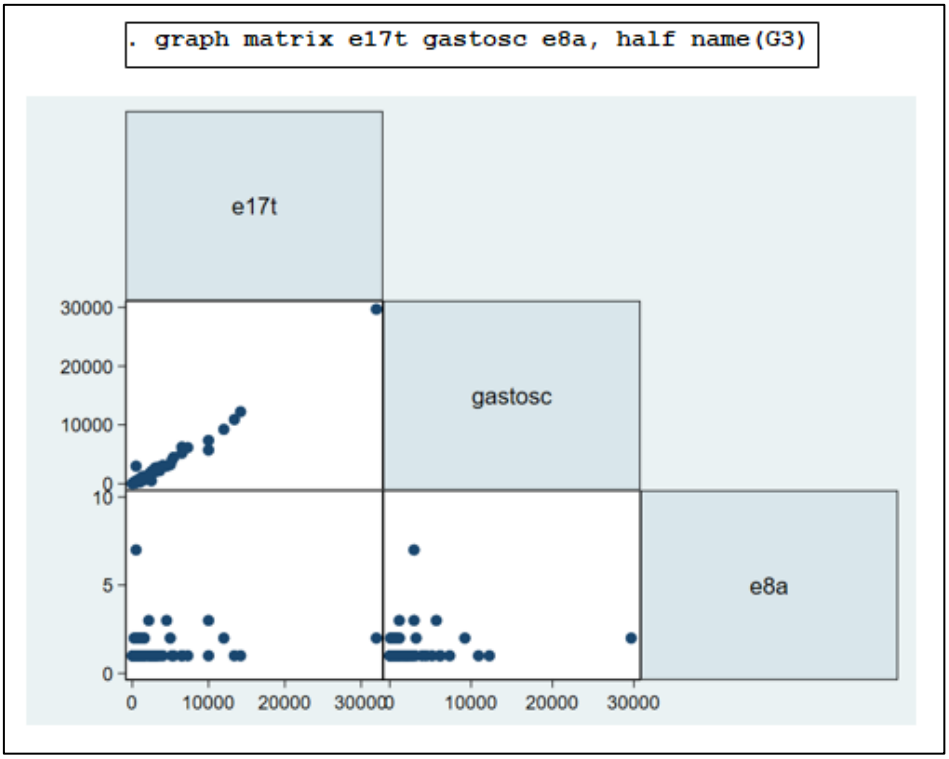


Figura 3.84. Grafica matricial entre las regresoras del modelo.

La gráfica muestra que las variables **ingresos** (***e17t***) y **gastos** (***gastosc***) tienen una correlación positiva, mientras que las demás gráficas de dispersión muestran una concentración.

Un aspecto en común en las tres gráficas es la existencia de algunos datos atípicos, lo que hace sospechar que el modelo de regresión puede presentar heterocedasticidad. Ahora aplicaremos la regla de Klein y el $R^2\ de\ Theil$.

Para aplicar la regla de Klein seleccionaremos a la variable gastos (***gastosc***) como la variable dependiente y al resto de regresoras como las variables independientes del siguiente modelo auxiliar.

$$C_i = \alpha_1 + \alpha_2 I_i + \alpha_3 N_i + e_i \text{ (3.7.38.)}$$

```
. reg gastosc e8a e17t

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(2, 83)        =   1852.06
       Model |  1.2322e+09     2   616089287   Prob > F        =    0.0000
    Residual |  27609997.6    83  332650.573   R-squared       =    0.9781
-------------+------------------------------   Adj R-squared   =    0.9776
       Total |  1.2598e+09    85    14821042   Root MSE        =    576.76

------------------------------------------------------------------------------
     gastosc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e8a |    174.1378   79.18576     2.20   0.031     16.64055    331.6351
       e17t |    .8664437   .0143958    60.19   0.000     .8378109    .8950764
      _cons |    -452.731   121.3812    -3.73   0.000    -694.1534   -211.3086
------------------------------------------------------------------------------
```

Figura 3.85. Regresión auxiliar para la regla de Klein.

El coeficiente de determinación de la regresión auxiliar (3.7.38.) es mayor al coeficiente de determinación del modelo original, por lo tanto, la multicolinealidad si está presente mediante esta variable. Ahora calcularemos el $R^2\ de\ Theil$ con las siguientes regresiones auxiliares.

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 C_i + v_i \text{ (3.7.39.)}$$

$$G_i = \theta_1 + \theta_2 C_i + \theta_3 N_i + v_i \text{ (3.7.40.)}$$

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 N_i + v_i \text{ (3.7.41.)}$$

```
. reg e25t3 e17t gastosc

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(2, 83)        =     21.75
       Model |    28444131     2  14222065.5   Prob > F        =    0.0000
    Residual |   54284702.3    83  654032.557   R-squared       =    0.3438
-------------+------------------------------   Adj R-squared   =    0.3280
       Total |   82728833.2    85  973280.391   Root MSE        =    808.72

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |   .3554823   .1317008     2.70   0.008     .0935346    .6174299
    gastosc |  -.2620203   .1496129    -1.75   0.084    -.5595944    .0355538
      _cons |   390.3489   106.0057     3.68   0.000     179.5078      601.19
------------------------------------------------------------------------------
```

Figura 3.86. Resultados de la regresión auxiliar (3.7.39.).

```
. reg e25t3 gastosc e8a

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(2, 83)        =     57.03
       Model |   47884358.7     2  23942179.4   Prob > F        =    0.0000
    Residual |   34844474.5    83  419812.946   R-squared       =    0.5788
-------------+------------------------------   Adj R-squared   =    0.5687
       Total |   82728833.2    85  973280.391   Root MSE        =    647.93

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gastosc |   .1165259   .0184549     6.31   0.000     .0798199    .1532318
        e8a |   678.5267   89.35951     7.59   0.000     500.7943    856.2592
      _cons |  -332.2723   134.9519    -2.46   0.016    -600.6862   -63.85831
------------------------------------------------------------------------------
```

Figura 3.87. Resultados de la regresión auxiliar (3.7.40.).

```
. reg e25t3 e17t e8a

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(2, 83)        =     69.43
       Model |   51778628.5     2  25889314.3   Prob > F        =    0.0000
    Residual |   30950204.7    83  372894.033   R-squared       =    0.6259
-------------+------------------------------   Adj R-squared   =    0.6169
       Total |   82728833.2    85  973280.391   Root MSE        =    610.65

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |    .113372   .0152418     7.44   0.000     .0830568    .1436873
        e8a |   691.1313   83.83892     8.24   0.000      524.379    857.8835
      _cons |  -405.2795   128.5139    -3.15   0.002    -660.8885   -149.6705
------------------------------------------------------------------------------
```

Figura 3.88. Resultados de la regresión auxiliar (3.7.41.).

Con estas regresiones auxiliares podremos calcular el R^2 *de Theil* siguiendo la siguiente fórmula.

$R^2\ de\ Theil = 0.7122 - [(0.7122 - 0.3438) + (0.7122 - 0.5788) + (0.7122 - 0.6259)] = 0.5635$ (3.7.42.)

El coeficiente de determinación de Theil contradice a todas las anteriores pruebas realizadas hasta el momento, debido a que su valor es 0.5635 y se puede inferir que el problema de multicolinealidad no es tan grave como aparenta.

Finalmente, se realizará una prueba *F* para comprobar la existencia de multicolinealidad en el modelo original. Para realizar este contraste se utilizará el siguiente modelo auxiliar: $C_i = \alpha_1 + \alpha_2 I_i + \alpha_3 N_i + e_i$, al cual se le planteará el siguiente contraste de hipótesis.

$H_0: No\ existe\ multicolinealidad$ (3.7.34.)

$$H_a: Existe\ multicolinealidad$$

El estadístico *F* calculado se consigue mediante.

$Fc = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)} = \frac{0.9781/(3-2)}{(1-0.9781)/(86-3+1)} = 3751.62$ (3.7.43.)

Mientras, el estadístico *F* tabulado es: $Ft^1_{84,0.05} = 3.95$. Al concluir que $|Fc| > Ft^1_{84,0.05}$ entonces rechazamos la hipótesis nula y aceptamos la hipótesis alternativa en la que se asume que efectivamente hay multicolinealidad en el modelo original provocado por la variable **gastos**($\boldsymbol{gastosc}$).

Tomando en cuenta la matriz de correlación, la matriz de gráficas de correlación, los índices VIF y TOL, la regla de Klein y el $R^2\ de\ Theil$, se puede concluir que los estimadores de las variables **gastos**($\boldsymbol{gastosc}$) **e ingresos**($\boldsymbol{e17t}$) del modelo original pueden estar influenciadas por la elevada correlación existente entre estas variables, estas sospechas se hacen más sólidas cuando al revisar los índices de VIF de cada variable, notamos que son mayores a 30, por lo tanto, si se requiere transformar o descartar variables estas serán consideradas. Algo a notar es que, pese a que las significancias individuales de los estimadores de estas variables pueden estar afectados el coeficiente de determinación del modelo original no se ha disparado como normalmente sucede en los modelos con multicolinealidad elevada.

La primera medida correctiva que se puede ejecutar para reducir la multicolinealidad en el modelo es aumentar el tamaño muestral, sin embargo al igual que

en la anterior regresión, no existen más observaciones sin datos vacíos en la ENAHO para aumentar el tamaño muestral. Otra medida correctiva que se puede ejecutar es aplicar una restricción a los estimadores, el problema con aplicar este método es la inexistencia de una restricción por parte del marco teórico, entonces este método no puede aplicarse. Otra opción posible sería descartar las variables que causan multicolinealidad, para ello especificaremos los siguientes modelos auxiliares.

$$G_i = \theta_1 + \theta_2 C_i + \theta_3 N_i + v_i \text{ (3.7.40.)}$$

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 N_i + v_i \text{ (3.7.41.)}$$

$G_i = \theta_1 + \theta_2 N_i + v_i$ (3.7.44.)

```
. reg e25t3 gastosc e8a

      Source |       SS       df       MS              Number of obs =      86
-------------+------------------------------           F(2, 83)      =   57.03
       Model |  47884358.7     2  23942179.4            Prob > F      =  0.0000
    Residual |  34844474.5    83  419812.946            R-squared     =  0.5788
-------------+------------------------------           Adj R-squared =  0.5687
       Total |  82728833.2    85  973280.391            Root MSE      =  647.93

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gastosc |   .1165259   .0184549     6.31   0.000     .0798199    .1532318
        e8a |   678.5267   89.35951     7.59   0.000     500.7943    856.2592
      _cons |  -332.2723   134.9519    -2.46   0.016    -600.6862   -63.85831
------------------------------------------------------------------------------
```

Figura 3.87. Resultados de la regresión auxiliar (3.7.40.).

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
         e8a |      1.02    0.978446
     gastosc |      1.02    0.978446
-------------+----------------------
    Mean VIF |      1.02
```

Figura 3.89. Índices de VIF y TOL de la regresión auxiliar (3.7.40.).

```
. reg e25t3 e17t e8a

      Source |       SS       df       MS              Number of obs =      86
-------------+------------------------------           F(2, 83)      =   69.43
       Model |  51778628.5     2  25889314.3            Prob > F      =  0.0000
    Residual |  30950204.7    83  372894.033            R-squared     =  0.6259
-------------+------------------------------           Adj R-squared =  0.6169
       Total |  82728833.2    85  973280.391            Root MSE      =  610.65

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |    .113372   .0152418     7.44   0.000     .0830568    .1436873
        e8a |   691.1313   83.83892     8.24   0.000      524.379    857.8835
      _cons |  -405.2795   128.5139    -3.15   0.002    -660.8885   -149.6705
------------------------------------------------------------------------------
```

Figura 3.88. Resultados de la regresión auxiliar (3.7.41.).

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
        e17t |      1.01    0.987317
         e8a |      1.01    0.987317
-------------+----------------------
    Mean VIF |      1.01
```

Figura 3.90. Índices de VIF y TOL de la regresión auxiliar (3.7.41.).

```
. reg e25t3 e8a

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(1, 84)        =     50.72
       Model |  31147303.5     1  31147303.5   Prob > F        =    0.0000
    Residual |  51581529.7    84   614065.83   R-squared       =    0.3765
-------------+------------------------------   Adj R-squared   =    0.3691
       Total |  82728833.2    85  973280.391   Root MSE        =    783.62

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e8a |   761.3625   106.9027     7.12   0.000     548.7747    973.9502
      _cons |  -220.2469   161.7978    -1.36   0.177    -541.9996    101.5057
------------------------------------------------------------------------------
```

Figura 3.91. Índices de VIF y TOL de la regresión auxiliar (3.7.44.).

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
         e8a |      1.00    1.000000
-------------+----------------------
    Mean VIF |      1.00
```

Figura 3.92. Índices de VIF y TOL de la regresión auxiliar (3.7.44.).

Estos modelos donde se descartan una o ambas variables causantes de multicolinealidad, muestran índices de VIF y TOL que aparentemente la multicolinealidad ha sido corregida, pero es necesario revisar detalladamente cada modelo. En cuanto al coeficiente de determinación podemos notar que el modelo (3.7.41.) es el mayor con respecto a los demás modelos especificados. En cuanto a la significancia global, los tres modelos presentan significancia global debido a que sus respectivos valores-p son menores al 5% de significancia. Por otro lado en cuanto a los signos esperados de los estimadores solamente los modelos (3.7.41.) y (3.7.44.) tienen estimadores que cumplen con sus signos esperados; esto nos puede indicar que el modelo (3.7.40.) podría ser descartado y en cuanto a sus significancias individuales de los estimadores, estos tienen significancia individual en sus modelos.

Estos datos nos llevan a la conclusión que el modelo (3.7.41.) puede ser la mejor opción para solucionar el problema de multicolinealidad, sin embargo este podría tener un sesgo de especificación ocasionando heterocedasticidad en el modelo auxiliar, por lo tanto se debe comprobar si efectivamente existe y de ser así entonces se procederá a determinar su corrección. La primera prueba que se contrastará es la prueba BP con el comando **estat hettest** y su opción **mtest**.

$$H_0: No\ existe\ heterocedasticidad$$

$$H_1: Existe\ heterocedasticidad$$

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of e25t3

         chi2(1)      =    78.98
         Prob > chi2  =   0.0000

. estat hettest e17t e8a, mtest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

-------------------------------------------------
    Variable |      chi2   df      p
-------------+-----------------------------------
        e17t |     90.55    1   0.0000 #
         e8a |     16.29    1   0.0001 #
-------------+-----------------------------------
simultaneous |     99.45    2   0.0000
-------------------------------------------------
                # unadjusted p-values
```

Figura 3.93. Prueba de BP de la regresión auxiliar (3.7.44.).

En la figura que muestra la prueba de BP para determinar la existencia de heterocedasticidad en el modelo auxiliar. Podemos determinar que efectivamente existe heterocedasticidad, y al igual que en el anterior modelo de regresión para los trabajadores independientes dedicados a actividades productivas/extractivas, el modelo no tiene un esquema de varianza conocido.

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     42.67
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	42.67	5	0.0000
Skewness	31.14	2	0.0000
Kurtosis	3.66	1	0.0558
Total	77.47	8	0.0000

Figura 3.94. Prueba general de Wwhite de heterocedasticidad de la regresión auxiliar (3.7.44.).

La prueba general de White nos permite llegar a la misma conclusión que hemos obtenido en la prueba BP. Rechazar la hipótesis nula puede ser ocasionado por la existencia de un sesgo de especificación por ello se realizara la prueba de White pura.

```
. predict um,resid

. gen umc=um^2

. gen x2c=e17t^2

. gen x3c=e8a^2
```

Figura 3.95. Generando regresores para la prueba pura de heterocedasticidad de White de la regresión auxiliar (3.7.44.).

```
. reg umc e17t e8a x2c x3c
```

Source	SS	df	MS	Number of obs	=	86
				F(4, 81)	=	19.69
Model	2.9432e+13	4	7.3579e+12	Prob > F	=	0.0000
Residual	3.0267e+13	81	3.7366e+11	R-squared	=	0.4930
				Adj R-squared	=	0.4680
Total	5.9698e+13	85	7.0233e+11	Root MSE	=	6.1e+05

umc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e17t	18.28551	35.034	0.52	0.603	-51.42116	87.99218
e8a	419972.8	254061.8	1.65	0.102	-85530.51	925476.1
x2c	.004001	.0013258	3.02	0.003	.0013631	.0066389
x3c	-33404.49	37255.86	-0.90	0.373	-107532	40722.99
_cons	-249410	254395.5	-0.98	0.330	-755577.1	256757.2

Figura 3.96. Prueba pura de heterocedasticidad de White de la regresión auxiliar (3.7.44.).

Con estos resultados podemos calcular el siguiente estadístico calculado.

$n * R_i^2 = 86 * 0.4930 = 42.40$ (3.7.45.)

Por otro lado, el estadístico tabulado es: $X^2_{4,0.05} = 9.49$, entonces al concluir que el estadístico calculado es mayor al estadístico tabulado, se infiere que según la prueba pura de heterocedasticidad de White el modelo auxiliar tiene heterocedasticidad.

Debido a que a la varianza heterocedástica del modelo auxiliar tiene un esquema desconocido, se procede a corregirla mediante los errores estándares robustos de White.

```
. reg e25t3 e17t e8a, robust

Linear regression                               Number of obs     =         86
                                                F(2, 83)          =      17.23
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6259
                                                Root MSE          =     610.65

------------------------------------------------------------------------------
             |               Robust
      e25t3  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t  |    .113372   .044726     2.53   0.013     .0244138    .2023303
        e8a  |    691.1313   132.546     5.21   0.000     427.5026     954.76
      _cons  |   -405.2795   176.1206   -2.30   0.024    -755.5764   -54.98273
------------------------------------------------------------------------------
```

Figura 3.97. Errores estándares robustos de la regresión auxiliar (3.7.44.).

A continuación, se presentarán las dos regresiones del modelo (3.7.41.) para corregir la multicolinealidad donde la primera regresión tiene errores hallados mediante MCO y la segunda tiene errores robustos.

$$G_i = -405.28_1 + 0.11_2 I_i + 691.13_3 N_i + v_i \text{ (3.7.41.)}$$

$$ee = (128.51) \quad (0.02) \quad (83.84)$$

$$t = -3.15 \qquad 7.44 \qquad 8.24$$

$$G_i = -405.28_1 + 0.11_2 I_i + 691.13_3 N_i + v_i \text{ (3.7.46.)}$$

$$ee = (176.12) \quad (0.04) \quad (176.12)$$

$$t = -2.30 \qquad 2.53 \qquad 5.21$$

En la regresión con errores robustos se ha logrado conservar los signos esperados de cada estimador, así como sus respectivas significancias individuales y su significancia global. Por lo tanto, se prefiere usar el modelo (3.7.46.) para corregir la multicolinealidad.

Pese a que el modelo (3.7.46.) puede ser una excelente opción para tratar la multicolinealidad en el modelo, podemos caer en un sesgo de especificación por subajuste debido a que el marco teórico no concibe el descarte de esta variable. Por tal motivo, se pondría en duda si el modelo (3.7.46.) corrige la multicolinealidad sin que conlleve a generar sesgos de especificación.

Finalmente, otra opción para solucionar el problema de la multicolinealidad en el modelo original es intentar transformar las variables. Sin embargo, al ser dos variables las causantes de este problema, la multicolinealidad en el modelo se hace compleja y no se lograría corregir oportunamente el modelo. Siguiendo la teoría que han propuesto (Gujarati & Porter, 2010), se ha optado por no plantearse algún método correctivo para la multicolinealidad.

- **Homocedasticidad.**

En esta sección se probará si el modelo cumple con el supuesto de homocedasticidad y se le pretenderá corregir, en caso se demuestre que el modelo especificado no cumple dicho supuesto. Para demostrar si el modelo cumple con el supuesto de homocedasticidad se hará uso de los métodos informales y formales. A continuación, se mostrarán los resultados obtenidos del modelo original y posteriormente las gráficas que se logran conseguir entre los residuos estimados y las variables del modelo.

```
. reg e25t3 e17t gastosc e8a
```

Source	SS	df	MS	Number of obs	=	86
				F(3, 82)	=	67.64
Model	58918846.4	3	19639615.5	Prob > F	=	0.0000
Residual	23809986.8	82	290365.692	R-squared	=	0.7122
				Adj R-squared	=	0.7017
Total	82728833.2	85	973280.391	Root MSE	=	538.86

e25t3	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e17t	.5539907	.0898667	6.16	0.000	.3752171	.7327642
gastosc	-.508537	.1025509	-4.96	0.000	-.7125434	-.3045306
e8a	779.6868	76.10672	10.24	0.000	628.2863	931.0873
_cons	-635.51	122.5403	-5.19	0.000	-879.2817	-391.7383

Figura 3.49. Regresión para los trabajadores independientes que se han dedicado a actividades comerciales.

Ahora se mostrarán gráficos de dispersión entre los residuos de este modelo y los valores estimados de la variable dependiente.

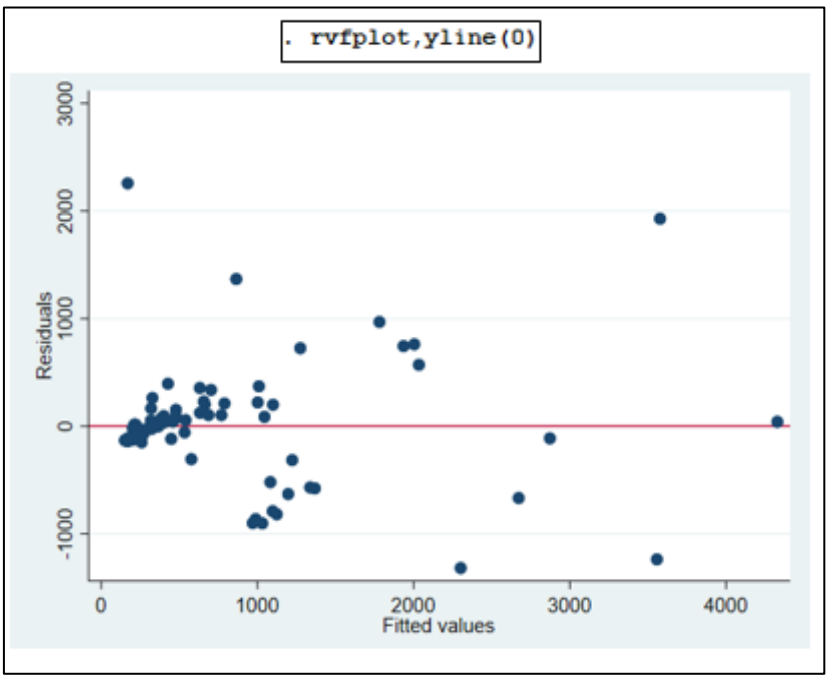


Figura 3.98. Grafica de dispersión entre los residuos y la variable dependiente estimada del modelo original.

En la figura que muestra una gráfica de dispersión entre los residuos y los valores estimados de la variable dependiente, del modelo sobre los trabajadores independientes dedicados a la actividad comercial, se puede apreciar una concentración en la esquina inferior izquierda y algunos puntos alejados de esta concentración, por lo tanto, al existir datos atípicos ya se puede tener sospechas que existe heterocedasticidad en el modelo.

A continuación, se muestran las gráficas de dispersión entre los residuos y las variables explicativas del modelo original.

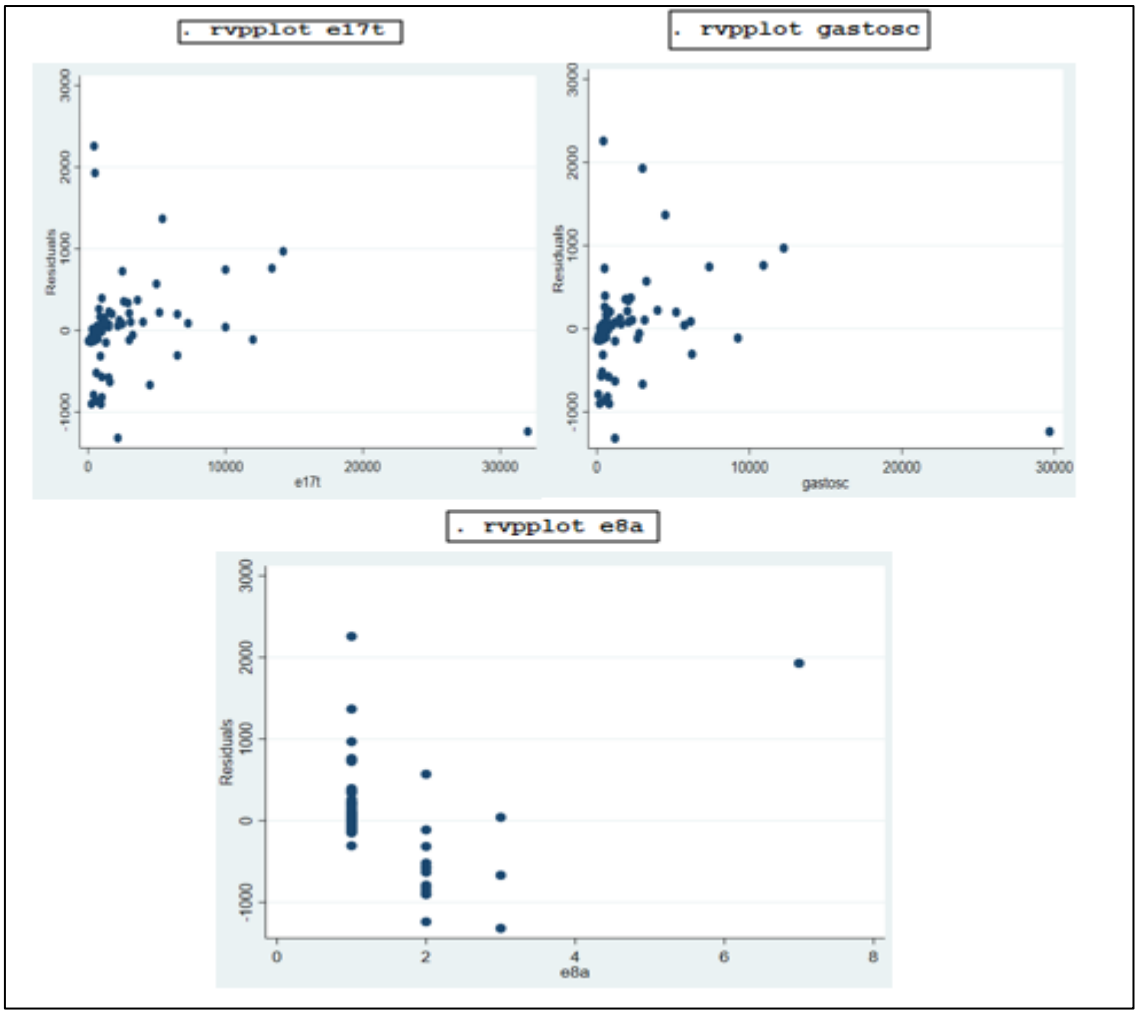


Figura 3.99. Grafica de dispersión entre los residuos y las variables explicativas del modelo original.

En los tres gráficos se puede observar una similitud, una concentración en la esquina inferior izquierda y algunos datos alejados. Por lo tanto, luego de revisar estas gráficas se puede sospechar que el modelo tiene heterocedasticidad y ya que en los tres gráficos que muestran una dispersión entre los residuos y las regresoras, posiblemente el esquema de la varianza del término de perturbación sea desconocida.

Para tener completa seguridad que el modelo efectivamente no cumple con el supuesto de heterocedasticidad, se complementará la información recibida de las gráficas anteriores con las pruebas formales mediante contraste de hipótesis, donde la prueba de hipótesis es.

$$H_0: No\ existe\ heterocedasticidad.$$

$$H_1: Existe\ heterocedasticidad.$$

Y se hará uso de las pruebas BP y de White. A continuación se muestran los resultados sobre la prueba BP.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of e25t3

         chi2(1)      =    43.96
         Prob > chi2  =   0.0000

. estat hettest e17t gastosc e8a,mtest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

-------------------------------------------
    Variable |      chi2   df      p
-------------+-----------------------------
        e17t |      9.27    1   0.0023 #
     gastosc |     15.20    1   0.0001 #
         e8a |     86.76    1   0.0000 #
-------------+-----------------------------
simultaneous |    105.42    3   0.0000
-------------------------------------------
                  # unadjusted p-values
```

Figura 3.100. Prueba general de BP de heterocedasticidad del modelo original.

Los resultados de la prueba BP nos permiten concluir que el modelo efectivamente contiene una varianza del término de perturbación heterocedástica, además, cabe recalcar que la prueba de BP para cada regresora señala que las tres variables son causantes de heterocedasticidad y entre las tres regresoras la variable **Número de trabajadores**($\boldsymbol{e8a}$) es posiblemente la mayor causante de heterocedasticidad en el modelo original.

Ahora veremos si con la prueba general de White de heterocedasticidad también se concluye que el modelo tiene heterocedasticidad.

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     32.18
         Prob > chi2  =    0.0002

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df       p
---------------------+-----------------------------
Heteroskedasticity |      32.18      9    0.0002
            Skewness |      18.34      3    0.0004
            Kurtosis |       3.08      1    0.0793
---------------------+-----------------------------
               Total |      53.60     13    0.0000
---------------------------------------------------
```

Figura 3.101. Prueba general de White de heterocedasticidad del modelo original.

Según la prueba general de White de heterocedasticidad, el modelo no cumple el supuesto de homocedasticidad y por lo tanto se debería ejecutar un método correctivo. Finalmente, se realizará la prueba de heterocedasticidad pura de White para corroborar la prueba general.

```
. predict uw,resid

. gen uwc=uw^2

. gen e17tc=e17t^2

. gen gastoscc=gastosc^2

. gen e8ac=e8a^2
```

```
. reg uwc e17t gastosc e8a e17tc gastoscc e8ac

      Source |       SS       df       MS              Number of obs =      86
-------------+------------------------------           F(6, 79)      =    6.97
       Model |  1.6437e+13     6  2.7396e+12           Prob > F      =  0.0000
    Residual |  3.1031e+13    79  3.9280e+11           R-squared     =  0.3463
-------------+------------------------------           Adj R-squared =  0.2966
       Total |  4.7469e+13    85  5.5846e+11           Root MSE      =  6.3e+05

------------------------------------------------------------------------------
         uwc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e17t |  -163.2519   241.3157    -0.68   0.501    -643.5788    317.075
    gastosc |   247.3555   263.7834     0.94   0.351    -277.6923   772.4033
        e8a |   287090.1   314503.1     0.91   0.364    -338912.7   913092.9
      e17tc |  -.0052127   .0167341    -0.31   0.756    -.0385211   .0280956
   gastoscc |   .0048947   .0184194     0.27   0.791    -.0317681   .0415576
       e8ac |   24437.68   51117.46     0.48   0.634    -77309.08   126184.4
      _cons |  -177069.7   297952.9    -0.59   0.554    -770130.2   415990.7
------------------------------------------------------------------------------
```

Figura 3.102. Prueba de heterocedasticidad pura de White del modelo original.

Con la figura anterior se calculan el siguiente estadístico calculado.

$n * R_i^2 = 86 * 0.3463 = 31.50$ (3.7.47.)

Y se usa el siguiente estadístico tabulado.

$X^2_{0.05,6} = 12.6$ (3.7.48.)

Entonces, al tener el estadístico calculado mayor al estadístico tabulado se puede rechazar la hipótesis nula y aceptar que el modelo no cumple el supuesto de homocedasticidad.

Ahora intentaremos aplicar los métodos correctivos apropiados para corregir la heterocedasticidad en el modelo. Aunque las tres regresoras pueden ser las causas de la heterocedasticidad en el modelo, la prueba de BP indica que la variable **Número de trabajadores**$(e8a)$ puede ser la mayor causante de heterocedasticidad en el modelo original, por ello aplicaremos MCP y MCF.

Para aplicar el MCP utilizaremos el componente [$weight$] del comando **reg**. Esta es la estructura de la sintaxis.

```
. reg e25t3 e17t gastosc e8a [aw=1/e8a]
(sum of wgt is 76.6428571429)

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(3, 82)        =     70.75
       Model |  37980958.6     3  12660319.5   Prob > F        =    0.0000
    Residual |  14672962.7    82  178938.569   R-squared       =    0.7213
-------------+------------------------------   Adj R-squared   =    0.7111
       Total |  52653921.3    85  619457.898   Root MSE        =    423.01

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |   .7969803   .0934965     8.52   0.000      .610986    .9829747
    gastosc |  -.7648209   .1071645    -7.14   0.000    -.9780052   -.5516366
        e8a |   481.4363   106.2112     4.53   0.000     270.1483    692.7243
      _cons |  -359.8885    128.776    -2.79   0.006     -616.065   -103.7119
------------------------------------------------------------------------------
```

Figura 3.103. Regresión del modelo original mediante Mínimos Cuadrados Ponderados.

Los resultados de la regresión que se muestra en la figura 3.103. Corresponden al siguiente modelo.

$$\left(\frac{G_i}{\sqrt{N_i}}\right) = \beta_1\left(\frac{1}{\sqrt{N_i}}\right) + \beta_2\left(\frac{I_i}{\sqrt{N_i}}\right) + \beta_3\left(\frac{C_i}{\sqrt{N_i}}\right) + \beta_4\sqrt{N_i} + \left(\frac{\mu_1}{\sqrt{N_i}}\right) \text{ (3.7.49.)}$$

Donde el modelo (3.7.49.) supone que la varianza del término de error corresponde al siguiente esquema: $E(\mu_i^2) = \sigma^2 N_i$, es decir que la varianza del termino de error depende de la variable N_i la cual corresponde a la variable **Números de trabajadores** ($\boldsymbol{e8a}$). Recuerde que los resultados del modelo estimado mediante MCP no se interpretan sino se reemplazan en el modelo original. Los resultados se expresan de la siguiente forma.

$$G_i = -359.89 + 0.80I_i - 0.76C_i + 481.43N_i + \mu_i \quad (3.7.50.)$$

$$ee = (128.78)\ \ (0.09)\ \ (0.11)\ \ (106.21)$$

$$t = -2.79 \quad 8.52 \quad -7.14 \quad 4.53$$

En la figura 3.103. Se puede ver que el componente [$\boldsymbol{weight}$] es el componente que le da a STATA la instrucción de efectuar una regresión con ponderaciones. En STATA existen cuatro posibles tipos de ponderaciones las cuales son: ponderaciones en frecuencias ($\boldsymbol{fweight}$), poblacional ($\boldsymbol{pweight}$), analítica ($\boldsymbol{aweight}$) y específica ($\boldsymbol{iweight}$), con estas especificaciones es fácil intuir que la regresión de MCP trabaja con una ponderación analítica. La regresión mediante MCP cuyos resultados se observan en el modelo (3.7.50.) mantiene los signos esperados de los estimadores, sus respectivas significancias individuales y su significancia global; y el coeficiente de determinación ha aumentado ligeramente, en consecuencia el error estándar de regresión, representado como $Root\ MSE = 423.01$, es menor al error estándar de regresión del modelo original. Algunos autores sugieren usar los errores robustos en los MCP, ya que supone estimadores más eficientes.

```
. reg e25t3 e17t gastosc e8a [aw=1/e8a], robust
(sum of wgt is 76.6428571429)

Linear regression                               Number of obs     =         86
                                                F(3, 82)          =      87.17
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7213
                                                Root MSE          =     423.01

------------------------------------------------------------------------------
             |               Robust
      e25t3  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e17t |   .7969803   .1466932     5.43   0.000     .5051608    1.0888
     gastosc |  -.7648209   .1709821    -4.47   0.000    -1.104959   -.424683
         e8a |   481.4363   278.8511     1.73   0.088    -73.28726    1036.16
       _cons |  -359.8885   270.9469    -1.33   0.188    -898.8881    179.1112
------------------------------------------------------------------------------
```

Figura 3.104. Regresión del modelo original mediante Mínimos Cuadrados Ponderados y errores robustos.

Y estos son sus resultados en el modelo especificado (3.7.50.) con errores robustos.

$G_i = -359.89 + 0.80I_i - 0.76C_i + 481.43N_i + \mu_i$ (3.7.51.)

$$ee = (270.95)\ \ (0.15)\ \ (0.17)\ \ (278.86)$$

$$t = -1.33 \qquad 5.43 \ \ -4.47 \quad 1.73$$

Los tres modelos mantienen estimadores que cumplen con sus signos esperados y tienen significancia individual, a excepción del modelo (3.7.51.) donde el estimador de la variable **Número de trabajadores** ($\boldsymbol{e8a}$). Por último, el modelo (3.7.51.) también muestra significancia global.

Veamos ahora cuál es el procedimiento para corregir la heterocedasticidad mediante **Mínimos Cuadrados Factibles**. Cabe recalcar que este método también comprende una extensión de los Mínimos Cuadrados Generalizados y una definición muy simple de este método es que consiste en *estimar* la varianza desconocida del término de perturbación utilizando la regresora del modelo que posiblemente esté causando la heterocedasticidad. La diferencia con los MCP, consiste en que estos últimos intentan acercarse a la varianza del término de error a través de una regresora.

(Adkins C. & Carter H., 2011) Explican que se debe tomar a la regresora que consideramos que es la culpable de causar heterocedasticidad y especificar una relación funcional entre la regresora y la varianza del termino de error. La función más común es la exponencial, la cual se especifica como:

$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_{i2} + \alpha_3 z_{i3} + \cdots + \alpha_k z_{ik})$ (3.7.52.)

Y dada la teoría expuesta anteriormente, z_{ik} son las regresoras del modelo original y α_1 son los parámetros. Este método correctivo toma el logaritmo natural del término de perturbación y lo sustituye en la varianza desconocida y le agrega un término de error diferente al modelo original. Por ejemplo, al asumir que z_{i2} es posiblemente la causante de heterocedasticidad entonces tenemos.

$\ln(\hat{\mu}_i^2) = \ln(\sigma_i^2) + e_i = \alpha_1 + \alpha_2 z_{i2} + e_i$ (3.7.53.)

(Adkins C. & Carter H., 2011) Determinan que $\hat{\mu}_i^2$ representa los valores del término de perturbación al cuadrado del modelo original, en este caso de (3.7.11.) cuyos

resultados se visualizan en la figura 3.49; e_i es el nuevo término de error en el modelo (3.7.53.) que representa ser el modelo de heterocedasticidad del modelo original.

Debido a que el método correctivo mediante MCF se basa en que el modelo (3.7.53.), debe estar correctamente especificado, por ello algunos autores como (Colin C. & Trivedi, 2009) Señalan que se puede combinar el MCF con los errores robustos con el fin de evitar que el modelo arroje resultados afectados por la mala especificación. Para lograrlo en STATA se debería introducir la opción **robust**.

A continuación se presentara cual es el método para corregir mediante MCF el modelo especificado sobre los trabajadores independientes que se han dedicado a la actividad comercial.

Empezamos especificando el modelo de heterocedasticidad que se usara para corregir el modelo (3.7.11), la cual es:

$\ln(\hat{\mu}_i^2) = \alpha_1 + \alpha_2 N_{i2} + e_i$ (3.7.54.)

Una vez especificado el modelo debemos calcular el logaritmo natural de la regresora que asumimos causa heterocedasticidad, en este caso de $\boldsymbol{e8a}$.

```
. gen z=ln(e8a)
```

Figura 3.105. Calculando el logaritmo natural de la variable $\boldsymbol{e8a}$.

Ahora realizamos la regresión del modelo original y posteriormente hallaremos los valores de sus respectivos residuos con el nombre ***ehat***.

```
. reg e25t3 e17t gastosc e8a
. predict ehat,resid
```

Figura 3.106. Calculando los residuos del modelo original.

Ahora generamos el logaritmo de los errores al cuadrado.

```
. gen ln_ehat_c=ln(ehat*ehat)
```

Figura 3.107. Calculando el logaritmo natural de los residuos al cuadrado del modelo original.

Ahora estimamos el modelo de heterocedasticidad (3.7.54.) utilizando las variables generadas ***ln_ehat_c*** y ***z***, las cuales son el logaritmo natural de los residuos del modelo al cuadrado y el logaritmo de la variable ***e8a*** respectivamente.

```
. reg ln_ehat_c z

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(1, 84)        =     24.36
       Model |  156.633711     1  156.633711   Prob > F        =    0.0000
    Residual |  540.209003    84  6.43105956   R-squared       =    0.2248
-------------+------------------------------   Adj R-squared   =    0.2155
       Total |  696.842714    85  8.19814958   Root MSE        =     2.536

------------------------------------------------------------------------------
    ln_ehat_c |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
            z |   3.724535   .7546936     4.94   0.000     2.223744    5.225326
        _cons |   9.262148   .3007048    30.80   0.000     8.664163    9.860132
------------------------------------------------------------------------------
```

Figura 3.108. Regresión entre el logaritmo de los residuos al cuadrado del modelo original y el logaritmo de *e8a*.

(Adkins C. & Carter H., 2011) Señalan que para obtener los estimadores de MCF se necesitan calcular el antilogaritmo de los valores estimados de la variable dependiente en el modelo (3.7.54.). Estos valores estimados, también llamados predichos o ajustados, se calculan con el comando **predict** y la opción **xb**. Posteriormente, se utilizará el comando **gen** para crear esa variable anti logarítmica.

```
. predict lnsig2, xb

. gen wt=exp(lnsig)
```

Figura 3.109. Obteniendo la variable que se usara para realizar la regresión con ponderaciones.

```
. reg e25t3 e17t gastosc e8a [aw=1/wt]
(sum of wgt is .006649835544)

      Source |       SS       df       MS      Number of obs   =        86
-------------+------------------------------   F(3, 82)        =     90.91
       Model |  31278868.9     3  10426289.6   Prob > F        =    0.0000
    Residual |  9404242.16    82   114685.88   R-squared       =    0.7688
-------------+------------------------------   Adj R-squared   =    0.7604
       Total |  40683111.1    85  478624.836   Root MSE        =    338.65

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       e17t |   .9139809   .0981305     9.31   0.000     .7187681    1.109194
    gastosc |  -.8739711   .1158369    -7.54   0.000    -1.104408   -.6435347
        e8a |   21.19958   281.3015     0.08   0.940    -538.3986    580.7978
      _cons |   70.16705   287.0261     0.24   0.807    -500.8193    641.1534
------------------------------------------------------------------------------
```

Figura 3.110. Resultados de la regresión original hallados mediante MCF.

Finalmente, la variable generada ***wt*** se utilizará para los estimadores de la regresión mediante MCF.

$$G_i = 70.17 + 0.92I_i - 0.87C_i + 21.19N_i + \mu_i \text{ (3.7.55.)}$$

$$ee = (287.03)\ (0.10)\ (0.12)\ (281.30)$$

$$t = 0.24 \quad 9.31 \quad -7.54 \quad 0.08$$

El modelo (3.7.55.) muestra resultados hallados mediante el método de los MCF, los estimadores en este modelo mantienen sus respectivos signos esperados y son significativos para el modelo, a excepción de la variable ***e8a*** ya que su valor-p es mayor a la significancia del 5%, por lo que se asume que la variable no es significativa. El modelo también mantiene la significancia global del modelo y un cambio positivo frente a los resultados de los modelos originales y el modelo corregido mediante MCP es que el coeficiente de determinación del modelo hallado mediante MCF es $R^2_{MCF} = 0.7688$, por lo que tiene una mejor bondad de ajuste que los modelos anteriores. Finalmente, recordemos que (Colin C. & Trivedi, 2009) Recomiendan usar los errores robustos de White en el modelo (3.7.55.) con la opción **robust**.

```
. reg e25t3 e17t gastosc e8a [aw=1/wt],robust
(sum of wgt is .006649835544)

Linear regression                               Number of obs     =         86
                                                F(3, 82)          =      93.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7688
                                                Root MSE          =     338.65

------------------------------------------------------------------------------
             |               Robust
      e25t3  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e17t |   .9139809   .0876269    10.43   0.000      .739663    1.088299
     gastosc |  -.8739711   .1042109    -8.39   0.000     -1.08128   -.6666625
         e8a |   21.19958   100.8186     0.21   0.834    -179.3608     221.76
       _cons |   70.16705   127.8678     0.55   0.585    -184.2027    324.5368
------------------------------------------------------------------------------
```

Figura 3.111. Resultados de la regresión original hallados mediante MCF y errores robustos de White.

$$G_i = 70.17 + 0.92I_i - 0.87C_i + 21.19N_i + \mu_i \text{ (3.7.56.)}$$

$$ee = (127.87)\ (0.09)\ (0.10)\ (100.82)$$

$$t = 0.55 \quad 10.43 \quad -8.39 \quad 0.21$$

La teoría que presenta (Adkins C. & Carter H., 2011) Indican que el modelo (3.7.56.) está libre de heterocedasticidad y correctamente especificado. Y mantiene las mismas características que en el modelo (3.7.55.), ya que se puede ver que los estimadores de las variables ***e17t*** y ***gastosc*** son significativos individualmente porque sus respectivos valores-p son inferiores a la significancia del 5%. Por otro lado, el estimador de la variable ***e8a*** no tiene significancia individual porque su valor-p es mayor al 5% de significancia. Además, el modelo (3.7.56.) conserva la significancia global debido a que el valor-p del estadístico *F* calculado es menor a una significancia del 5%.

Finalmente, para corregir la heterocedasticidad en el modelo original también conviene tomar los errores robustos del modelo original.

```
. reg e25t3 e17t gastosc e8a, robust

Linear regression                               Number of obs     =         86
                                                F(3, 82)          =      56.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7122
                                                Root MSE          =     538.86

------------------------------------------------------------------------------
             |               Robust
      e25t3  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e17t |   .5539907   .1491688     3.71   0.000     .2572464    .8507349
     gastosc |   -.508537   .1772043    -2.87   0.005    -.8610527   -.1560213
         e8a |   779.6868   204.7257     3.81   0.000     372.4222    1186.951
       _cons |    -635.51   202.8463    -3.13   0.002    -1039.036   -231.9841
------------------------------------------------------------------------------
```

Figura 3.112. Resultados de la regresión original corregido de heterocedasticidad mediante los errores robustos de White.

$$G_i = -635.51 + 0.55I_i - 0.51C_i + 779.65N_i + \mu_i \text{ (3.7.57.)}$$

$$ee = (202.65)\ \ (0.15)\ \ (0.18)\ \ (204.65)$$

$$t = 0.55 \quad 10.43 \quad -8.39 \quad 0.21$$

En las siguientes tablas podemos ver la información resumida de todos los modelos planteados para corregir la heterocedasticidad.

Modelo especificado	Estimador	Ingreso (I_i)	Gastos (C_i)	Número de trabajadores(N_i)
MCO	$\hat{\beta}_k$	0.55	-0.51	779.69
(3.7.11.)	ee	(0.09)	(0.10)	(76.11)
	t	6.16	-4.96	10.24
MCG	$\hat{\beta}_k$	0.79	-0.76	481.44
(3.7.50)	ee	(0.10)	(0.11)	(106.21)
	t	8.52	-7.14	4.53
MCG con errores de White	$\hat{\beta}_k$	0.79	-0.76	481.44
	ee	(0.15)	(0.17)	(278.86)
(3.7.51.)	t	5.43	-4.47	1.73
MCF	$\hat{\beta}_k$	0.91	-0.87	21.20
(3.7.55.)	ee	(0.10)	(0.11)	(281.30)
	t	9.31	-7.54	0.08
MCF con errores de White	$\hat{\beta}_k$	0.91	-0.87	21.20
	ee	(0.09)	(0.10)	(100.82)
(3.7.56.)	t	10.43	-8.39	0.21
Errores de White	$\hat{\beta}_k$	0.55	-0.51	779.69
(3.7.57.)	ee	(0.15)	(0.18)	(204.73)
	t	3.71	-2.87	3.81

Tabla 3.19. Resultados de los modelos especificados para corregir la heterocedasticidad (1).

Estas tablas muestran los resultados de los distintos métodos correctivos aplicados al modelo original. Nos indican, que el mejor método para corregir la heterocedasticidad, son los estimadores de MCF con los errores robustos de White (3.7.5.) ya que conserva los signos esperados de los estimadores, tiene significancia global, una mejor bondad de ajuste tal como muestra su coeficiente de determinación comparado con los coeficientes de determinación de otros modelos y los estimadores tienen significancia individual a excepción del estimador que acompaña a la variable ***e8a***. Además, pese a que posiblemente este modelo esté influenciado por la presencia de multicolinealidad se ha

Tabla 3.20. Resultados de los modelos especificados para corregir la heterocedasticidad (2).

- **Actividad prestadora de servicios.**

Para terminar esta sección, comprobaremos si el modelo estimado para los trabajadores que se han dedicado a prestar servicios cumple los supuestos de MCO y de no ser así entonces se presentará sus respectivos métodos correctivos.

Tomando en cuenta los resultados del modelo estimado, que se muestran a continuación, se comprobará la existencia del cumplimiento de los supuestos de independencia entre los regresores y homocedasticidad.

```
. reg e25t3 e20t gastoss e8a

      Source |       SS           df       MS      Number of obs   =        87
-------------+----------------------------------   F(3, 83)        =    651.25
       Model |  98242834.8         3  32747611.6   Prob > F        =    0.0000
    Residual |  4173593.14        83  50284.2547   R-squared       =    0.9592
-------------+----------------------------------   Adj R-squared   =    0.9578
       Total |  102416428         86  1190888.7    Root MSE        =    224.24

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e20t |   .9820805   .0247914    39.61   0.000     .9327713    1.03139
    gastoss |  -.9630771   .0404672   -23.80   0.000    -1.043565   -.8825894
        e8a |   29.12427   45.43964     0.64   0.523    -61.25334    119.5019
      _cons |   17.79308    59.2564     0.30   0.765    -100.0655    135.6517
------------------------------------------------------------------------------
```

Figura 3.54. Regresión para los trabajadores independientes que se han dedicado a actividades prestadoras de servicios.

- **Multicolinealidad.**

Empecemos visualizando la matriz de correlación de las variables usadas en el modelo.

```
. corr e25t3 e20t gastoss e8a
(obs=87)

             |   e25t3     e20t  gastoss      e8a
-------------+------------------------------------
      e25t3  |   1.0000
       e20t  |   0.7815   1.0000
    gastoss  |   0.4010   0.8766   1.0000
        e8a  |   0.0977   0.4313   0.5825   1.0000
```

Figura 3.113. Matriz de correlación de las variables en el modelo (3.7.21.).

La matriz de correlación indica que el coeficiente de correlación entre las variables **ingresos** $(e20t)$ y **gastos** $(gastoss)$ es mayor frente a las demás coeficientes de correlación, no obstante, ningún coeficiente de correlación entre las regresoras supera el 0.90 por lo que podemos intuir que los estimadores del modelo no están tan influenciados por la multicolinealidad imperfecta en el caso que existiera. En otras palabras, es posible que el modelo está libre de multicolinealidad.

Ahora veamos los índices de VIF y TOL de los estimadores de las variables del modelo especificado.

```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
     gastoss |     5.55    0.180292
        e20t |     4.50    0.222111
         e8a |     1.58    0.633600
-------------+----------------------
    Mean VIF |     3.88
```

Figura 3.114. Índices de VIF y TOL de las variables en el modelo (3.7.21.).

Los índices de VIF y TOL de los estimadores de las variables del modelo especificado son menores a 10, por lo tanto se puede intuir que los estimadores del modelo son estables y no se encuentran influenciados por la multicolinealidad. A continuación, se presenta la gráfica matricial de dispersión entre las regresoras del modelo.

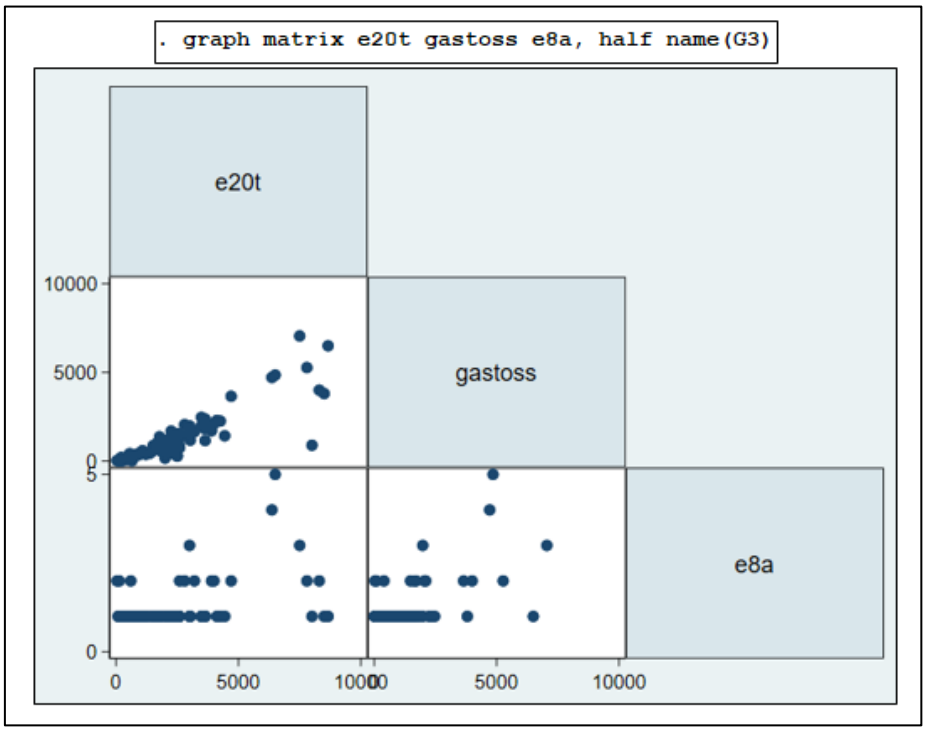


Figura 3.115. Grafica matricial sobre la correlación de las variables en el modelo (3.7.21.).

En la gráfica se puede ver que las variables **ingresos** $(\boldsymbol{e20t})$ y **gastoss**$(\boldsymbol{gastoss})$ se puede observar un patrón positivo por lo que se puede intuir que existe una correlación entre estas variables. Por otro lado, las gráficas de dispersión entre las demás variables no se distingue un patrón claro, aunque se observan datos atípicos, indicando que es posible que exista heterocedasticidad.

Siguiendo con los métodos de comprobación de multicolinealidad en el modelo, ahora se procederá a efectuar la verificación mediante la regla de Klein y el R^2 *de Theil*. Para realizar la comprobación mediante la regla de Klein se tomara el siguiente modelo auxiliar.

$$C_i = \alpha_1 + \alpha_2 I_i + \alpha_3 N_i + e_i \text{ (3.7.58.)}$$

```
. reg gastoss e20t e8a

      Source |       SS       df       MS      Number of obs   =        87
-------------+------------------------------   F(2, 84)        =    190.95
       Model |  139606762      2  69803380.8   Prob > F        =    0.0000
    Residual |   30706118     84  365549.024   R-squared       =    0.8197
-------------+------------------------------   Adj R-squared   =    0.8154
       Total |  170312880     86  1980382.32   Root MSE        =    604.61

------------------------------------------------------------------------------
     gastoss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e20t |   .5224036   .0349166    14.96   0.000     .4529682     .591839
         e8a |   528.6118   108.0904     4.89   0.000     313.6621    743.5614
       _cons |  -694.8566   140.6348    -4.94   0.000    -974.5244   -415.1888
------------------------------------------------------------------------------
```

Figura 3.116. Resultados de la regresión auxiliar (3.7.58.).

En la figura 3.116. Se observa que el coeficiente de determinación del modelo auxiliar es 0.8197, mientras que el modelo original tiene un coeficiente de determinación de 0.9592, por lo tanto, si aplicamos la regla de Klein nos damos cuenta que no existe multicolinealidad en el modelo original, ya que su coeficiente de determinación es mayor al coeficiente de determinación del modelo auxiliar.

Ahora para realizar la comprobación mediante el R^2 *de Theil* especificaremos las siguientes regresiones auxiliares.

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 C_i + v_i \text{ (3.7.59.)}$$

$$G_i = \theta_1 + \theta_2 C_i + \theta_3 N_i + v_i \text{ (3.7.60.)}$$

$$G_i = \theta_1 + \theta_2 I_i + \theta_3 N_i + v_i \text{ (3.7.61.)}$$

```
. reg e25t3 e20t gastoss

      Source |       SS       df       MS      Number of obs   =        87
-------------+------------------------------   F(2, 84)        =    983.57
       Model |  98222177.6      2  49111088.8   Prob > F        =    0.0000
    Residual |  4194250.39     84  49931.5523   R-squared       =    0.9590
-------------+------------------------------   Adj R-squared   =    0.9581
       Total |  102416428     86  1190888.7   Root MSE        =    223.45

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        e20t |    .97886   .0241916    40.46   0.000     .9307523    1.026968
     gastoss |  -.9508667   .0355771   -26.73   0.000    -1.021616   -.8801177
       _cons |   47.0944   37.56875     1.25   0.213    -27.61518     121.804
------------------------------------------------------------------------------
```

```
. reg e25t3 gastoss e8a

      Source |       SS       df       MS      Number of obs   =        87
-------------+------------------------------   F(2, 84)        =      9.77
       Model |  19334557.6     2  9667278.81   Prob > F        =    0.0002
    Residual |  83081870.3    84  989069.885   R-squared       =    0.1888
-------------+------------------------------   Adj R-squared   =    0.1695
       Total |  102416428    86  1190888.7   Root MSE        =    994.52

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    gastoss |   .4038858   .0937506     4.31   0.000     .2174525    .5903191
       e8a |  -335.6971   197.3441    -1.70   0.093    -728.1375    56.7433
      _cons |    1163.01   229.4058     5.07   0.000     706.8109    1619.208
------------------------------------------------------------------------------
```

Figura 3.117. Resultados de la regresión auxiliar (3.7.59.).

Figura 3.118. Resultados de la regresión auxiliar (3.7.60.).

```
. reg e25t3 e20t e8a

      Source |       SS       df       MS      Number of obs   =        87
-------------+------------------------------   F(2, 84)        =     89.73
       Model |  69762373.4     2  34881186.7   Prob > F        =    0.0000
    Residual |  32654054.6    84  388738.745   R-squared       =    0.6812
-------------+------------------------------   Adj R-squared   =    0.6736
       Total |  102416428    86  1190888.7   Root MSE        =    623.49

------------------------------------------------------------------------------
      e25t3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      e20t |   .4789655   .0360071    13.30   0.000     .4073616    .5505695
       e8a |  -479.9696   111.4662    -4.31   0.000    -701.6324   -258.3068
      _cons |   686.9936    145.027     4.74   0.000     398.5914    975.3958
------------------------------------------------------------------------------
```

Figura 3.119. Resultados de la regresión auxiliar (3.7.61.).

Con los resultados de las regresiones auxiliares podremos hallar el $R^2\ de\ Theil$ tomando sus respectivos coeficientes de determinación.

$$R^2\ de\ Theil = 0.9592 - (0.9592 - 0.9590) - (0.9592 - 0.1888) - (0.9592 - 0.6812) = -0.0894 \quad (3.7.62.)$$

Ya que el $R^2\ de\ Theil$ es cercano a 0 podemos argumentar que no existe multicolinealidad en el modelo especificado. Finalmente, se comprobará la existencia de multicolinealidad en el modelo mediante el contraste de la prueba F. Para realizar este contraste se utilizará el siguiente modelo auxiliar: $C_i = \alpha_1 + \alpha_2 I_i + \alpha_3 N_i + e_i$, el cual se le planteará el siguiente contraste de hipótesis.

$$H_0: No\ existe\ multicolinealidad$$

$$H_1: Existe\ multicolinealidad$$

El estadístico F calculado se consigue mediante.

$$Fc = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)} = \frac{0.8197/(3-2)}{(1-0.8197)/(87-3+1)} = 193.22 \text{ (3.7.63.)}$$

Por otro lado el estadístico F tabulado es $Ft^{1}_{85,0.05} = 3.95$ entonces al tener $|Fc| > Ft$, rechazamos la hipótesis nula y asumimos que la variable ***gastoss*** puede ser causante de multicolinealidad en el modelo.

No obstante, tomando en cuenta los índices de VIF y TOL, el coeficiente de correlación no sea tan elevado, los estimadores del modelo original son significativos y que el coeficiente de determinación sea elevado, podemos concluir que las regresoras están correlacionadas entre sí, pero no influyen en los estimadores. Por lo tanto, no debería plantearse ejecutar un método correctivo porque la multicolinealidad no está afectando a la regresión del modelo original.

- **Homocedasticidad.**

En esta sección comprobaremos si el modelo cumple el supuesto de homocedasticidad mediante los métodos informales y formales. En el caso que exista heterocedasticidad en el modelo se deberá ejecutar un método correctivo. Para realizar los métodos informales se usarán los gráficos de dispersión entre los residuos y los valores ajustados de la variable dependiente y con cada regresor del modelo.

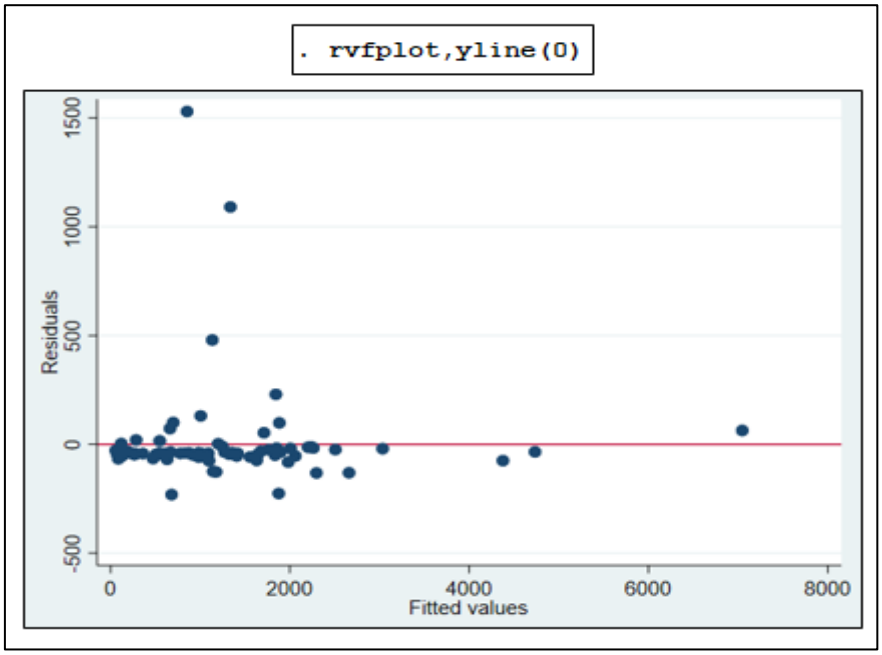


Figura 3.120. Grafica de dispersión entre los residuos y los valores ajustados de la variable dependiente.

La grafica de dispersión anterior nos indica que existen datos atípicos en el modelo, por lo tanto podría haber heterocedasticidad. Ahora veamos las gráficas de dispersión entre los residuos con las regresoras.

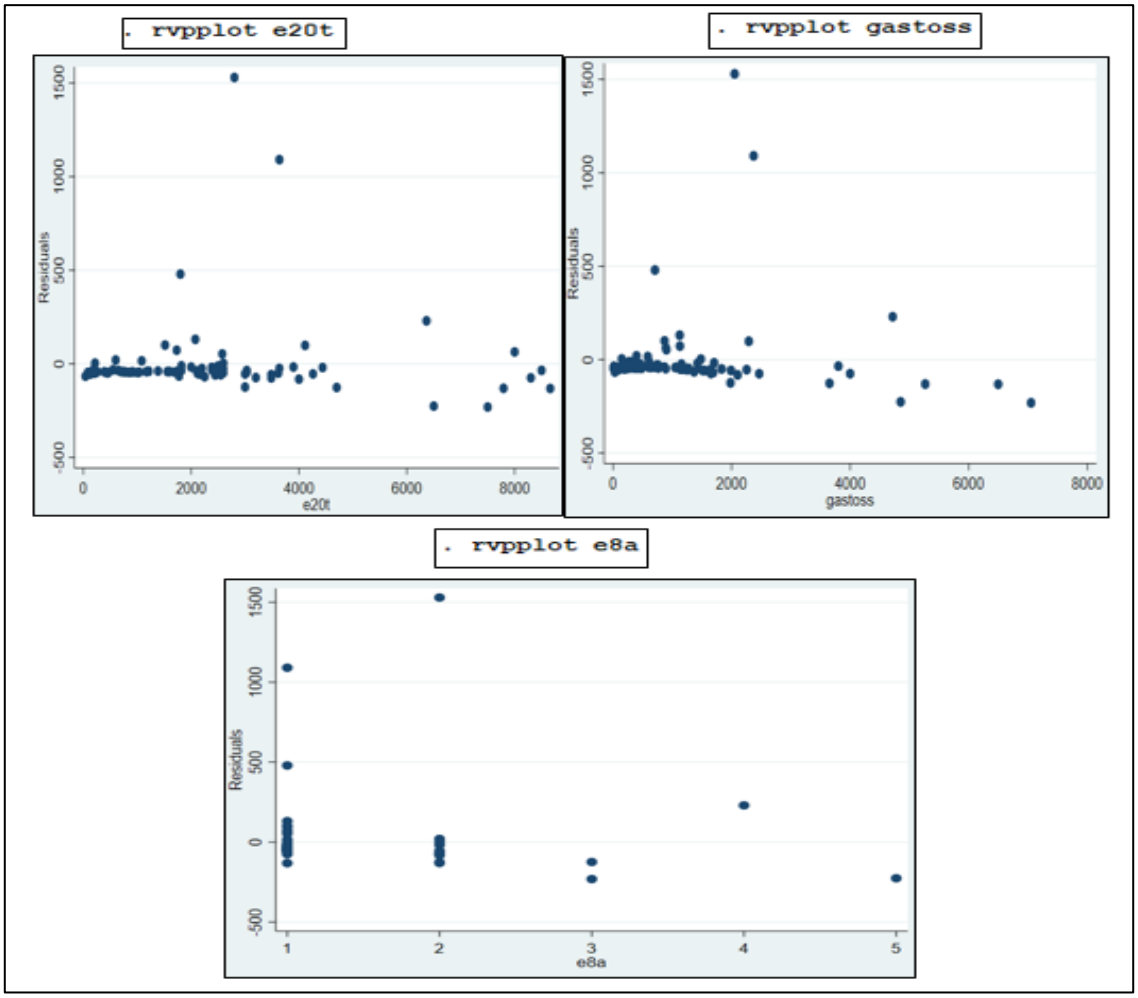


Figura 3.121. Grafica de dispersión entre los residuos y los regresores del modelo.

En los gráficos de dispersión entre los residuos y los regresores del modelo vistos anteriormente, podemos notar la existencia de datos atípicos por ello se puede sospechar que las regresoras del modelo pueden causar heterocedasticidad en el modelo.

Ahora comprobaremos si el modelo cumple el supuesto de homocedasticidad mediante las pruebas de BG y de White con el contraste de hipótesis, las cuales son:

$$H_0: No\ existe\ heterocedasticidad$$

$$H_1: Existe\ heterocedasticidad$$

Recordemos que la prueba BG asume que los residuos del modelo deben seguir la distribución normal y ya que este modelo no tiene residuos que siguen una distribución normal, entonces los resultados de la prueba BG puede verse afectada. No obstante STATA permite realizar la prueba BG sin asumir que los residuos sigan la distribución normal mediante las opciones **iid** y **fstat**, la primera opción utiliza el estadístico chi-cuadrado (X^2) y la segunda opción utiliza el estadístico F calculado.

```
. estat hettest,fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of e25t3

         F(1 , 85)     =     0.07
         Prob > F      =   0.7965
. estat hettest e20t,fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e20t

         F(1 , 85)     =     0.35
         Prob > F      =   0.5558

. estat hettest gastoss,fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: gastoss

         F(1 , 85)     =     1.24
         Prob > F      =   0.2692

. estat hettest e8a,fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e8a

         F(1 , 85)     =     1.13
         Prob > F      =   0.2912
```

Figura 3.122. Prueba BG de heterocedasticidad con el estadístico *F*.

Mediante la prueba BG usando el estadístico F podemos notar que el valor-p en cada prueba de hipótesis es mayor al 5% de significancia, por lo tanto, podemos aceptar la hipótesis nula y asumir que no existe heterocedasticidad.

A continuación, utilizaremos la opción **iid** la cual utiliza el estadístico chi-cuadrado para realizar la prueba BG sin asumir que los residuos siguen una distribución normal.

```
. estat hettest e20t,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e20t

         chi2(1)      =     0.36
         Prob > chi2  =   0.5504

.
. estat hettest gastoss,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: gastoss

         chi2(1)      =     1.25
         Prob > chi2  =   0.2639

.
. estat hettest e8a,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: e8a

         chi2(1)      =     1.14
         Prob > chi2  =   0.2858
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of e25t3

         chi2(1)      =     0.07
         Prob > chi2  =   0.7936
```

Figura 3.123. Prueba BG de heterocedasticidad con el estadístico chi-cuadrado.

Según la prueba BG de heterocedasticidad utilizando el estadístico chi-cuadrado visto en la figura anterior, podemos aceptar la hipótesis nula y asumir que el modelo no tiene heterocedasticidad, ya que sus respectivos valores-p son mayores al 5% de significancia. Ahora ejecutemos la prueba general de heterocedasticidad de White, para determinar si existe o no heterocedasticidad en el modelo.

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =    10.92
         Prob > chi2  =   0.2810

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	10.92	9	0.2810
Skewness	3.46	3	0.3259
Kurtosis	1.66	1	0.1974
Total	16.05	13	0.2467

Figura 3.124. Prueba general de White de heterocedasticidad.

La prueba general de White de heterocedasticidad indica que el valor-p es mayor a la significancia del 5%, en consecuencia, se acepta la hipótesis nula y se asume que no existe heterocedasticidad en el modelo. Para estar más seguros, comprobemos el resultado con la prueba pura de White de heterocedasticidad.

```
. predict u,resid

. gen uc=u^2

. gen x2c=e20t^2

. gen x3c=gastoss^2

. gen x4c=e8a^2

. reg uc e20t gastoss e8a x2c x3c x4c
```

Source	SS	df	MS	Number of obs	=	87
				F(6, 80)	=	1.45
Model	6.6067e+11	6	1.1011e+11	Prob > F	=	0.2075
Residual	6.0900e+12	80	7.6125e+10	R-squared	=	0.0979
				Adj R-squared	=	0.0302
Total	6.7506e+12	86	7.8496e+10	Root MSE	=	2.8e+05

uc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
e20t	-36.66244	72.68288	-0.50	0.615	-181.306	107.9811
gastoss	193.1917	111.7636	1.73	0.088	-29.22484	415.6083
e8a	252278.4	182457.8	1.38	0.171	-110824.2	615381
x2c	-.000939	.0073188	-0.13	0.898	-.015504	.0136259
x3c	-.0216209	.0149122	-1.45	0.151	-.051297	.0080553
x4c	-48692.41	35730.77	-1.36	0.177	-119798.9	22414.1
_cons	-234195	169242.5	-1.38	0.170	-570998.3	102608.3

Figura 3.125. Prueba pura de White de heterocedasticidad.

Con estos resultados construimos el siguiente estadístico calculado.

$n * R^2 = 87 * 0.0979 = 8.44$ (3.7.64.)

El cual se distribuye según el siguiente estadístico tabulado.

$X_6^2 = 12.60$ (3.7.65.)

Podemos notar que el estadístico calculado es menor al estadístico tabulado, entonces se acepta la hipótesis nula y se asume que el modelo si cumple con el supuesto de homocedasticidad mediante la prueba pura de heterocedasticidad de White.

Luego de haber contrastado estas hipótesis y demostrar que el modelo cumple con el supuesto de homocedasticidad, entonces no es necesario plantearse algún método correctivo.

3.7.1.7. Interpretación de los resultados.

En esta última sección interpretaremos los resultados finales obtenidos en la sección anterior, cuyos resultados provienen de las regresiones que cumplen los supuestos de MCO. En la siguiente tabla se muestra un resumen sobre los estimadores obtenidos de las regresiones.

	Actv. Productiva (3.7.37.)			Actv. Comercial (3.7.56.)			Actv. Prestadora de servicios (3.7.11.)		
	$\hat{\beta}_k$	ee	t	$\hat{\beta}_k$	ee	t	$\hat{\beta}_k$	ee	t
Constante	−91.2	(187.19)	−0.49	70.17	(127.87)	10.43	17.80	(59.26)	39.61
Ingresos	0.52	(0.41)	1.27	0.91	(0.09)	−8.39	0.98	(0.02)	−23.80
Gastos	−0.59	(0.38)	−1.51	−0.87	(0.10)	0.21	−0.96	(0.04)	0.64
Número de trabajadores	387.06	(296.03)	1.31	21.20	(127.87)	0.55	29.13	(45.44)	0.30
Número de observaciones	27			86			87		
Coeficiente de determinación	0.7141			0.7688			0.9592		
Significancia global	17.10			93.68			651.25		
Error estándar de regresión	639.09			338.65			224.24		

Tabla 3.21. Resultados de los modelos especificados para explicar el nivel de ganancias netas de los trabajadores independientes según las actividades que se dedican en el distrito de Chiclayo en 2018.

El modelo especificado ha sido estimado según las actividades que se han realizado los trabajadores independientes y se han seleccionado a las variables ***ingresos***, ***gastos*** y el ***número de trabajadores*** de los trabajadores independientes para explicar su nivel de **ganancias netas**. Las regresiones muestran resultados notoriamente diferentes

entre las actividades. Podemos notar que el modelo con una mejor bondad de ajuste es la que explica a los trabajadores dedicados a la actividad prestadora de servicios y a la vez es el único modelo que no ha presentado violaciones a los supuestos de no multicolinealidad y homocedasticidad, por lo que se puede asumir que este modelo tiene los estimadores más confiables entre los tres modelos especificados. Por otro lado, solamente en los modelos dedicados a las actividades comerciales y de servicios el intercepto es positivo por lo que las ganancias netas de los trabajadores en estas actividades han crecido, mientras los trabajadores dedicados a la actividad productiva tienen ganancias netas decrecientes ya que su intercepto es negativo.

En cuanto, a los estimadores de las regresoras de los trabajadores independientes dedicados a la actividad de servicios, tienen el estimador de la variable ***ingresos*** mayor con respecto a las los demás modelos, entonces podemos inferir que en el distrito de Chiclayo los trabajadores independientes reciben más ingresos si se dedican a la actividad de servicios. No obstante, también son los que más afectados tienen sus ganancias netas si gastan más unidades monetarias. En conclusión, los trabajadores independientes que se dedican a actividades comerciales y de servicios perciben más ganancias netas.

3.7.2. Ejemplo con el uso de datos de series temporales.

Ahora se presentará un ejemplo de cómo construir un modelo econométrico con datos de series temporales. No se pondrá énfasis ni en el planteamiento del problema ni en el marco teórico, con el fin de dar más espacio a la explicación de cómo utilizar el método de estimación de MCO con datos de series temporales.

Las series temporales son usadas en su amplitud para explicar el comportamiento de variables macroeconómicas centrándose en su evolución en el tiempo y como han sido influenciadas por otras variables. Es importante tener en cuenta el **tiempo** en este tipo de modelos econométricos debido a que las variables dependen de sí mismas en tiempos pasados y muchas veces se necesita capturar el efecto del mismo para explicar las relaciones entre las variables.

En econometría, el término “tiempo” hace referencia a los procesos aleatorios que influyen sobre las variables económicas, en palabras más simples, la idea es representar sucesos no previstos como crisis, epidemias, guerras, efecto climático, etc.; y cómo estos sucesos afectan a las variables económicas. La finalidad de las series temporales es

replicar los procesos aleatorios y predecir el comportamiento futuro de las variables usando información pasada.

En este ejemplo se especificará un modelo econométrico que explicará el comportamiento de las importaciones peruanas desde el año 1999 al año 2019, con una frecuencia trimestral y en millones de soles.

3.7.2.1. Especificación del modelo econométrico.

Este es el modelo econométrico que se utilizará para explicar el comportamiento que han tenido las importaciones peruanas desde el primer trimestre del 1999 hasta el último trimestre del 2019.

$$IMP_t = \hat{\beta}_1 + \hat{\beta}_2 PBI_t + \hat{\beta}_3 INDP_t + \hat{\beta}_4 IBI_t + \hat{\mu}_t \quad (3.7.66.)$$

- IMI_t: Importaciones totales en millones de soles (2007=100).
- PBI_t: Producto Bruto Interno peruano en millones de soles (2007=100).
- $INDP_t$: Índice de protección.
- IBI_t: Inversión Bruta Interna en millones de soles (2007=100).
- μ_t : Término de perturbación, proceso aleatorio que recoge los efectos capturados de variables no introducidas en el modelo econométrico pero que influyen en la variable dependiente.

Estas variables se pueden encontrar en las series estadísticas del BCRP a excepción del índice de protección, la cual debe ser construida con la siguiente fórmula.

$$INDP_t = \frac{IMI.\ INSUMOS + IMI.\ BIENES\ DE\ CAPITAL}{IMP.\ TOT} \quad (3.7.67.)$$

Los estimadores que acompañan a las variables utilizadas en el modelo econométrico tienen los siguientes signos esperados.

- PBI_t: Esta variable es considerada en este modelo como la variable ingreso de la economía peruana, por lo que tiene una relación directa con las importaciones.
- $INDP_t$: Esta variable es construida con otras variables las cuales son importación de insumos, importación de bienes de capital e importación total, tal como muestra la fórmula (3.7.67.). Se le ha incluido en el modelo debido

a que representa el nivel de protección de la economía peruana, por lo que tiene una relación inversa con las importaciones.

- IBI_t: La inversión bruta interna es una variable que está incluida debido a que en las economías dependientes con poca industrialización, como la economía peruana, dependen de la inversión para que aumenten sus importaciones. Por lo tanto, es una relación directa.

3.7.2.2. Acceso a la base de datos.

Se ha recalcado que las variables $IMP_t, PBI_t \ y \ IBI_t$ deben tener el año base 2007, por lo que en las series estadísticas del BCRP se buscarán las siguientes series estadísticas.

- IBI_t: PN02531AQ – PBI por tipo de gasto (millones S/ 2007) – Demanda Interna – Inversión Bruta Inversión.
- PBI_t: PN02538AQ – PBI por tipo de gasto (millones S/ 2007) – PBI.
- IMP_t : PN02537AQ – PBI por tipo de gastos (millones S/ 2007) – Importaciones.

La variable $INDP_t$ no se encuentra de forma literal en la serie estadística, por eso aquí se muestran cuáles son las variables que se podrán utilizar para construirla:

- $IMP.\ INSUMOS_t$: Importaciones según su uso o destino económico – valores FOB (millones US$) – Insumos.
- $IMP.\ BIENES\ DE\ CAPITAL_t$: Importaciones según uso o destino económico – valores FOB (millones US$) – Bienes de Capital.
- $IMP.\ TOT_t$: Importaciones según uso o destino económico - valores FOB (millones US$) - Total Importaciones.

Una vez construida la base de datos en STATA, se debería empezar indicando a STATA que se trabajará con datos de series temporales. Primero debemos generar la variable ***trimestre*** que recoge información sobre los trimestres de todos los años. El comando **gen** será utilizado para crear la variable ***trimestre***.

```
. gen trimestre=yq(1999,1)+t-1
```

Figura 3.126. Generando variable trimestre (1).

En la figura 3.126; se utiliza el comando **gen** para crear la variable ***trimestre*** usando el componente **yq** el cual es una función de datos temporales que indica a STATA

crear una variable que contenga información trimestral usando los indicadores que están entre sus paréntesis; el primer indicador es el año con el que se empieza la serie y el segundo indicador es el número del trimestre. Revisemos la base de datos para comprobar que se ha creado dicha variable.

tiempo	trimestre
1	156
2	157
3	158
4	159
5	160
6	161
7	162
8	163
9	164
10	165

Figura 3.127. Generando variable trimestre (2).

Para cambiar el formato de la variable generada se utiliza el comando **format**.

```
. format trimestre %tq
```

Figura 3.128. Generando variable trimestre (3).

El componente **%tq** transforma el formato original de la serie en un formato que muestra el año y el trimestre que se ha creado. En la base de datos, los valores de la variable ***trimestre*** se observan siguiendo un formato trimestral.

tiempo	trimestre
1	1999q1
2	1999q2
3	1999q3
4	1999q4
5	2000q1
6	2000q2
7	2000q3
8	2000q4
9	2001q1
10	2001q2

Figura 3.129. Generando variable trimestre (3).

Ahora que ya tenemos la variable ***trimestre,*** ya podemos instruir a STATA que se trabajará con datos de tiempo trimestrales con el comando **tsset**.

```
. tsset trimestre
        time variable:  trimestre, 1999q1 to 2019q4
                delta:  1 quarter
```

Figura 3.130. Datos de series trimestrales.

Ahora construiremos la variable ***indp_v*** la cual representa el índice de protección especificado en el modelo, para ello utilizaremos el comando **gen** y las variables ***impi***, ***impbc*** y ***impdolar*** las cuales representan las importaciones en dólares de insumos, importaciones en dólares de bienes de capital e importaciones totales en dólares respectivamente.

```
. gen indpd=( impi+ impbc)/ impdolar
```

Figura 3.131. Construyendo la variable índice de protección (1).

La variable ***indpd*** en la figura 3.131. Es el índice de protección en dólares por lo que se utilizará la variable ***tc_v*** que representa el tipo de cambio trimestral, para convertir el índice en soles.

```
. gen indp_v=indpd*tc_v
```

Figura 3.132. Construyendo la variable índice de protección (2).

Ahora ya se ha logrado obtener el índice de protección en soles. Posteriormente, STATA permite la ejecución de gráficos de línea que ayuda a analizar cómo evoluciona la serie temporal con respecto al tiempo. Para realizar esta gráfica de línea se debe digitar el comando **tsline**.

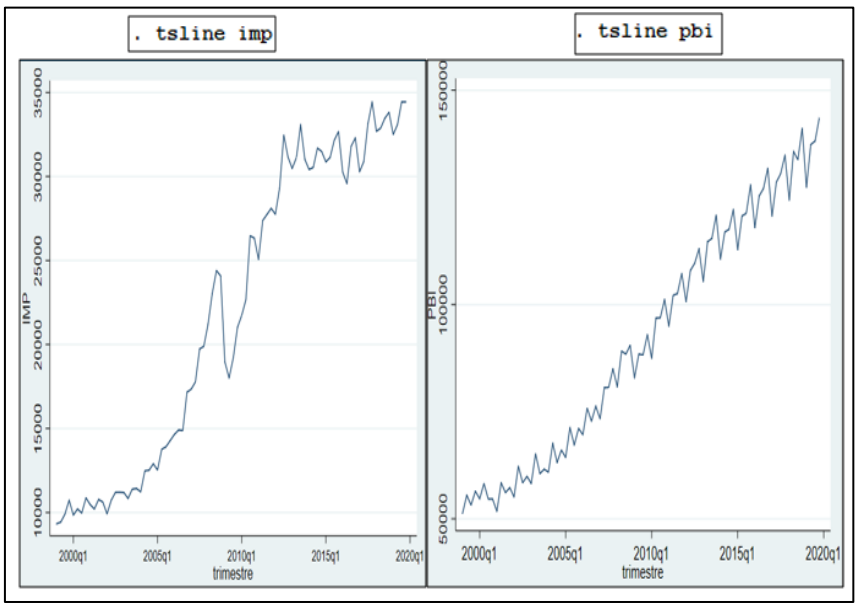


Figura 3.133. Grafica de línea de las variables *imp* y *pbi*.

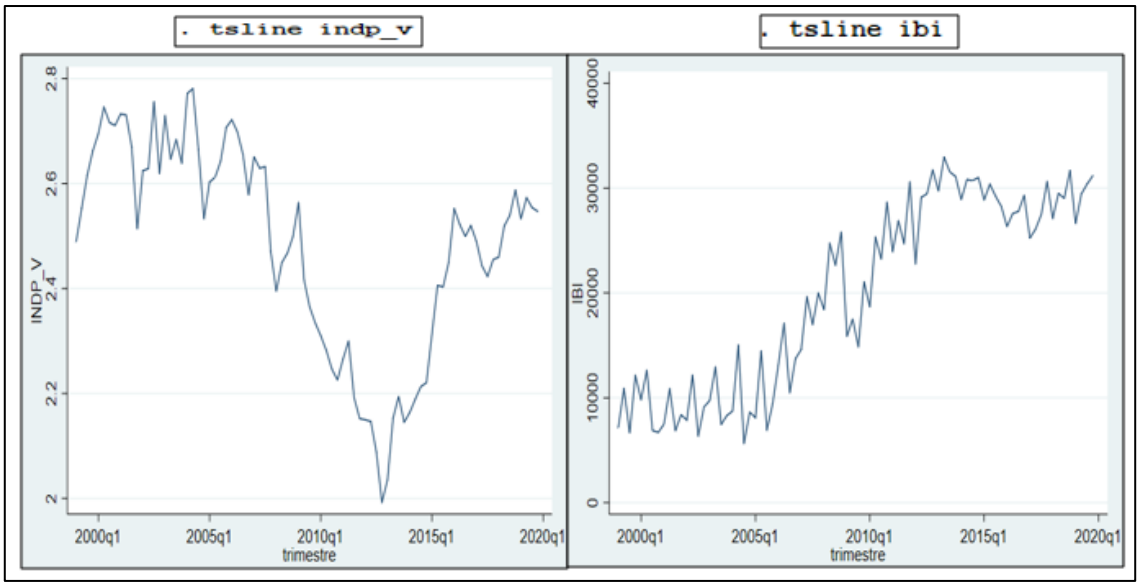


Figura 3.134. Grafica de línea de las variables *indp_v* y *ibi*.

Estas gráficas muestran cómo las variables seleccionadas evolucionan según la función del tiempo, en este caso de forma trimestral. Según las gráficas las variables ***imp***, ***pbi*** e ***ibi*** tienen una tendencia creciente, es decir aumentan sus valores conforme aumentan los trimestres. Además, se puede ver que no son variables estacionarias ya que tienen picos en toda la línea, de hecho, la variable ***pbi*** muestra esto a la perfección. Asumir que las variables no son estacionarias implica concluir que su media y su varianza no son constantes en el tiempo por lo que podría existir la posibilidad que el modelo tenga autocorrelación.

Por otro lado, la variable ***indp_v*** tiene una tendencia negativa muy notoria desde el tercer trimestre del año 2005 hasta el cuarto trimestre del año 2012, por lo que en este periodo el índice de protección ha sido inferior y por lo tanto las importaciones debieron aumentar.

Estas gráficas indican que las variables no son estacionarias, y debido a su condición de no estacionariedad podrían afectar a las variables del modelo. Algunos autores recomiendan usar sus respectivos logaritmos y reemplazarlos en el modelo original, con el fin de revertir la no estacionariedad en las variables. De tal forma que el modelo (3.7.66.) se transforma en el siguiente modelo.

$$lIMP_t = \hat{\beta}_1 + \hat{\beta}_2 lPBI_t + \hat{\beta}_3 lINDP_t + \hat{\beta}_4 lIBI_t + \hat{e}_t \text{ (3.7.68.)}$$

Al aplicar logaritmos en ambos lados de la igualdad, el modelo original pasa a transformarse en un modelo log-log, como se observa en el modelo especificado (3.7.68.).

Con el fin de notar la diferencia entre ambos modelos, veamos un cuadro que sirve de resumen sobre los principales estadísticos descriptivos de las variables entre los modelos (3.7.66.) y (3.7.68.). El comando **sum** será requerido para la generación de tal cuadro.

```
. sum imp pbi ibi indp_v

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         imp |         84     21939.9     9102.559    9323.021   34442.83
         pbi |         84     91867.7     28439.78    51214.63   143698.5
         ibi |         84    19970.95     9124.595    5663.859   32947.09
```

Figura 3.135. Cuadro descriptivo de las variables del modelo (3.7.66.).

```
. sum lpbi lindp_v libi

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        lpbi |         84    11.37845    .3210931    10.84378   11.87547
     lindp_v |         84    .9056098    .0829861    .6894119   1.022755
        libi |         84    9.769755    .5515062    8.641861   10.40266
```

Figura 3.136. Cuadro descriptivo de las variables del modelo (3.7.68.).

En la figura 3.135. Se logra visualizar un resumen que brinda información sobre los descriptivos de las variables del modelo (3.7.66.). Según la tabla creada con el comando **sum,** la variable explicativa ***pbi*** tiene la desviación estándar más alta, por lo que esta variable podría ocasionar problemas en el momento de estimar el modelo estándar muy ínfima, entonces se puede entender que la variable tiende a tener valores constantes.

Mientras, que la figura 3.136. Muestra información sobre los estadísticos descriptivos de las variables del modelo (3.7.68.) y a diferencia de las variables del modelo (3.7.66.), la desviación estándar es muy inferior, lo que significa que los valores de los logaritmos de las variables tienden a mantenerse constante en el tiempo.

Revisemos en STATA sus respectivos gráficos de línea, los cuales mostraran cómo evolucionan las series temporales de las variables logarítmicas.

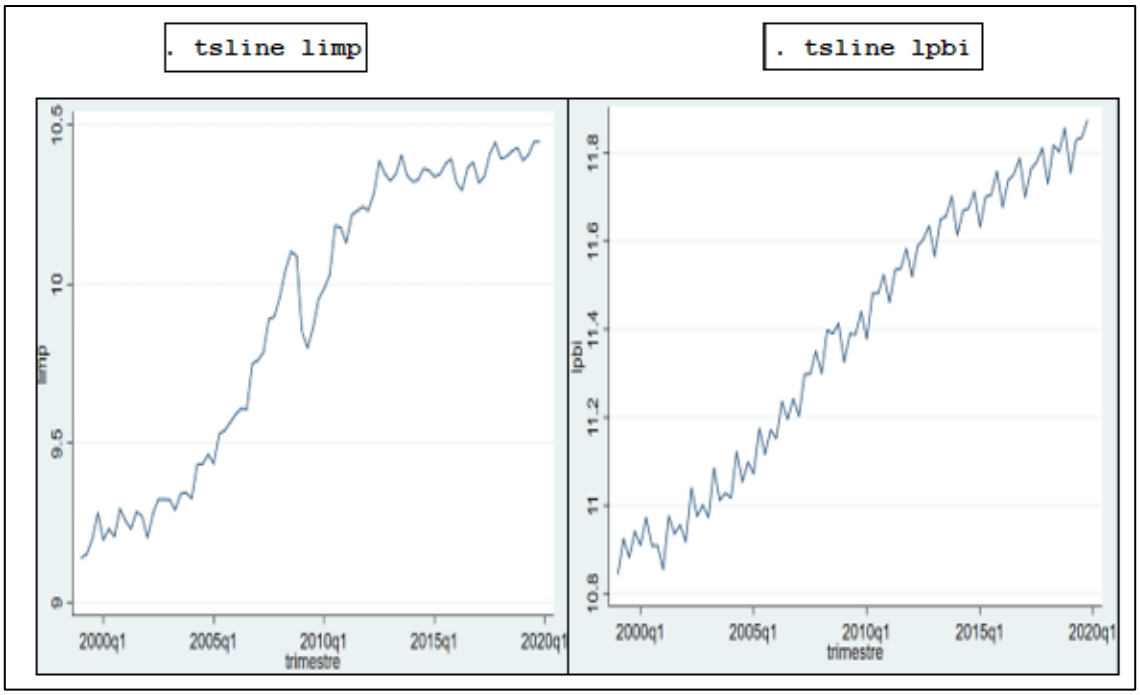


Figura 3.137. Gráficos de línea de las variables *limp* y *lpbi.*

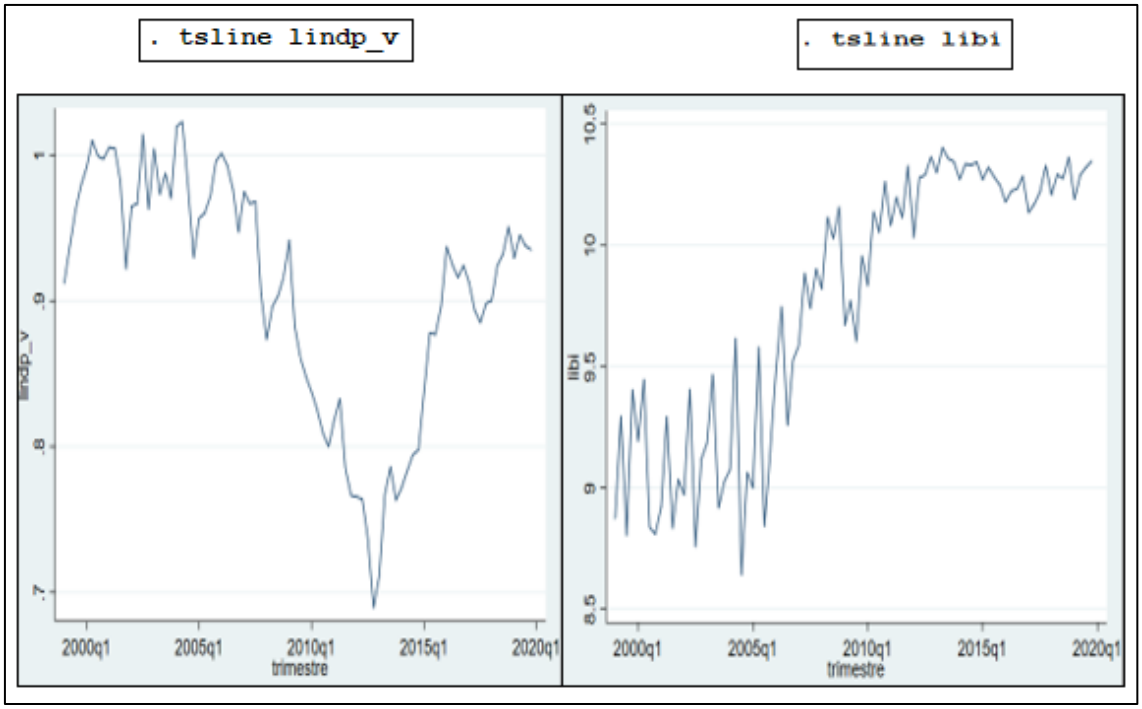


Figura 3.138. Gráficos de línea de las variables *lindp_v* y *libi.*

En las gráficas de las figuras anteriores se observa la evolución de las variables del modelo (3.7.68.) en el tiempo. Si comparamos los valores en el eje horizontal de los gráficos de las variables logarítmicas con respecto a los valores del eje horizontal de los gráficos de las variables lineales, nos daremos cuenta de la reducción entre ambas de

forma notoria, en consecuencia a dicha reducción, los valores de las variables logarítmicas tienden a ser constantes en el tiempo.

No obstante, las variables logarítmicas mantienen tendencias similares a las variables del modelo (3.7.66.). Esto significa que las variables no son estacionarias en su media, y solo la variable ***lpbi*** parecer ser estacionaria en su varianza, en consecuencia a que la dispersión a lo largo de la grafia de línea no muestra picos notablemente diferenciados.

La estacionariedad de las variables puede influir en sus resultados. Con el fin de mostrar tales influencias, se harán las regresiones en los modelo (3.7.66.) y (3.7.68.) y se mostrarán como las interpretaciones entre ambas regresiones son distintas. No obstante, sólo se utilizará al modelo (3.7.66.) para explicar de forma lineal a las importaciones.

3.7.2.3. Estimación de los coeficientes de regresión.

Ahora realizaremos la regresión del modelo (3.7.66.) mediante MCO utilizando el comando **reg**.

```
. reg imp pbi indp_v ibi

      Source |       SS           df       MS      Number of obs   =        84
-------------+----------------------------------   F(3, 80)        =   1498.25
       Model |  6.7568e+09         3  2.2523e+09   Prob > F        =    0.0000
    Residual |   120261734        80  1503271.68   R-squared       =    0.9825
-------------+----------------------------------   Adj R-squared   =    0.9819
       Total |  6.8771e+09        83  82856574.5   Root MSE        =    1226.1

------------------------------------------------------------------------------
         imp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |   .2382184   .0152821    15.59   0.000     .2078061    .2686307
       indp_v |  -5965.741   1093.501    -5.46   0.000    -8141.878   -3789.605
         ibi |   .1717964   .0567696     3.03   0.003     .0588213    .2847715
       _cons |    11429.6    2872.71     3.98   0.000     5712.724    17146.48
------------------------------------------------------------------------------
```

Figura 3.139. Regresión mediante MCO del modelo (3.6.66).

$$IMP_t = 11429.6 + 0.24PBI_t - 5965.74INDP_t + 0.17IBI_t + \hat{\mu}_t \quad (3.7.69.)$$

$$ee = (2872.71) \quad (0.02) \quad (1093.50) \quad (0.06)$$

$$t = 3.98 \quad 15.59 \quad -5.46 \quad 3.03$$

La estimación del modelo (3.7.68.) mediante MCO calcula estimadores que cumplen los signos esperados, y sus respectivos valores-p son menores a un nivel de significancia del 5%, en consecuencia las variables tienen significancia individual.

El modelo también tiene significancia global debido a que el valor-p del estadístico F es menor a una significancia del 5%. Además, cuenta con una excelente bondad de ajuste, pero al ser 0.9825 podemos sospechar que posiblemente el modelo tenga multicolinealidad.

Tomando en cuenta el ceteris paribus, los estimadores se pueden interpretar de la siguiente manera.

PBI_t: Si el PBI peruano aumenta en un millón de soles, entonces las importaciones peruanas aumentan en 0.23 millones de soles.

$INDP_t$: Si el índice de protección en soles aumenta en una unidad, entonces las importaciones peruanas disminuyen en 5965.74 millones de soles.

IBI_t: Si la inversión bruta interna aumenta en un millón de soles, entonces las importaciones peruanas aumentan en 0.17 millones de soles.

Ahora veamos los resultados de la regresión del modelo con variables logarítmicas calculados mediante MCO con el comando **reg**.

```
. reg limp lpbi lindp_v libi

      Source |       SS           df       MS      Number of obs   =        84
-------------+----------------------------------   F(3, 80)        =   1263.52
       Model |  17.5006251         3  5.83354169   Prob > F        =    0.0000
    Residual |  .369352654        80  .004616908   R-squared       =    0.9793
-------------+----------------------------------   Adj R-squared   =    0.9786
       Total |  17.8699777        83  .215300937   Root MSE        =    .06795

------------------------------------------------------------------------------
        limp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpbi |    1.219585   .0654788    18.63   0.000     1.089278    1.349892
    lindp_v |   -.7453668   .1253454    -5.95   0.000    -.9948122   -.4959214
        libi |    .0534616   .0421295     1.27   0.208    -.0303787     .137302
       _cons |   -3.827133   .4470966    -8.56   0.000    -4.716883   -2.937382
------------------------------------------------------------------------------
```

Figura 3.140. Regresión mediante MCO del modelo (3.6.68).

$$lIMP_t = -3.82 + 1.22lPBI_t - 0.74lINDP_t + 0.05lIBI_t + \hat{e}_t \text{ (3.7.70.)}$$

$$ee = (0.44) \quad (0.07) \quad (0.13) \quad (0.04)$$

$$t = -8.56 \quad 18.63 \quad -5.95 \quad 1.27$$

Según la figura 3.140., el modelo (3.7.70.) concluimos que tiene significancia global, al revisar el valor-p del estadístico F calculado es menor al 5%. Además, los estimadores que acompañan a las variables son significativas individualmente debido a que sus respectivos valores-p son menores a una significancia del 5% a excepción del estimador que acompaña a la variable ***libi***

El modelo también presenta una buena bondad de ajuste, tal como señala el coeficiente de determinación que es igual a 97.93%. Suponiendo el ceteris paribus, los estimadores se interpretan de la siguiente forma.

$lPBI_t$: Si el PBI peruano aumenta en una unidad porcentual entonces las importaciones peruanas aumentan en 1.21%.

$lINDP_t$: Si el índice peruano de protección aumentan en una unidad porcentual entonces las importaciones peruanas disminuyen en 0.74%.

$lIBI_t$: Si la inversión bruta de inversión aumentan en unidad porcentual entonces las importaciones peruanas aumentan en 0.05%

3.7.2.4. Evaluación del cumplimiento de los supuestos.

En esta sección se verificará si el modelo original cumple con los supuestos de MCO sobre independencia entre los regresores, homocedasticidad y no autocorrelación.

- **Modelo original.**
  - **No multicolinealidad.**

Para empezar a comprobar si existe multicolinealidad en el modelo, veamos la matriz de correlación entre las variables del modelo.

```
. corr pbi indp_v ibi
(obs=84)

             |      pbi    indp_v      ibi
-------------+---------------------------
         pbi |   1.0000
     indp_v |  -0.5591    1.0000
         ibi |   0.9367   -0.7184    1.0000
```

Figura 3.141. Matriz de correlación de las variables en el modelo (3.7.69.).

La matriz de correlación muestra que el coeficiente de correlación entre las variables ***ibi*** y ***pbi*** es el más alto de todos, entonces se puede sospechar que estas variables

están causando multicolinealidad en el modelo. Para estar más seguro veamos los índices VIF y TOL de las variables.

```
. vif

    Variable |      VIF       1/VIF
-------------+----------------------
         ibi |    14.81    0.067499
         pbi |    10.43    0.095883
       indp_v |     2.64    0.378332
-------------+----------------------
    Mean VIF |     9.30
```

Figura 3.142. Índice VIF y TOL de las variables en el modelo (3.7.69.).

El índice VIF de la variable ***ibi*** se encuentra entre 10 y 30, por tal motivo se puede asumir que el estimador de esta variable puede estar influenciada por la existencia de multicolinealidad imperfecta generada por esta variable, pero como no es mayor a 30 no supone ser un problema que amerite plantearse un método correctivo para la multicolinealidad.

Posteriormente, se procede a mostrar las gráficas de correlación entre los regresores,

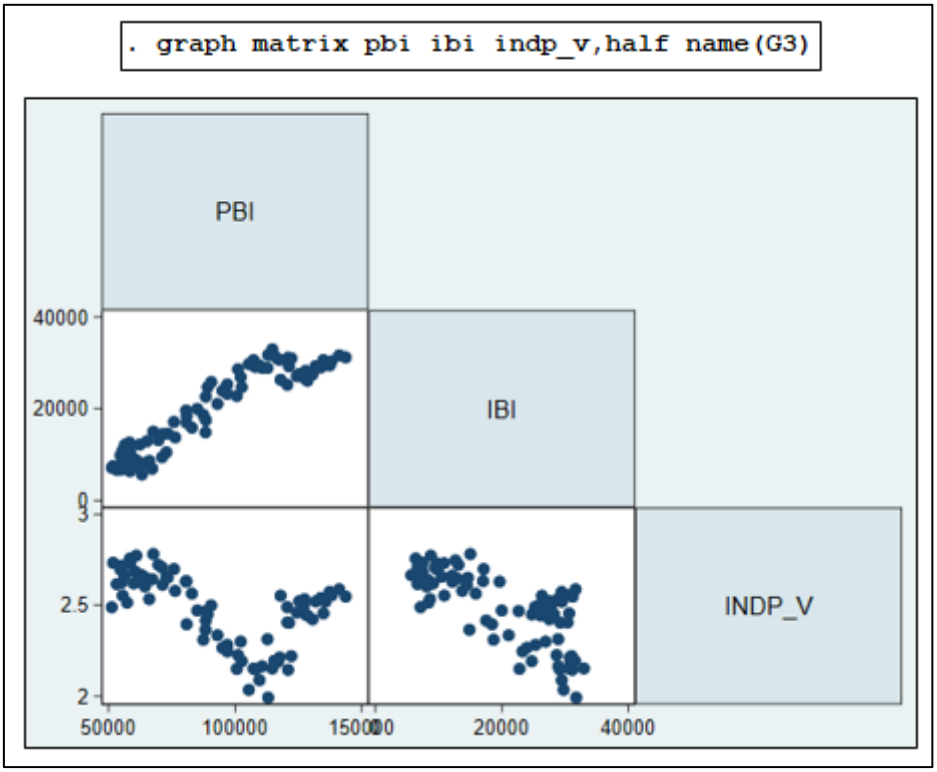


Figura 3.143. Grafica de correlación de las variables en el modelo (3.7.69.).

del modelo (3.7.69.) entre las cuales se puede visualizar que la gráfica de correlación

entre las variables ***ibi*** y ***pbi*** tiene un patrón ascendente, mientras que en las demás gráficas de correlación se muestra un patrón descendente aunque difícilmente se puede notar.

Los resultados anteriores pueden complementarse al realizarse la regla de Klein y el $R^2\ de\ Theil$. Para poder utilizar la regla de Klein se ejecuta la siguiente regresión auxiliar.

$$IBI_t = \alpha_1 + \alpha_2 PBI_t + \alpha_3 INDP_t + v_t \text{(3.7.71.)}$$

```
. reg ibi pbi indp_v

      Source |       SS           df       MS      Number of obs   =        84
-------------+----------------------------------   F(2, 81)        =    559.50
       Model |  6.4440e+09         2  3.2220e+09   Prob > F        =    0.0000
    Residual |   466450633        81  5758649.79   R-squared       =    0.9325
-------------+----------------------------------   Adj R-squared   =    0.9308
       Total |  6.9104e+09        83  83258236.7   Root MSE        =    2399.7

------------------------------------------------------------------------------
         ibi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |    .2497155   .0111709    22.35   0.000      .227489    .271942
       indp_v |   -12916.06   1587.774    -8.13   0.000    -16075.24   -9756.892
       _cons |    29084.03   4601.099     6.32   0.000     19929.28    38238.77
------------------------------------------------------------------------------
```

Figura 3.144. Resultado del modelo (3.7.71.).

Aplicando la regla de Klein se puede notar que el coeficiente de determinación del modelo auxiliar (3.7.71.) es muy cercano al coeficiente de determinación del modelo original (3.7.69.), entonces se intuye que el modelo original posiblemente tiene el estimador de la variable ***ibi*** influenciado por la presencia de multicolinealidad.

Ahora, para utilizar el efecto del $R^2\ de\ Theil$ se realizarán las siguientes regresiones auxiliares.

$$IMP_t = \alpha_1 + \alpha_2 PBI_t + \alpha_3 INDP_t + v_t \text{ (3.7.72.)}$$

$$IMP_t = \alpha_1 + \alpha_2 IBI_t + \alpha_3 INDP_t + v_t \text{ (3.7.73.)}$$

$$IMP_t = \alpha_1 + \alpha_2 PBI_t + \alpha_3 IBI_t + v_t \text{ (3.7.74.)}$$

```
. reg imp pbi indp_v
```

Source	SS	df	MS	Number of obs	=	84
				F(2, 81)	=	2037.58
Model	6.7431e+09	2	3.3715e+09	Prob > F	=	0.0000
Residual	134028563	81	1654673.62	R-squared	=	0.9805
				Adj R-squared	=	0.9800
Total	6.8771e+09	83	82856574.5	Root MSE	=	1286.3

imp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pbi	.2811186	.005988	46.95	0.000	.2692044	.2930329
indp_v	-8184.675	851.1077	-9.62	0.000	-9878.112	-6491.238
_cons	16426.13	2466.365	6.66	0.000	11518.84	21333.42

Figura 3.145. Resultado del modelo (3.7.72.).

```
. reg imp ibi indp_v
```

Source	SS	df	MS	Number of obs	=	84
				F(2, 81)	=	533.13
Model	6.3916e+09	2	3.1958e+09	Prob > F	=	0.0000
Residual	485540187	81	5994323.3	R-squared	=	0.9294
				Adj R-squared	=	0.9277
Total	6.8771e+09	83	82856574.5	Root MSE	=	2448.3

imp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ibi	.9926932	.042338	23.45	0.000	.9084539	1.076932
indp_v	1996.494	1930.722	1.03	0.304	-1845.036	5838.024
_cons	-2839.833	5437.411	-0.52	0.603	-13658.58	7978.91

Figura 3.146. Resultado del modelo (3.7.73.).

```
. reg imp pbi ibi
```

Source	SS	df	MS	Number of obs	=	84
				F(2, 81)	=	1647.46
Model	6.7121e+09	2	3.3560e+09	Prob > F	=	0.0000
Residual	165005037	81	2037099.22	R-squared	=	0.9760
				Adj R-squared	=	0.9754
Total	6.8771e+09	83	82856574.5	Root MSE	=	1427.3

imp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pbi	.1992736	.0157296	12.67	0.000	.1679765	.2305706
ibi	.3794733	.0490265	7.74	0.000	.2819259	.4770207
_cons	-3945.345	648.4956	-6.08	0.000	-5235.648	-2655.042

Figura 3.147. Resultado del modelo (3.7.74.).

Con los respectivos coeficientes de determinación de cada modelo auxiliar se calcula el $R^2\ de\ Theil$.

$$R^2\ de\ Theil = 0.9825 - (0.9825 - 0.9805) - (0.9825 - 0.9294) - (0.9825 - 0.9760) = 0.9209 \text{ (3.7.75.)}$$

El $R^2\ de\ Theil$ es muy cercano al coeficiente de determinación del modelo original, entonces se entiende que la multicolinealidad está presente pero no necesariamente está influyendo de sobremanera.

Por último, se realizará el contraste mediante la prueba de hipótesis usando la prueba *F* del modelo (3.7.71.).

$$H_0: No\ existe\ multicolinealidad$$

$$H_1: Existe\ multicolinealidad$$

El estadístico *F* calculado se halla mediante.

$$Fc = \frac{R_i^2/(k-2)}{(1-R_i^2)/(n-k+1)} = \frac{0.9325/(3-2)}{(1-0.9325)/(84-3+1)} = 1132.81 \text{ (3.7.76.)}$$

Mientras que el estadístico *F* tabulado se puede hallar en STATA mediante la instrucción **disp** señalando sus grados de libertad y el nivel de significancia del 5%.

```
. disp invF(3-2,84-3+1,1-0.05)
3.9573883
```

Figura 3.148. Resultado del modelo (3.7.71.).

$$Ft^2_{82,0.05} = 3.95 \text{ (3.7.77.)}$$

En vista que $|Fc| > Ft$ entonces se rechaza la hipótesis nula y se concluye que existe multicolinealidad en el modelo original.

A través de las diversas pruebas de multicolinealidad se ha diagnosticado que el modelo original presenta multicolinealidad. No obstante, los índices de VIF manifiesta que los estimadores pueden estar influenciados por la presencia de multicolinealidad pero no supone que sean altamente inestables, por lo que se concluye que no es necesario plantearse un método correctivo para tratar la presencia de multicolinealidad en el modelo.

- **Homocedasticidad.**

A continuación, se hará uso de los métodos informales y formales para probar si el modelo cumple el supuesto de homocedasticidad, en caso contrario, se planteará cuál debe ser la medida correctiva a seguir.

Los métodos informales para detectar la presencia de heterocedasticidad en el modelo, los conforman el gráfico de dispersión entre los residuos del modelo con los valores de la variable dependiente estimada, y también los gráficos de dispersión entre los residuos del modelo con los valores de las variables explicativas del modelo.

Los métodos formales están conformados, por las pruebas de hipótesis testeadas mediante los métodos de BG y de White.

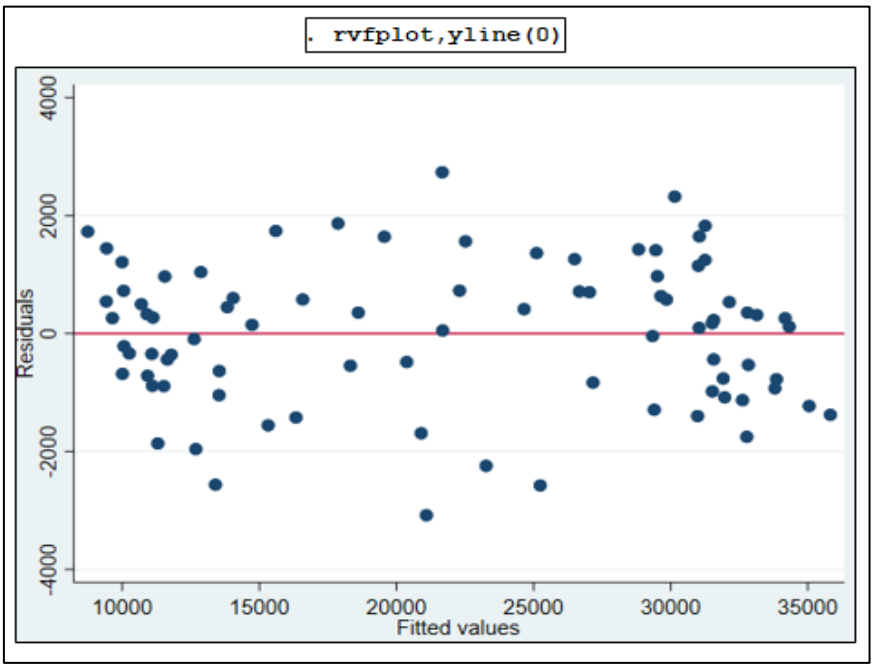


Figura 3.149. Gráfico de dispersión entre los residuos y los valores estimados de la variable dependiente del modelo (3.7.69.).

Esta gráfica muestra cómo están distribuidos los residuos con los valores estimados de la variable dependiente, y no se vislumbra ningún patrón ni mucho menos ningún dato atípico. Por lo que, a simple vista se puede pensar que el modelo no presente heterocedasticidad.

Veamos cuales son los gráficos de dispersión entre los residuos y los valores de las regresoras.

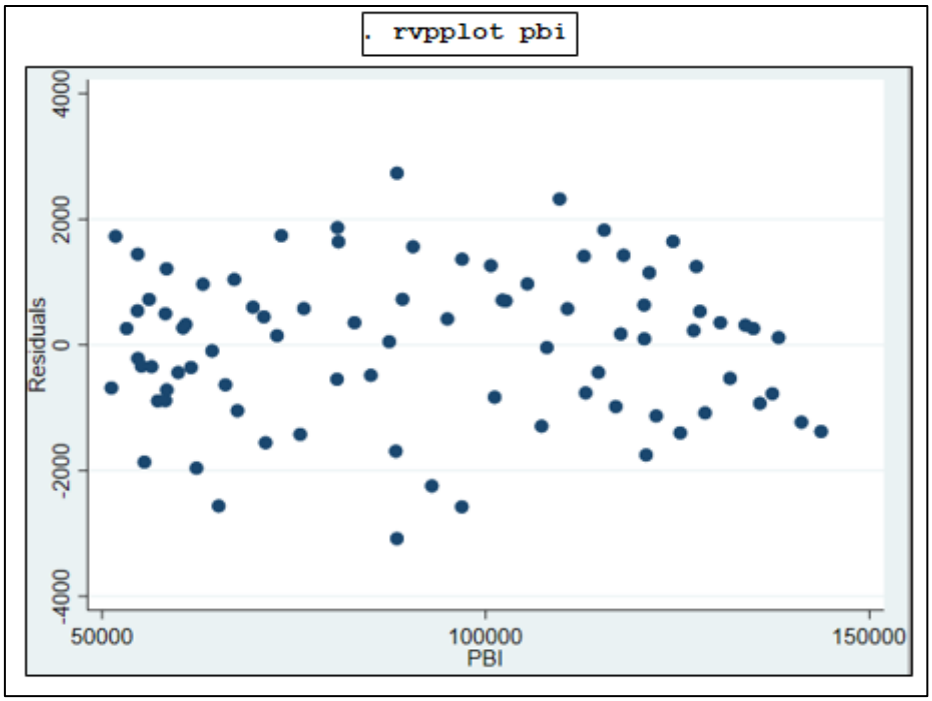


Figura 3.150. Gráfico de dispersión entre los residuos y los valores de la variable regresora *pbi* del modelo (3.7.69.).

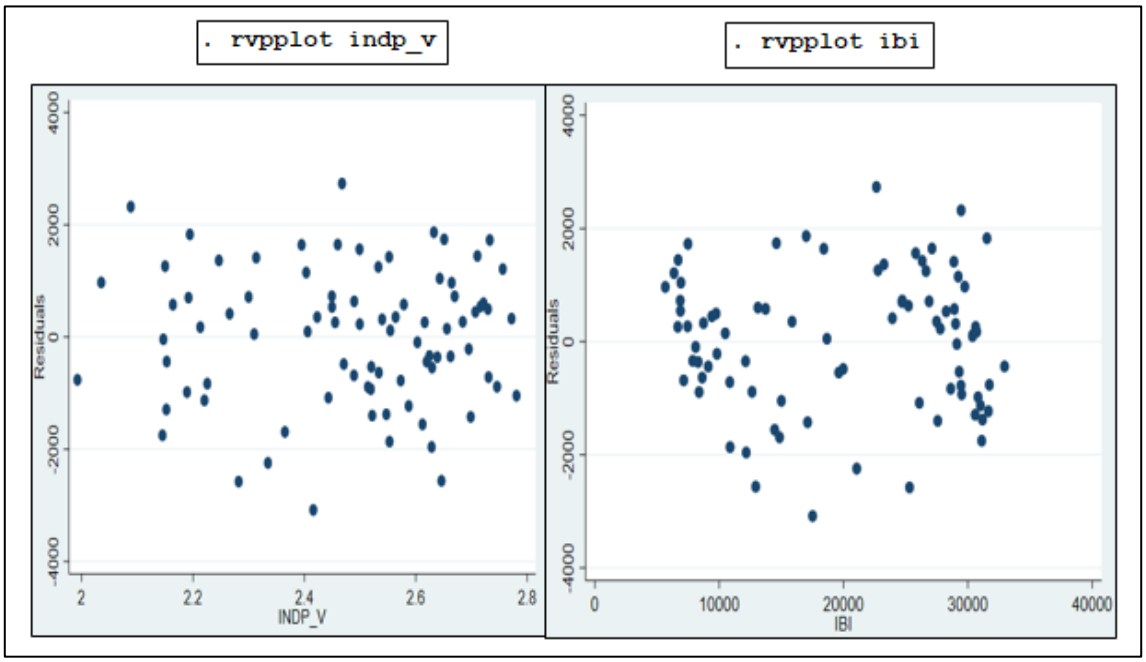


Figura 3.151. Gráfico de dispersión entre los residuos y los valores de la variable regresora *indp_v* y *ibi* del modelo (3.7.69.).

Los gráficos de dispersión entre los residuos y los valores de las regresoras no muestran la existencia de ningún patrón, ni tampoco la existencia de datos atípicos. En

consecuencia, se puede intuir que los regresoras no causan heterocedasticidad. Cabe mencionar que estas gráficas también se utilizan como métodos informales para diagnosticar autocorrelación, entonces se puede argumentar que el modelo posiblemente está libre de heterocedasticidad y autocorrelación.

Estas pruebas informales contrastadas mediante gráficas, serán corroboradas por los resultados que se obtengan de los métodos formales mediante el contraste de hipótesis por las pruebas BG y de White. Ambas pruebas siguen la siguiente prueba de hipótesis.

$.H_0: No\ existe\ heterocedasticidad.$

$$H_1: Existe\ heterocedasticidad.$$

Los resultados de la prueba BG se muestran en la siguiente figura.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of imp

         chi2(1)      =     0.01
         Prob > chi2  =   0.9147

. estat hettest pbi indp_v ibi,mtest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance

-------------------------------------------
    Variable |      chi2    df      p
-------------+-----------------------------
         pbi |      0.02     1    0.8896 #
       indp_v |      1.07     1    0.3018 #
         ibi |      0.19     1    0.6669 #
-------------+-----------------------------
simultaneous |      2.71     3    0.4387
-------------------------------------------
                  # unadjusted p-values
```

Figura 3.152. Prueba de BG del modelo (3.7.69.).

La prueba de BG nos permite aceptar la hipótesis nula y asumir que no existe heterocedasticidad en el modelo.

Para reafirmar que no existe heterocedasticidad en el modelo, se contrastará mediante la prueba de White.

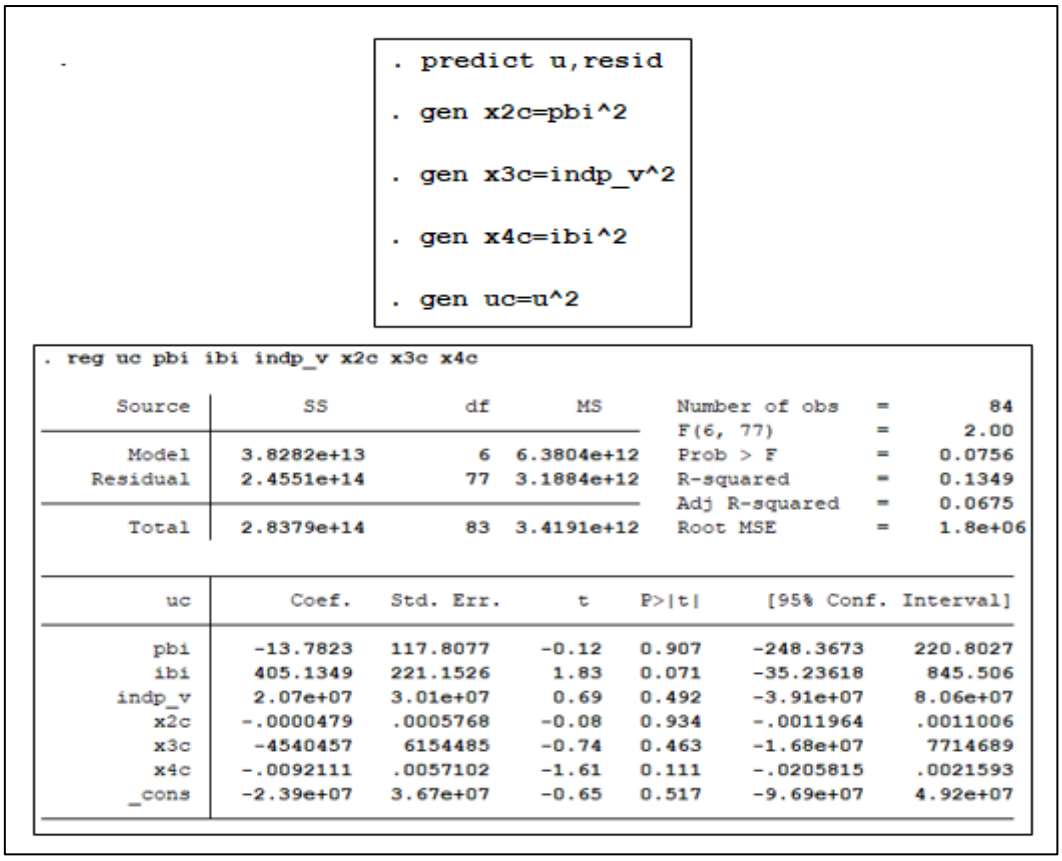
```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =     11.76
         Prob > chi2  =    0.2273

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	11.76	9	0.2273
Skewness	4.64	3	0.2002
Kurtosis	1.12	1	0.2896
Total	17.52	13	0.1766

Figura 3.153. Prueba general de White de heterocedasiticidad del modelo (3.7.69.).

La prueba general de heterocedasticidad de White nos indica que debemos aceptar la hipótesis nula y por tanto asumir que el modelo está libre de heterocedasticidad.

Esta prueba puede ser ratificada mediante la prueba pura de heterocedasticidad de White, que será contrastada usando los cuadrados de las regresoras y utilizando una regresión auxiliar, donde el cuadrado de los residuos será la variable dependiente y estará explicada por las regresoras y los cuadrados de las regresoras.



```
. predict u,resid

. gen x2c=pbi^2

. gen x3c=indp_v^2

. gen x4c=ibi^2

. gen uc=u^2
```

```
. reg uc pbi ibi indp_v x2c x3c x4c
```

Source	SS	df	MS	Number of obs	=	84
				F(6, 77)	=	2.00
Model	3.8282e+13	6	6.3804e+12	Prob > F	=	0.0756
Residual	2.4551e+14	77	3.1884e+12	R-squared	=	0.1349
				Adj R-squared	=	0.0675
Total	2.8379e+14	83	3.4191e+12	Root MSE	=	1.8e+06

uc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pbi	-13.7823	117.8077	-0.12	0.907	-248.3673	220.8027
ibi	405.1349	221.1526	1.83	0.071	-35.23618	845.506
indp_v	2.07e+07	3.01e+07	0.69	0.492	-3.91e+07	8.06e+07
x2c	-.0000479	.0005768	-0.08	0.934	-.0011964	.0011006
x3c	-4540457	6154485	-0.74	0.463	-1.68e+07	7714689
x4c	-.0092111	.0057102	-1.61	0.111	-.0205815	.0021593
_cons	-2.39e+07	3.67e+07	-0.65	0.517	-9.69e+07	4.92e+07

Figura 3.154. Prueba pura de White de heterocedasiticidad del modelo (3.7.69.).

Se obtiene el siguiente estadístico calculado.

$n * R^2 = 84 * 0.1349 = 11.33$ (3.7.78.)

Y el siguiente estadístico tabulado.

$X^2_{6,0.05} = 12.59$ (3.7.79.)

Después de haber hallado el estadístico calculado y el estadístico tabulado, se observa que el primero es menor al segundo, entonces no se rechaza la hipótesis nula y se asume que según la prueba de heterocedasticidad pura de White, el modelo (3.7.69.) está libre de heterocedasticidad.

Luego de haber contrastado, mediante los métodos formales e informales, y de no haber encontrado inferencias que nos indiquen que el modelo esté violando el supuesto de homocedasticidad; entonces, no se ejecutará ningún método correctivo para tratar la heterocedasticidad.

- **No autocorrelación.**

Los métodos informales para detectar la autocorrelación son los gráficos de dispersión vistos en las figuras 3.149., 3.150. Y 3.151., y como ya se había dicho anteriormente, su interpretación indica que el modelo está libre de autocorrelación. No obstante, siguiendo la teoría que proponen (Gujarati & Porter, 2010) debemos graficar los residuos en una gráfica de línea, la cual será realizada mediante la instrucción **tsline**. La variable ***u*** representa a los residuos, como se puede ver en la figura 3.154.

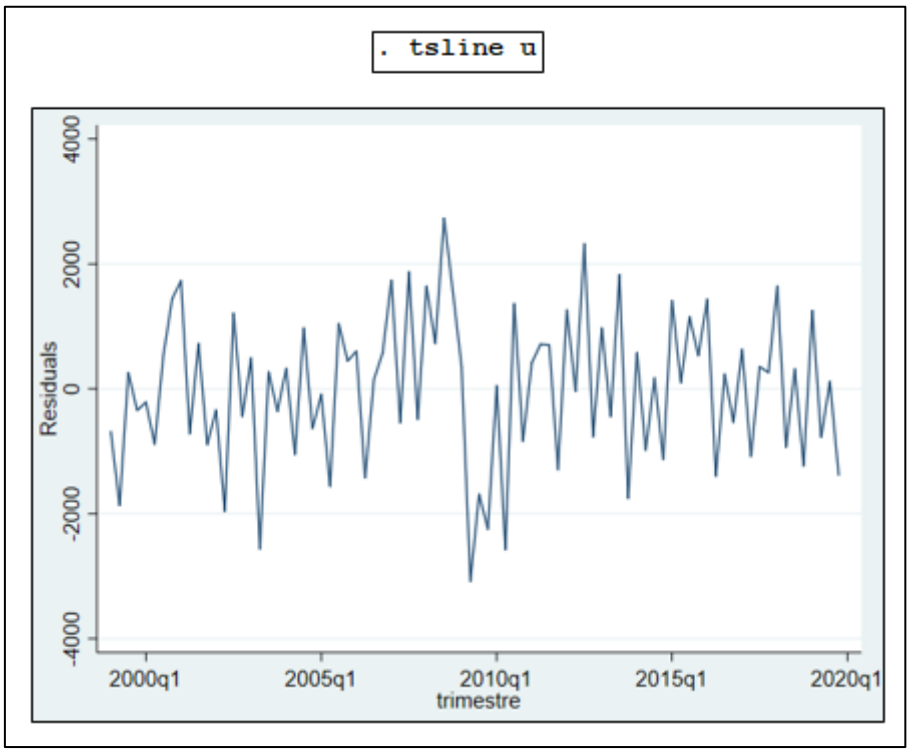


Figura 3.155. Grafica de dispersión de los residuos del modelo (3.7.69.).

La gráfica de línea en la figura 3.155. Muestra la evolución en el tiempo de los residuos. Aparentemente, no existe una tendencia, salvo por una: los valores de los trimestres de los años 2008 y 2009. Según la gráfica, se puede sospechar que el modelo está libre de autocorrelación. Además, (Gujarati & Porter, 2010) Recomiendan una gráfica de línea de los residuos estandarizados con respecto al tiempo.

En STATA, los residuos estandarizados se generan de forma similar a los residuos, con la ayuda del comando **predict** y la opción **rstandard**. Cabe recalcar, que el comando **predict** es un comando de postestimación que toma la última regresión calculada con el comando **reg**, por lo que, se debe realizar después de la regresión del modelo (3.7.69.) para generar los residuos y/o los valores que se requieran hallar con el comando **predict**.

Veamos los pasos para realizar una gráfica de línea los residuos estandarizados.

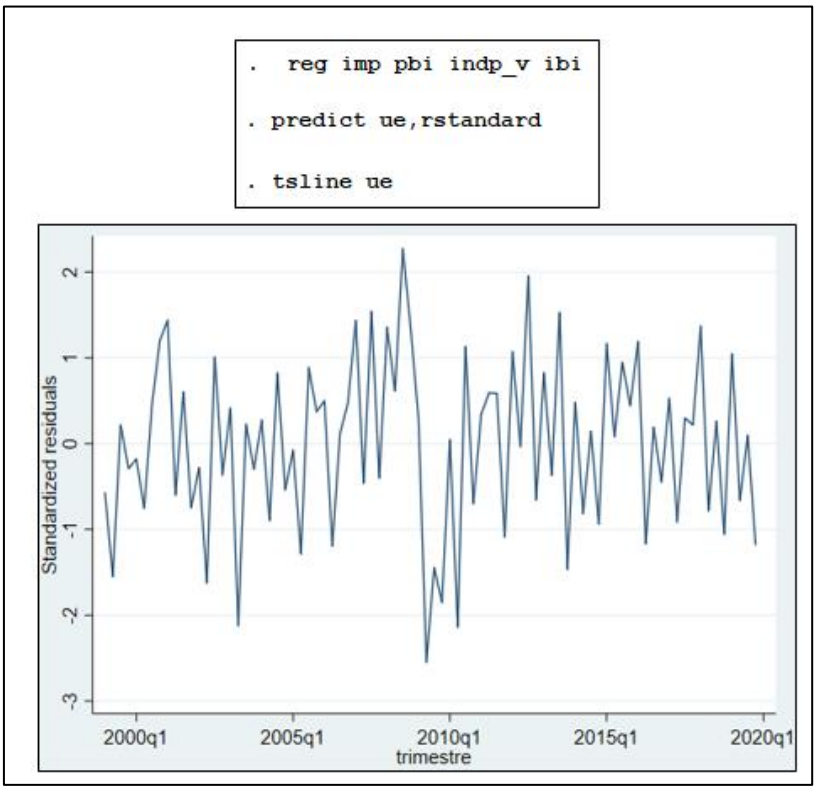


Figura 3.156. Grafica de dispersión de los residuos estandarizados del modelo (3.7.69.).

De forma similar a la anterior gráfica, en la gráfica de línea de los residuos estandarizados no se observa ningún patrón de tendencia. Según (Gujarati & Porter, 2010) El diagnóstico visual de ambas gráficas se interpretan como la ausencia de autocorrelación en el modelo.

Las variables económicas tienden a estar correlacionadas entre sí con sus valores pasados, entonces no se puede confiar plenamente en estos diagnósticos visuales. Por esta razón, se deben ejecutar los métodos formales con el fin de realizar un diagnóstico concluyente de autocorrelación.

El modelo que se elegirá para explicar la posible autocorrelación será un *AR(1)*, como se ve en el siguiente modelo.

$$\hat{\mu}_t = p\hat{\mu}_{t-1} + e_t \quad (3.7.80.)$$

El modelo (3.7.80.) indica que los residuos del modelo original (3.7.69.) siguen un esquema $AR(1)$, lo cual asume que los residuos están correlacionados con sus valores pasados un periodo. Si el p es cercano a 0, entonces no existe autocorrelación en el modelo, por otro lado, si p es cercano a 1 o -1, entonces la autocorrelación puede ser positiva o negativa respectivamente. En la siguiente figura se muestran los resultados de tal esquema *AR(1)* mediante MCO.

```
. reg u L.u,noconstant

      Source |       SS           df       MS      Number of obs   =        83
-------------+----------------------------------   F(1, 82)        =      1.84
       Model |  2628647.34         1  2628647.34   Prob > F        =    0.1787
    Residual |   117164669        82  1428837.43   R-squared       =    0.0219
-------------+----------------------------------   Adj R-squared   =    0.0100
       Total |   119793317        83  1443292.97   Root MSE        =    1195.3

------------------------------------------------------------------------------
           u |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           u |
         L1. |   -.149026    .109872    -1.36   0.179    -.3675964    .0695445
------------------------------------------------------------------------------
```

Figura 3.157. Resultados del esquema de *AR(1)* de los residuos del modelo (3.7.69.).

Según la figura anterior, $p = -0.14$, entonces, al estar *p* cerca de 0 **es posible que el esquema *AR(1)* indique que los residuos no están correlacionados con sus valores retardados un período**, lo que significa que no existe autocorrelación en los residuos del modelo original con respecto a sus valores rezagados un período. En el caso que los residuos y los residuos rezagados un período estén correlacionados, esta correlación sería negativa.

Analicemos la instrucción que se ve en la figura 3.157. Se observa que la regresión del modelo $AR(1)$ se ha calculado con el comando **reg** y la variable ***u***, la cual representa a los valores de los residuos y es la variable dependiente en la regresión, por otro lado, se ha utilizado al operador **L.** como la preinstrucción que indica a STATA que tome el primer rezago de la variable ***u*** como variable independiente y la opción **noconstant** ordena que no calcule el término constante. De la figura, se estima el siguiente resultado

$$\hat{\mu}_t = -0.14\hat{\mu}_{t-1} + e_t \text{ (3.7.81.)}$$

En este modelo la significancia global, el coeficiente de determinación y el error estándar del estimador p no son tan relevantes para el análisis. Sin embargo, la significancia individual del estimador p si debería ser revisada.

Si se requiere realizar una regresión utilizando un esquema *AR(2)* para explicar la autocorrelación de los residuos, entonces se añade la preinstrucción **L2.**

```
. reg u L.u L2.u, noconstant
```

Source	SS	df	MS	Number of obs	=	82
				F(2, 80)	=	9.56
Model	22436137.5	2	11218068.8	Prob > F	=	0.0002
Residual	93879168	80	1173489.6	R-squared	=	0.1929
				Adj R-squared	=	0.1727
Total	116315306	82	1418479.34	Root MSE	=	1083.3

u	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
u						
L1.	-.0997467	.1008798	-0.99	0.326	-.3005039	.1010105
L2.	.4093958	.1006857	4.07	0.000	.2090249	.6097668

Figura 3.158. Resultados del esquema de *AR(2)* de los residuos del modelo (3.7.69.).

¿Qué indica la figura 3.158.? Esta figura señala, que los residuos rezagados dos periodos pueden tener correlación positiva con los residuos del modelo, si tomamos en cuenta que $\hat{p}_2$ presenta significancia individual y $\hat{p}_1$ no es significativo. No obstante, al no estar lo suficientemente cercano a 0, no es concluyente. Estos resultados se representan como.

$$\hat{\mu}_t = -0.09\hat{\mu}_{t-1} + 0.41\hat{\mu}_{t-2} + e_t \text{ (3.7.82.)}$$

Estos esquemas autorregresivos utilizados para explicar la autocorrelación en el modelo, pueden ser corroborados graficando la dispersión entre los residuos con sus respectivos p rezagos. En este caso, existe sospecha que haya una posible autocorrelación entre sus residuos y los residuos rezagos uno o dos periodos.

Generamos sus valores rezagados respectivos con el comando **gen** y el operador **L.** y **L2.** para indicar que se están requiriendo la generación del primer y segundo retardo de los residuos, respectivamente. Posteriormente, con el comando **scatter** ordenamos a STATA que realice la gráfica de dispersión entre las variables. Se espera a que no exista ningún patrón en las gráficas de dispersión, ya que de no existir autocorrelación en el

modelo entonces no tiene que notarse la existencia de correlación, ni de forma positiva ni negativa.

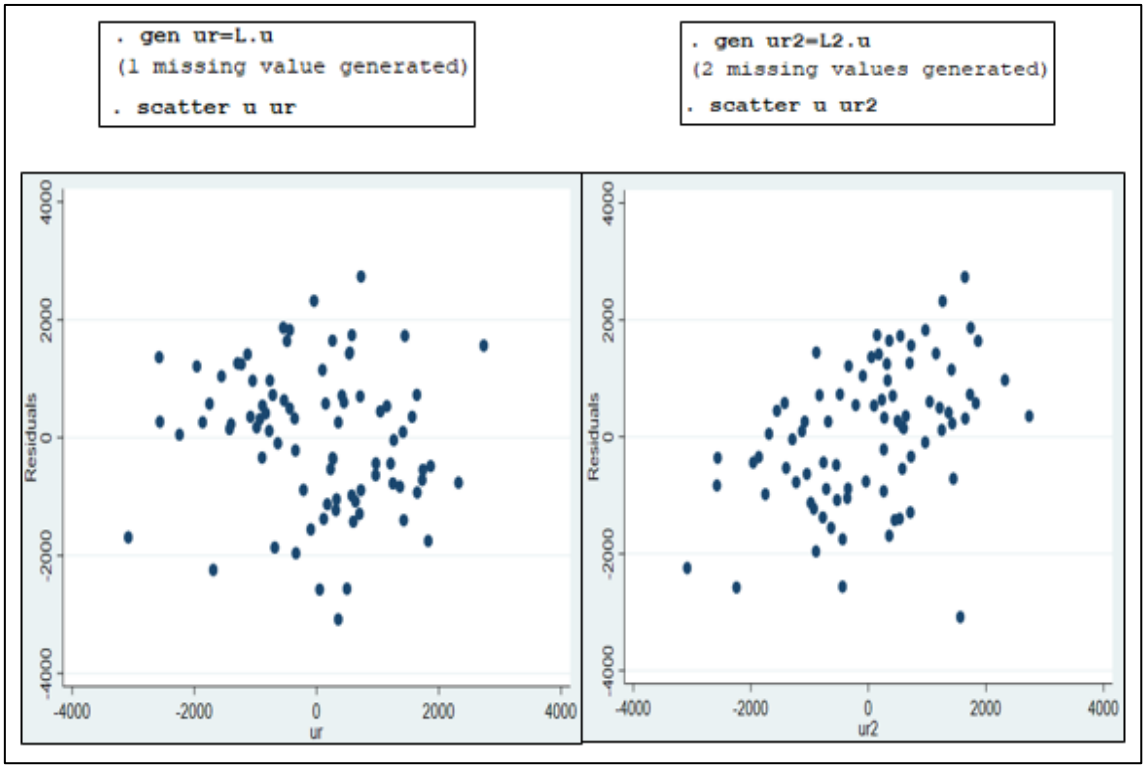


Figura 3.159. Gráficos de dispersión entre los residuos del modelo (3.7.69.) y sus residuos rezagados uno y dos periodos.

En la figura 3.159. Se muestran dos gráficos, a la izquierda se encuentra la gráfica de dispersión entre los residuos y sus valores rezagados un período y a la derecha se encuentra la gráfica de dispersión entre los residuos y sus valores rezagados dos periodos. En la gráfica de la izquierda se encuentra una nube de dispersión ligeramente decreciente, mientras que en la gráfica de la derecha es más notorio un patrón ascendente, entonces se intuye que es posible que el modelo tenga residuos autocorrelacionados en su segundo período.

Para confirmar las sospechas que se han obtenido de las gráficas, se deben realizar los métodos formales mediante la prueba de hipótesis con los contrastes de Durbin-Watson, alternativo de Durbin y de Breusch-Godfrey (BG). Estos son los contrastes de hipótesis.

$$H_0: No\ existe\ autocorrelación$$

$$H_1: Existe\ autocorrelación$$

Recordemos que para aplicar la test de Durbin los modelos deben cumplir los siguientes requerimientos:

- El modelo original debe incluir el intercepto.
- La muestra usada del modelo original no debe tener datos faltantes.
- Las regresoras del modelo no son variables estocásticas.
- Ml modelo no incluye a la variable dependiente rezagada como regresora y los residuos siguen un $AR(1)$.

Este último requerimiento significa que la prueba Durbin-Watson solamente comprueba si los residuos están correlacionados con sus valores rezagados un periodo. El comando **dwatson** muestra el resultado del contraste de Durbin-Watson y como comando de postestimación, debe ser introducido después de realizar la regresión del modelo original.

```
. reg imp pbi ibi indp_v

. estat dwatson

Durbin-Watson d-statistic(  4,    84) =  2.273641
```

Figura 3.160. Prueba de Durbin-Watson (1).

La prueba de Durbin-Watson indica que $dw = 2.27$, por lo que al ser cercano a 2 no rechazamos la hipótesis nula y asumimos que el modelo no tiene autocorrelación en el esquema autorregresivo de primer orden. En la tabla de Durbin-Watson se encuentran los siguientes estadísticos tabulados que corresponden a los límites superior e inferior, $d_U = 1.721$ $d_L = 1.575$. Con estos estadísticos se construye el siguiente diagrama.

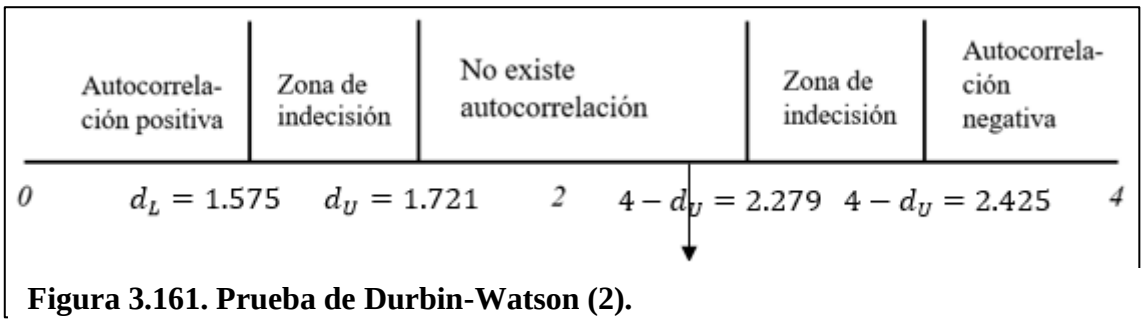


Figura 3.161. Prueba de Durbin-Watson (2).

En la figura, se observa $d_U < dw < d_L$, por lo que no se rechaza la hipótesis nula y se asume que no existe autocorrelación en el modelo siguiendo un esquema $AR(1)$, según el contraste de Durbin-Wattson.

En el caso que el modelo original incluya a la variable dependiente rezagada como una regresora, entonces la prueba de Durbin-Watson hubiera arrojado un resultado equivocado. En estos modelos la prueba para determinar si existe o no autocorrelación es la prueba alternativa de Durbin, que se puede ejecutar en STATA con el comando **durbinalt**. La prueba alternativa de Durbin se realiza después de ejecutar la regresión del siguiente modelo econométrico.

$$IMP_t = \hat{\beta}_1 + \hat{\beta}_2 PBI_t + \hat{\beta}_3 INDP_t + \hat{\beta}_4 IBI_t + \hat{\beta}_5 IMP_{t-1} + \hat{\mu}_t \text{ (3.7.83.)}$$

La regresión del modelo (3.7.83.) se realiza con el comando **reg** y usando el operador **L.** en la variable dependiente, como una regresora en la instrucción ordenada en STATA.

```
. reg imp pbi ibi indp_v L.imp
```

Source	SS	df	MS	Number of obs	=	83
				F(4, 78)	=	1879.47
Model	6.6470e+09	4	1.6618e+09	Prob > F	=	0.0000
Residual	68964751.3	78	884163.478	R-squared	=	0.9897
				Adj R-squared	=	0.9892
Total	6.7160e+09	82	81902343.4	Root MSE	=	940.3

imp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pbi	.1174554	.0198251	5.92	0.000	.0779866	.1569241
ibi	.1096735	.0445332	2.46	0.016	.0210146	.1983324
indp_v	-3133.234	933.3393	-3.36	0.001	-4991.37	-1275.098
imp L1.	.4824277	.0636465	7.58	0.000	.3557172	.6091382
_cons	6300.532	2352.236	2.68	0.009	1617.59	10983.47

Figura 3.162. Resultados de la regresión (3.7.83.).

Con este resultado se hace la prueba alternativa de Durbin.

```
. estat durbinal

Durbin's alternative test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	0.488	1	0.4850

H0: no serial correlation

Figura 3.163. Resultados de la prueba alternativa de Durbin del modelo (3.7.83.).

La prueba alternativa de Durbin indica que el valor-p es mayor a una significancia del 5%, por lo que se acepta la hipótesis nula y el modelo (3.7.83.) estaría libre de autocorrelación.

Estas pruebas formales solo nos han permitido conocer si los residuos del modelo no están correlacionados con sus valores rezagados un periodo, sin embargo, anteriormente hemos notado en las gráficas la posibilidad que los residuos estén correlacionados con sus valores rezagados en dos periodos. Es necesario realizar la prueba de BG para conocer si los residuos están correlacionados con sus propios valores en uno o más periodos rezagados. El comando que se utiliza para realizar esta prueba en STATA es **estat bgodfrey**, y ya que se pide contrastar si los residuos dependen de sus valores rezagados dos periodos se usará la opción **lags**, cuya función es indicar a STATA el número de rezagos que se quiere contrastar. Al tratarse de un comando de postestimación, volveremos a ejecutar la regresión del modelo (3.7.69.).

```
. reg imp pbi ibi indp_v
```

```
. estat bgodfrey, lags(1 2)

Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	2.399	1	0.1214
2	15.633	2	0.0004

Figura 3.164. Resultados de la prueba de BG para el modelo (3.7.69.).

Observando los valores-p que se ven en la figura, podemos argumentar que este contraste indica el mismo resultado que muestra la prueba de Durbin-Watson. Sin embargo, la prueba de BG también está señalando que efectivamente los residuos están correlacionados con sus valores pasados rezagados en dos periodos. En conclusión, esta prueba indica que los residuos del modelo (3.7.69.) siguen un esquema $AR(2)$, entonces el modelo presenta autocorrelación. ¿Qué significa que el modelo tenga autocorrelación en dos periodos rezagados? Significa que los residuos provenientes del modelo (3.7.69.) dependen de sus propios valores rezagados dos periodos.

Debido a que es posible que los estimadores estén afectados por la existencia de autocorrelación en el modelo, se debería aplicar el método indicado para corregir la violación del supuesto de no autocorrelación.

STATA permite corregir la presencia de autocorrelación en un modelo mediante el estimador de MCF de C-O en dos pasos con el comando **prais** y las opciones **corc** y **twostep**, la primera opción indica a STATA que ejecute el método de C-O y la segunda opción ordena que el proceso iterativo de C-O sea en dos pasos. En el caso que se requiere recuperar el primer dato mediante la transformación de P-W se debe mantener **twosetp**. En las siguientes figuras se muestran los resultados de ambos métodos correctivos.

```
. prais imp pbi indp_v ibi, corc twostep

Iteration 0:  rho = 0.0000
Iteration 1:  rho = -0.1490

Cochrane-Orcutt AR(1) regression -- twostep estimates

      Source |       SS       df       MS              Number of obs =      83
-------------+------------------------------           F(3, 79)      = 1990.05
       Model |  8.7325e+09     3  2.9108e+09           Prob > F      =  0.0000
    Residual |  115553032     79  1462696.61           R-squared     =  0.9869
-------------+------------------------------           Adj R-squared =  0.9864
       Total |  8.8481e+09    82   107903185           Root MSE      =  1209.4

------------------------------------------------------------------------------
         imp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |    .2253674   .0142306    15.84   0.000      .197042    .2536928
       indp_v |   -5227.619   1015.131    -5.15   0.000    -7248.188   -3207.051
         ibi |    .2274967   .0542763     4.19   0.000     .1194625    .3355309
       _cons |    9673.384   2650.997     3.65   0.000     4396.707    14950.06
-------------+----------------------------------------------------------------
         rho |   -.149026
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    2.273641
Durbin-Watson statistic (transformed) 2.040507
```

Figura 3.105. Resultados del método de corrección de C-O en dos pasos del modelo (3.7.69.).

En la parte superior de los resultados del comando se pueden observar el valor de cada p según el número de iteraciones, en este caso, solamente hay una iteración cuyo p es -0.14 la cual corresponde al esquema $AR(1)$ del modelo original.

$$\hat{\mu}_t = -0.14\hat{\mu}_{t-1} + e_t \text{ (3.7.81.)}$$

Entonces, con (3.7.81.) se construye el siguiente modelo.

$$IMP_t - pIMP_{t-1} = (1-p)\hat{\beta}_1 + \hat{\beta}_2(PBI_t - pPBI)_{t-1} + \hat{\beta}_3(INDP_t - pINDP_{t-1}) + \hat{\beta}_4(IBI_t - pIBI_{t-1}) + e_t \text{ (3.7.84.)}$$

$$IMP_t^* = \hat{\beta}_1 + \hat{\beta}_2 PBI_t^* + \hat{\beta}_3 INDP_t^* + \hat{\beta}_4 IBI_t^* + e_t \text{ (3.7.85.)}$$

También se puede ver la expresión “Cochrane-Orcutt”, indicado el método correctivo al cual se le atribuyen los resultados hallados. Observando la figura, los resultados estimados de (3.7.85.) son:

$$IMP_t^* = 9673.39 + 0.23PBI_t^* - 5227.62INDP_t^* + 0.23IBI_t^* + e_t \text{ (3.7.86.)}$$

$$ee = (2650.99) \quad (0.014) \quad (1015.13) \quad (0.054)$$

$$t = 3.65 \quad 15.84 \quad -5.15 \quad 4.19$$

Finalmente, en la parte más inferior de la figura se observan los estadísticos calculados de dw del modelo original (3.7.69.9) y transformado (3.7.86.). En los cuales, se observa que, el estadístico calculado dw del modelo original es mayor al transformado. ¿Qué significa que el estadístico calculado dw del modelo original sea mayor al estadístico calculado dw del modelo transformado? Recordemos que, el modelo original no tiene autocorrelación en un periodo rezagado y tanto los métodos C-O y P-W solo corrigen a los modelos que tienen autocorrelación en el primer rezago, entonces es probable que (3.7.86.) no esté corrigiendo la autocorrelación. Además, el método C-O pierde la primera observación, apliquemos el método P-W que recupera la primera observación.

```
. prais imp pbi indp_v ibi, twostep

Iteration 0:  rho = 0.0000
Iteration 1:  rho = -0.1490

Prais-Winsten AR(1) regression -- twostep estimates

      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(3, 80)      = 2070.20
       Model |  8.9890e+09     3  2.9963e+09           Prob > F      =  0.0000
    Residual |   115789867    80  1447373.34           R-squared     =  0.9873
-------------+------------------------------           Adj R-squared =  0.9868
       Total |  9.1048e+09    83   109696714           Root MSE      =  1203.1

------------------------------------------------------------------------------
         imp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |   .2250473   .0141338    15.92   0.000     .1969202    .2531744
      indp_v |  -5166.667   998.4949    -5.17   0.000    -7153.735   -3179.599
         ibi |   .2301378    .053595     4.29   0.000     .1234804    .3367952
       _cons |   9494.31   2599.651     3.65   0.000      4320.84    14667.78
-------------+----------------------------------------------------------------
         rho |  -.149026
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    2.273641
```

Figura 3.166. Resultados del método de corrección de P-W del modelo (3.7.69.).

$$IMP_t^* = 9494.31 + 0.23PBI_t^* - 5166.66INDP_t^* + 0.23IBI_t^* + e_t \text{ (3.7.87.)}$$

$$ee = (2599.65) \quad (0.014) \quad (998.50) \quad (0.054)$$

$$t = 3.65 \quad 15.92 \quad -5.17 \quad 4.29$$

Con el comando **prais** y la opción **twosetp** se logra obtener los resultados recuperando la primera observación perdida en el método anterior con el método P-W. Por lo general, el método de P-W brinda mejores resultados que el método C-O debido a que el método P-W logra recuperar la primera observación.

Se puede notar en que el estadístico calculado *dw* de P-W es ligeramente mayor al estadístico calculado *dw* que muestra el método de C-O. No obstante, es posible que este método no esté corrigiendo la autocorrelación en el modelo original, ya que el modelo original no tiene autocorrelación en un periodo rezagado, sino en el segundo periodo rezagado.

STATA también permite el cálculo de los resultados de los métodos iterativos tanto de C-O como de P-W. En ambas instrucciones excluimos la opción **twosetp** y STATA generará tantas iteraciones como crea necesario.

```
. prais imp pbi indp_v ibi, corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = -0.1490
Iteration 2:  rho = -0.2417
Iteration 3:  rho = -0.2879
Iteration 4:  rho = -0.3062
Iteration 5:  rho = -0.3127
Iteration 6:  rho = -0.3149
Iteration 7:  rho = -0.3156
Iteration 8:  rho = -0.3159
Iteration 9:  rho = -0.3159
Iteration 10:  rho = -0.3160
Iteration 11:  rho = -0.3160
Iteration 12:  rho = -0.3160
Iteration 13:  rho = -0.3160

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      83
-------------+------------------------------           F(3, 79)      = 2661.91
       Model |  1.1477e+10     3  3.8258e+09           Prob > F      =  0.0000
    Residual |   113542853    79   1437251.3           R-squared     =  0.9902
-------------+------------------------------           Adj R-squared =  0.9898
       Total |  1.1591e+10    82   141354100           Root MSE      =  1198.9

------------------------------------------------------------------------------
         imp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |      .2136   .0130921    16.32   0.000     .1875408    .2396591
       indp_v |  -4469.728   923.7429    -4.84   0.000    -6308.393   -2631.064
         ibi |   .2800068   .0505591     5.54   0.000     .1793714    .3806422
       _cons |   7824.414   2394.729     3.27   0.002     3057.825       12591
-------------+----------------------------------------------------------------
         rho |   -.315978
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    2.273641
Durbin-Watson statistic (transformed) 1.743623
```

Figura 3.167. Resultados del método de corrección del iterativo de C-O del modelo (3.7.69.).

El método iterativo de C-O brinda los siguientes resultados en sus estimadores.

$$IMP_t^{**} = 7824.41 + 0.21PBI_t^{**} - 4469.73INDP_t^{**} + 0.28IBI_t^{**} + e_t^{**} \text{ (3.7.88.)}$$

$$ee = (2394.73) \quad (0.013) \quad (923.744) \quad (0.051)$$

$$t = 3.27 \quad 16.32 \quad -4.84 \quad 5.54$$

El resultado más importante que se observa en la figura son los estadísticos calculados de *dw* tanto del modelo original (3.7.69.) como del modelo transformado (3.7.88.), cuyos valores respectivos son 2.27 y 1.74, de forma similar en los anteriores modelos transformados se intuye que el modelo (3.7.87.) no podría ser el idóneo, debido a que sigue el supuesto que en caso que haya autocorrelación en el modelo, esta debe ser de primer orden para que sea válido su uso. Por último, en la figura se muestra que STATA ha considerado hasta 13 iteraciones para encontrar el p indicado.

El método iterativo de P-W se realiza ejecutando el comando **prais** y excluyendo las

```
. prais imp pbi indp_v ibi

Iteration 0:  rho = 0.0000
Iteration 1:  rho = -0.1490
Iteration 2:  rho = -0.2468
Iteration 3:  rho = -0.2931
Iteration 4:  rho = -0.3104
Iteration 5:  rho = -0.3163
Iteration 6:  rho = -0.3181
Iteration 7:  rho = -0.3187
Iteration 8:  rho = -0.3189
Iteration 9:  rho = -0.3190
Iteration 10:  rho = -0.3190
Iteration 11:  rho = -0.3190
Iteration 12:  rho = -0.3190
Iteration 13:  rho = -0.3190

Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(3, 80)      = 2800.66
       Model |  1.1934e+10     3  3.9779e+09           Prob > F      =  0.0000
    Residual |   113627520     80    1420344           R-squared     =  0.9906
-------------+------------------------------           Adj R-squared =  0.9902
       Total |  1.2047e+10     83   145148803           Root MSE      =  1191.8

------------------------------------------------------------------------------
         imp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pbi |   .2132352   .0129735    16.44   0.000     .1874172    .2390532
       indp_v |  -4428.44   908.6202    -4.87   0.000    -6236.651   -2620.228
         ibi |   .2821536   .0498855     5.66   0.000     .1828783     .381429
       _cons |   7710.649   2350.277     3.28   0.002     3033.448    12387.85
-------------+----------------------------------------------------------------
         rho |  -.3189871
------------------------------------------------------------------------------

Durbin-Watson statistic (original)    2.273641
Durbin-Watson statistic (transformed) 1.766439
```

Figura 3.168. Resultados del método de corrección del iterativo de P-W del modelo (3.7.69.).

Con el modelo iterativo de P-W se obtienen los estimadores:

$$IMP_t^{**} = 7710.65 + 0.21PBI_t^{**} - 4428.44INDP_t^{**} + 0.28IBI_t^{**} + e_t^{**} \text{ (3.7.89.)}$$

$$ee = (2350.28) \quad (0.013) \quad (908.62) \quad (0.050)$$

$$t = 3.28 \quad 16.44 \quad -4.87 \quad 5.66$$

Y al igual que el método iterativo de C-O, el método P-W muestra un mayor estadístico calculado dw del modelo original con respecto al estadístico calculado dw del modelo transformado. Por lo que, este modelo tampoco parece ser el mejor para corregir la autocorrelación del modelo original.

Si los métodos iterativos de C-O y P-W no son válidos para corregir la autocorrelación en el modelo, se debe utilizar el método de los errores de Newey también conocido como los errores CHA (errores consistentes con heterocedasticidad y autocorrelación). La ventaja de este método correctivo frente a los anteriores, es que este permite indicar el número de rezagos que dependen los residuos del modelo, siendo en este caso dos periodos máximos retardados. El comando requerido es **newey** y la opción para indicar el número de rezagos a utilizar es **lag**.

```
. newey imp pbi indp_v ibi,lag(2)

Regression with Newey-West standard errors      Number of obs     =         84
maximum lag: 2                                  F(  3,        80) =    2431.11
                                                Prob > F          =     0.0000
```

imp	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
pbi	.2382184	.0182355	13.06	0.000	.2019287	.2745081
indp_v	-5965.741	1266.283	-4.71	0.000	-8485.724	-3445.758
ibi	.1717964	.0718278	2.39	0.019	.0288546	.3147383
_cons	11429.6	3179.869	3.59	0.001	5101.459	17757.74

Figura 3.169. Resultados del método correctivo de los errores CHA modelo (3.7.69.).

$$IMP_t = 11429.6 + 0.24PBI_t - 5965.74INDP_t + 0.17IBI_t + \hat{\mu}_t \text{ (3.7.90.)}$$

$$ee = (3179.87) \quad (0.01) \quad (1266.28) \quad (0.07)$$

$$t = 3.59 \quad 13.06 \quad -4.71 \quad 2.39$$

Este método de corrección es una extensión del método correctivo de White aplicado para la heterocedasticidad, por tal motivo, los estimadores se mantienen, pero

sus respectivos errores estándares cambian lo suficiente para mostrar los resultados de los estimadores sin presencia de autocorrelación. Podemos notar que siguiendo el modelo corregido mediante los errores de CHA, los estimadores mantienen sus respectivas significancias individuales y el modelo conserva su significancia global. Este método no solo sirve para resolver problemas de autocorrelación con residuos que siguen esquemas autorregresivos superior al primer orden, sino también para resolver aquellos modelos que tienen heterocedasticidad y autocorrelación.

Para finalizar, cabe recalcar que los métodos C-O y P-W son los idóneos para corregir la autocorrelación cuando los residuos siguen esquemas $AR(1)$. No obstante, si este modelo especificado no tiene residuos que sigan un esquema $AR(1)$, el método de los errores CHA será utilizado para corregir la autocorrelación en el modelo.

3.7.2.5. Interpretación de los resultados.

Antes de interpretar los resultados se mostrarán dos tablas en el que se puede observar un resumen sobre la información de los modelos especificados para solucionar el problema de autocorrelación en el modelo.

Modelo especificado	Variables			
		Producto Bruto Interno	Índice de Protección	Inversión Bruta Interna
MCO (3.7.69.)	$\hat{\beta}_k$	0.24	-5965.74	0.17
	ee	(0.02)	(1093.50)	(0.17)
	t	15.59	-5.46	3.03
C-O dos pasos (3.7.86.)	$\hat{\beta}_k$	0.23	-5227.62	0.23
	ee	(0.01)	(1015.13)	(0.05)
	t	15.84	-5.15	4.19
P-W (3.7.87.)	$\hat{\beta}_k$	0.23	-5166.67	0.23
	ee	(0.01)	(998.50)	(0.05)
	t	15.92	-5.17	4.29
	$\hat{\beta}_k$	0.21	-4469.73	0.28

Método iterativo C-O (3.7.88)	ee	(0.01)	(923.74)	(0.05)
	t	16.32	-4.84	5.54
	$\hat{\beta}_k$	0.21	-4428.44	0.28
Método iterativo P-W (3.7.89)	ee	(0.01)	(908.62)	(0.05)
	t	16.44	-4.87	5.66
	$\hat{\beta}_k$	0.24	-5965.74	0.17
Errores CHA (3.7.90.)	ee	(0.02)	(1266.28)	(0.07)
	t	13.06	-4.71	2.39

Tabla 3.22. Información de los estimadores de los modelos especificados para corregir al modelo que explica a las importaciones.

Modelo especificado	MCO (3.7.69.)	C-O dos pasos (3.7.86.)	P-W (3.7.87.)	Método iterativo C-O (3.7.88.)	Método iterativo P-W (3.7.89.)	Errores CHA (3.7.90.)
Número de observaciones	84	83	84	83	84	84
Estadístico F calculado	1498.25	1990.05	2070.2	2661.91	2800.66	2431.11
Coeficiente de determinación	98.25%	98.69%	98.73%	99.02%	99.06%	
Error Estándar de la Regresión	1226.1	1209.4	1203.1	1198.9	1191.8	
Coeficiente de autocovarianza (p)	-0.14	-0.14	-0.14	-0.31	-0.31	

] **Tabla 3.23. Información de los modelos especificados para corregir al modelo que explica a las importaciones.**

Las anteriores tablas resumen información sobre los modelos con los que se ha intentado corregir la autocorrelación en el modelo original (3.7.69.) y se puede apreciar

que los modelos tienen significancia global y sus respectivos estimadores son significativos individualmente. A demás los modelos tienen una buena bondad de ajuste.

La información más importante son sus respectivos estadísticos dw y sus coeficientes de autocovarianza (p), mediante estos dos últimos datos se concluye que los métodos C-O y P-W no están resolviendo la autocorrelación en el modelo. Tal como se ya explicó, esto sucede porque los métodos C-O y P-W sólo son válidos si los residuos del modelo siguen un $AR(1)$. La condición mencionada no ha sido cumplida en el caso del modelo original (3.7.69.), ya que según la prueba BG, los residuos del modelo original (3.7.69.) están correlacionados entre sí con dos periodos rezagados. Entonces, el método que permite obtener mejores estimadores libres de autocorrelación es el método de estimación del modelo con errores CHA (3.7.90.).

En cuanto a los estimadores del modelo (3.7.90.), estos se interpretan de la misma forma que se interpretan los estimadores del modelo (3.7.69.).

4. Análisis de Regresión Lineal con Variable Dependiente Cualitativa

En ocasiones, los modelos microeconométricos se especifican para explicar variables cuantitativas como el nivel de ingresos de los trabajadores, número de asegurados en una zona, tasa de natalidad de un país, etc. Ciertamente, en la mayoría de los casos los modelos tienen variables dependientes cuantitativas que recogen información sobre datos numéricos de la población, no obstante, existen algunos modelos que utilizan a variables cualitativas para explicar ciertos rasgos de una sociedad o una persona. Las variables que recogen información sobre características, rasgos o condiciones de la unidad de estudio se les denomina **variables cualitativas** y su uso se ha extendido en las últimas décadas tanto en variables explicativas como en variables explicadas.

(Pucutay V., 2002) Explica que en la investigación sobre las sociedades y sus indicadores de vida se ha extendido el uso de modelos econométricos que reúnen un conjunto de variables explicativas, sean cualitativas o cuantitativas, para explicar a una realidad problemática o cierto fenómeno económico que son capturados en información cualitativa. Por ejemplo, es muy común utilizar modelos econométricos con variable dependiente binaria para explicar las causas pobreza en una determinada sociedad, de hecho, este el motivo por el cual su uso ha sido ampliado, ya que permite constatar cual es el efecto de un conjunto de variables explicativas que causan cierta condición o característica generalizada en la población.

En este capítulo se explicará el uso de los modelos que utilizan variables dependientes cualitativas y se detallará porque en ciertas investigaciones son mejores para explicar que un modelo de regresión clásico.

4.1. Conceptos Previos

4.1.1. Modelos de elección discreta.

(Uriel & Aldás, 2005) Definen a los modelos de elección discreta como aquellos modelos que usan a variables cualitativas como variables dependientes. También señalan que estos modelos están relacionados ampliamente con el análisis discriminante, y el uso de este tipo de modelos tiene ventajas frente a los modelos de regresión clásica, ya que permite obtener resultados eficientes y válidos usando menos supuestos. (Greene, 2012) Señala que el término "elección discreta" hace referencia a que estos modelos realizan un análisis de elección individual, por ejemplo, dadas algunas variables ¿Se debería comprar un seguro o no, en tiempos de elecciones?, ¿Cuál es el candidato de preferencia dadas algunas condiciones?, ¿Cuáles son los gustos y preferencias entre las marcas de bienes y/o servicios si consideramos sus ingresos, gastos o entre otros? Obviamente, estas preguntas pueden dar entre dos o más respuestas y no son variables socioeconómicas como tal sino más bien indicadores.

Otra ventaja que tienen los modelos de elección discreta es el cálculo de probabilidades sobre la ocurrencia o el cumplimiento de la variable estudiada. Esto quiere decir que los métodos econométricos no solo miden los efectos cuantitativos, sino también hallan las probabilidades que ejercen las variables explicativas sobre la variable dependiente, así lo especifica (Greene, 2012).

(Pérez L., 2012) Conceptualiza los modelos de elección discreta en la siguiente cita.

> *"Cuando la variable dependiente es una variable discreta que refleja decisiones individuales en las que el conjunto de elección está formado por alternativas separadas y mutuamente excluyentes estamos ante los modelos de elección discreta."* (Pérez L., 2012)

Por último, estos modelos son un tipo de modelo con variable dependiente limitada (VDL). Otros tipos de modelos con VDL son los modelos censurados, modelos truncados y de conteo.

4.1.2. Modelo de elección binaria.

Revisemos las preguntas que hicimos anteriormente, la primera pregunta tiene dos posibles respuestas "sí" y "no", ahora supongamos que hemos tomado un conjunto de variables explicativas para explicar cuáles son los factores que determinan que una persona compre o no un seguro, entonces estamos ante un **modelo de regresión de**

variable dependiente de elección binomial, también llamado **modelo de regresión binomial**, **modelo de elección binaria** o **modelo dicotómico**.

(Gujarati & Porter, 2010) Plantean otro ejemplo, suponiendo que para estudiar la participación de la fuerza laboral en una sociedad se dispone de la variable **PFL** la cual puede tomar dos posibles respuestas.

$$PFL = \begin{cases} PFL = 1 \rightarrow Si\ la\ persona\ está\ trabajando \\ PFL = 0 \rightarrow Si\ la\ persona\ no\ está\ trabajando \end{cases} \quad (4.1.1.)$$

Lo expresado en (4.1.1.) significa que la variable **PFL**, dependiendo de la realidad de cada persona, puede tomar dos posibles valores: "1" la persona se encuentra trabajando y "0" la persona no se encuentra trabajando. Entonces, **se denomina modelos de elección binaria a aquellos modelos que toman un conjunto de regresoras para explicar a una variable dependiente binomial**.

Por lo general, los valores que se les asigna a una variable dicotómica para indicar que cumplen una condición o característica y para señalar que no cumple la condición deseada, son los valores "1" y "0" respectivamente, no obstante, estos valores son totalmente arbitrarios y los investigadores pueden elegir los valores que crean conveniente. Para efecto de esta guía, al momento de construir las variables dicotómicas se utilizarán a los valores "1" para indicar que las unidades de estudio cumplen una característica o condición estudiada y "0" para señalar que las unidades de estudio no cumplen la condición estudiada. Las variables que solo admiten dos posibles valores se les conoce como **variables dicotómicas** o **variables Dummy**.

(Uriel & Aldás, 2005) Indican algunos ejemplos de temas de investigación que se pueden realizar con este tipo de modelos.

- **Elección de tenencia de vivienda**. Se suponen solo dos posibilidades: comprar (1) o pagar un alquiler (0). En este caso la característica estudiada es explicar los factores que determinan a que una persona pueda comprar una casa.
- **Referéndum de la constitución europea**. Votar sí (1) o no (0). En este ejemplo el tema estudiado es el referéndum en la constitución europea y la característica de interés es el "sí".
- **Consumidor de una determinada marca**. Si el usuario compra la marca señalada (1) caso contrario (0).

Analíticamente, una variable dependiente se representa como.

$$Y_i = \begin{cases} 1 \rightarrow Prob(Y_i = 1) = P_i \\ 0 \rightarrow Prob(Y_i = 0) = 1 - P_i \end{cases} \text{(4.1.2.)}$$

(4.1.2.) Significa que Y_i tiene una de probabilidad de P_i que sea igual a 1, por otro lado, tiene la probabilidad de $(1 - P_i)$ que Y_i sea igual a 0. Este es el tema principal de este capítulo y en las siguientes secciones se entrará en detalle sobre los métodos econométricos que se siguen para calcular los resultados. Por último, (Greene, 2012) Añade la siguiente función sobre los modelos de elección binaria.

> *"Con el propósito de estudiar el comportamiento individual, construiremos modelos que vinculen la decisión o el resultado con un conjunto de factores, al menos en un espíritu de regresión."* (Greene, 2012)

(Greene, 2012) También señala otra forma de expresar (4.1.2.)

$$Prob(Y = j) = F(efectos\ relativos, parámetros) \text{ (4.1.3.)}$$

Lo que (4.1.3.) da a entender es que la probabilidad que *Y* sea igual a *j*, donde *j* puede ser 1 o 0, está en función de los efectos relativos y de los parámetros del modelos especificado.

4.2. Modelos con Variables Dependientes Dicotómicas

¿En qué se diferencia el modelo de regresión lineal clásico con el modelo de elección binaria? La primera respuesta que podemos dar a esta pregunta es que el MRLC utiliza variables cuantitativas para designar a la variable dependiente, mientras que el modelo de elección binaria utiliza a variables Dummy como la variable dependiente. Otra respuesta que podemos dar tiene que ver con la distribución de los errores, en los MRLC deben seguir la distribución normal, mientras que los modelos de elección binaria pueden utilizar tres tipos de distribución en sus respectivos errores y son: distribución de Bernoulli, distribución logística acumulada y distribución normal **y dependiendo del tipo de distribución que siguen los errores de un modelo de elección binaria, tenemos tres tipos de modelos de elección binaria.**

- **Modelos de Probabilidad Lineal**. Los errores siguen la distribución de Bernoulli.
- **Modelos Logit**. Los errores siguen una distribución logística acumulada.

- **Modelos Probit**. Los errores siguen una distribución normal.

(Uriel & Aldás, 2005) Exponen las siguientes propiedades econométricas que siguen los modelos de elección binaria. Con la expresión (4.1.2.) se puede calcular la siguiente esperanza de Y_i.

$$E(Y_i) = 0 * (1 - P_i) + 1 * P_i = P_i \text{ (4.2.1.)}$$

(4.2.1.) Significa que la media o valor esperado de Y_i es igual a la probabilidad que tiene Y_i a ser igual a 1. (Uriel & Aldás, 2005) Prosiguen su explicación suponiendo que Y_i está explicado por un conjunto de regresoras. Recomiendan expresarlo tomando a la función Z_i, con el fin de evitar confundirlo con el MRLC.

$$Z_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} = [1 \quad X_{2i} \quad \dots \quad X_{3i}] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \text{ (4.2.2.)}$$

Entonces, la esperanza condicionada a las regresoras es.

$$E(Y_i | X_{2i}, X_{3i}, \dots, X_{ki}) = F(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}) = F(Z_i) \text{ (4.2.3.)}$$

Al especificar el modelo econométrico con la función Z_i vista en (4.2.3.) tenemos.

$$Y_i = E(Y_i | X_{2i}, X_{3i}, \dots, X_{ki}) + \mu_i = F(Z_i) + \mu_i \text{ (4.2.4.)}$$

Según el término de error μ_i en (4.2.4.) que sigue una distribución determinada, se elige el tipo de modelos de elección binaria.

A continuación, se presentarán detalles sobre cada tipo de modelo de elección binaria.

4.2.1. Modelos de Probabilidad Lineal.

El modelo de probabilidad lineal (MPL), en términos simples, se define como un MRLC cuyos errores μ_i y variable dependiente Y_i siguen una distribución de Bernoulli. Teniendo el siguiente modelo econométrico.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i \text{ (4.2.5.)}$$

Donde Y_i es una variable dicotómica, entonces podemos expresar su esperanza condicional como la probabilidad condicional que sea igual a 1 dado los efectos recogidos en X_i. (Gujarati & Porter, 2010) Denotan lo anterior como.

$E(Y_i|X_i) = \Pr(Y_i = 1|X_i)$ (4.2.6.)

Ya que el MPL sigue la estructura de un MRLC, entonces a (4.2.6.) podemos agregar lo siguiente.

$E(Y_i|X_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i = \Pr(Y_i = 1|X_i)$ (4.2.7.)

Además, debemos tener en cuenta el supuesto de exogeneidad representado como $E(\mu_i|X_i) = 0$ entonces (4.2.7.) se reescribe como.

$E(Y_i|X_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} = \Pr(Y_i = 1|X_i)$ (4.2.8.)

Al tener en cuenta (4.2.8.), la teoría econométrica indica que Y_i sigue una distribución de Bernoulli y a su vez la mayoría de variables aleatorias que siguen dicha distribución están determinadas por la probabilidad que la variable sea 1, mientras que las regresoras en su mayoría siguen la distribución binomial, según (Gujarati & Porter, 2010).

(Uriel & Aldás, 2005) Explican que este tipo de modelos tienen algunas ventajas, como: su facilidad al momento de calcular los resultados y no requieren asumir el cumplimiento de los supuestos que nos brindan estimadores MELI. No obstante, este modelo tiene más inconvenientes que ventajas.

- **El término de error no sigue una distribución normal.**

Esta es la principal desventaja y al mismo tiempo, el motivo por el cual no es recomendado utilizar este modelo para obtener los estimadores. Tal como se mencionó anteriormente, la variable dependiente y el término de error siguen una distribución de Bernoulli, la cual no es la deseada para obtener estimadores MELI. (Gujarati & Porter, 2010) Indican que la distribución normal no es inherente a la variable dependiente ni tampoco al término de error en los MPL, ya que estas variables solo pueden tomar dos valores "1" o "0". Además, muestran las distribuciones de probabilidades de μ_i, los cuales son.

$$\begin{matrix} Cuando\ Y_i = 1 \rightarrow \mu_i = 1 - \beta_1 - \beta_2 X_i \rightarrow Prob = P_i \\ Cuando\ Y_i = 0 \rightarrow \mu_i = \beta_1 - \beta_2 X_i \rightarrow Prob = (1 - P_i) \end{matrix} \quad (4.2.9.)$$

Y (Pucutay V., 2002) Muestra la gráfica de probabilidades.

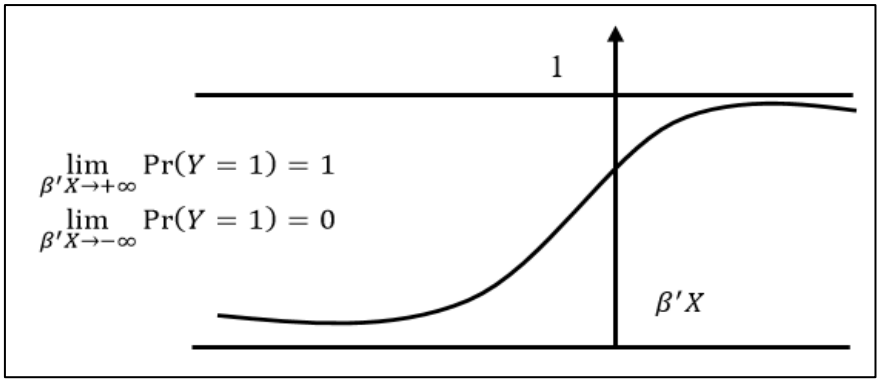


Figura 4.1. Modelo de Probabilidad Lineal.

Si las probabilidades siguen una distribución parecida a la figura anterior, entonces se puede interpretar que bastaría con una función de distribución acumulada para obtener estimadores MELI, este tema será abordado más adelante. (Gujarati & Porter, 2010) También mencionan que tomando en cuenta la propiedad de consistencia de los estimadores, el MPL en muestras grandes puede producir estimadores con distribución normal, por lo que en muestras grandes se podría tomar en cuenta este método de estimación que sigue los mismos pasos que el MCO.

- **Varianzas heterocedásticas**.

A raíz que los errores y la variable dependiente siguen una distribución de Bernoulli entonces la media y la varianza de las variables son p y $p(1-p)$ respectivamente, según (Gujarati & Porter, 2010) p es la probabilidad de tener éxito o $Y_i = 1$.

(Greene, 2012) Establece que el MPL produce varianzas heterocedásticas en medida que depende de los estimadores, por lo que hace que el término de error pierda la propiedad de constancia, ¿Qué se puede hacer al respecto? Una respuesta precipitada sería aplicar el método MCF o MCP, no obstante, (Greene, 2012) Menciona que esta práctica no es la recomendada en la siguiente cita.

> *"Un defecto grave es que sin ajustes ad hoc con las perturbaciones, no podemos estar seguros de que las predicciones de este modelo realmente se verán como probabilidades. No podemos limitar $X'\beta$ al intervalo* [0,1]. *Tal modelo produce probabilidades sin sentido y variaciones negativas. Por estas razones, el modelo de probabilidad lineal se usa con menos frecuencia, excepto como base para la comparación con otros modelos más apropiados."* (Greene, 2012)

La cita que se recoge de Green, da a entender que, la varianza heterocedástica inherente en la mayoría de MPL puede ocasionar que las probabilidades estimadas del

modelo no se encuentren en el intervalo $[0,1]$, por lo que no tendría sentido ni justificación realizar el MPL para estimar y calcular las probabilidades de un modelo. Entonces, debido a que el empleo de MCF no es una opción válida en la mayoría de MPL, la otra respuesta sería emplear el método de corrección de errores de White, no obstante (Gujarati & Porter, 2010) Advierte que este método de corrección debería tratarse con cuidado y solo podría ser aplicado en muestras grandes.

Para finalizar, (Colin C. & Trivedi, 2005) Comparan el método de estimación de MCO, que es el método de estimación para el MPL, con el método de Máxima Verosimilitud (MV) y definen que el método MPL con el método de estimación MCO producen estimadores inestables o ineficientes debido a que las observaciones con $X_i'\beta$ que están cercanas a 0 o 1 han sido asignados con mayor peso que el resto de observaciones y aunque el MPL con errores estándar heterocedásticos pueden ser una herramienta útil para el análisis de datos debido a su facilidad y sencillez, es mejor utilizar los modelos **logit** o **probit**, cuyos métodos de estimación se basan en el método de estimación MV. (Uriel & Aldás, 2005) Complementan lo anterior afirmando que los valores cercanos a 0 y 1 tienen varianzas más pequeñas.

- **No cumplimiento de $0 \leq \boldsymbol{E}(\boldsymbol{Y}|\boldsymbol{X}) \leq 1$.**

Si recordamos que P_i es la probabilidad que $Y_i = 1$ y se puede definir como la esperanza condicional de Y_i dado X_i, (4.2.8.) se puede replantear como.

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} = \Pr(Y_i = 1|X_i) = P_i \quad (4.2.9.)$$

Y ya que P_i se trata de una probabilidad sus valores solamente deben estar entre los valores del intervalo $[0,1]$. Entonces, la teoría econométrica indica que los MPL pueden ocasionar probabilidades que se encuentren fuera del intervalo, en consecuencia no habría sentido ni justificación realizar este método si los resultados están equivocados. (Gujarati & Porter, 2010) Denominan a este problema como el **verdadero problema con la estimación del MPL por MCO** y explica que se debe a que el método MCO no toma en cuenta la restricción sobre el valor de las probabilidades.

Entonces, en vista que el MPL puede conducirnos a resultados equivocados es conveniente plantearse la siguiente pregunta ¿Cuáles son las alternativas al MPL? La respuesta a esta pregunta son los **modelos de probabilidad no lineales** y estos modelos

comprenden a los modelos **logit** y **probit**. (Wooldrige, 2009) Muestra la siguiente función para los modelos logit y probit.

$$P(Y_i = 1|X_{2i}, X_{3i}, \dots, X_{ki}) = G(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}) \text{ (4.2.10.)}$$

Explica que la función G asume que los valores de los estimadores están estrictamente en el intervalo $[0,1]$, de esta forma se asegura que con el modelo de probabilidad no lineal se puedan obtener estimadores correctos, es por esto que es recomendado este tipo de modelos ampliamente en lugar del MPL.

La función G puede hacer que los errores sigan dos tipos de distribución, si el término de error sigue una distribución logística entonces estamos ante un modelo logit y por otro lado si el término de error sigue una distribución normal entonces se está usando el modelo probit, también llamado normit.

4.2.2. Modelos Logit.

Este modelo fue propuesto por Joseph Berkson y fue quien acuñó el término "logit" para referirse a este tipo de modelos que siguen una distribución logística. Los modelos logit siguen la siguiente función G.

$$P_i = G(Z) = \frac{\exp(Z)}{[1+\exp(Z)]} = \Lambda(Z) \text{ (4.2.11.)}$$

Y en su forma extendida (4.2.11.) se escribe como.

$$P_i = G(Z) = \frac{e^Z}{[1+e^Z]} = \frac{e^{(\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\cdots+\beta_k X_{ki})}}{[1+e^{(\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\cdots+\beta_k X_{ki})}]} \text{ (4.2.12.)}$$

(Greene, 2012) Nombra a la función (4.2.12.) como la **función de distribución logística acumulada** y tiene una campana de distribución simétrica, lo cual es lo deseado. (Gujarati & Porter, 2010) Mencionan dos rasgos importantes sobre el modelo logit: el primero es que a medida que Z se encuentra comprendido en los números reales, P_i se mantiene en el rango $[0,1]$, y la segunda es que la probabilidad P_i no depende de las regresoras, sin embargo, el modelo no es lineal ni en las regresoras ni en los estimadores, en consecuencia, el método MCO para estimar los estimadores resulta ser incorrecto.

(Gujarati & Porter, 2010) Continúan explicando que al tener a P_i como la probabilidad que $Y_i = 1$, entonces $1 - P_i$ que es la probabilidad que $Y_i = 0$ se escribe como.

$$1 - P_i = \frac{1}{[1+e^{Z}]} \text{ (4.2.13.)}$$

Con (4.2.11.) y (4.2.13.) se pueden calcular los llamados **coeficientes de razón (odds ratio), también llamado razón de apuestas** de la siguiente forma.

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z}}{1+e^{-z}} \text{ (4.2.14.)}$$

(Colin C. & Trivedi, 2005) Definen en términos simples los odds ratio como la medición de la probabilidad que $Y_i = 1$ en relación a la probabilidad que $Y_i = 0$ y brinda el siguiente ejemplo. Supongamos que en un estudio farmacéutico se quiere probar la efectividad de una droga farmacéutica, donde $Y_i = 1$ denota supervivencia del paciente y $Y_i = 0$ denota que no ha sobrevivido, y toma a la dosis de la droga estudiada como una regresora. Si el odds ratio fuese igual a 2 podemos interpretar el resultado como la probabilidad de supervivencia es dos veces mayor que la probabilidad de no sobrevivir.

(Escobar M., Fernández M., & Bernardi, 2012) Señalan que también se pueden calcular la razón $Y_i = 0$ frente a $Y_i = 1$. Aunque es poco usual y más se utiliza la razón anterior. A continuación, se expresa la razón de $Y_i = 0$ frente a $Y_i = 1$.

$$\frac{\Pr\,(Y_i=0)}{\Pr\,(Y_i=1)} = \frac{\Pr\,(Y_i=0)}{1-\Pr\,(Y_i=0)} \text{ (4.2.15.)}$$

Por ejemplo, si (4.2.15.) fuese $\frac{0.63}{0.37} = 1.7$ entonces se interpreta como: es 1.7 veces más probable que $Y_i = 0$ que $Y_i = 1$, según (Escobar M., Fernández M., & Bernardi, 2012).

Al convertir (4.2.14.) a logaritmos obtenemos lo siguiente.

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} \text{ (4.2.16.)}$$

(Uriel & Aldás, 2005) Señalan que la probabilidad P_i es una función no lineal de los estimadores, mientras el logaritmo de los odds ratio es una función lineal de los estimadores. (Gujarati & Porter, 2010) Indican que L_i se le denomina **logit** y de ahí proviene su nombre. Para (Colin C. & Trivedi, 2005) Los economistas deberían interpretar (4.2.14.) o (4.2.16.), ya que el estimador implica ser una semielasticidad. Además suponiendo que un estimador del modelo logit es 0.1 entonces a medida que el

regresor aumenta una unidad, la razón de probabilidades (odds ratio) aumenta en 0.1; la interpretación de los estimadores se explicara a más detalle en las siguientes secciones.

A manera de conclusión, el modelo logit se resume en la siguiente cita.

> *"De modo que el modelo de regresión logística es equivalente al modelo de regresión lineal con la diferencia de que transforma la variable dependiente en el logaritmo de su razón, para conseguir así que varíe de $-\infty$ a $+\infty$ y sobre ese valor estima la ecuación de la regresión."* (Escobar M., Fernández M., & Bernardi, 2012)

(Gujarati & Porter, 2010) Mencionan algunas características sobre los modelos logit.

- Si la probabilidad P_i va desde 0 a 1, el logit L estará comprendida entre $-\infty$ y $+\infty$.
- Aunque L sea lineal en las regresoras, las probabilidades no son lineales.
- Si L fuese positivo entonces el valor de las regresoras aumentan las probabilidades que $Y_i = 1$. Mientras si L fuese negativo entonces la probabilidad que $Y_i = 1$ disminuye si los valores de las regresoras incrementan.

4.2.3. Modelos Probit.

El surgimiento de los modelos probit o normit se le atribuye al bioestadístico americano Chester Ittner Bliss, quien en 1934 propuso este método de estimación para los problemas biológicos. En la actualidad, ha sido ampliamente utilizado en la ciencia económica.

Al igual que el modelo logit, el modelo probit parte desde (4.2.10.)

$$P(Y_i = 1|X_{2i}, X_{3i}, \dots, X_{ki}) = G(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}) \text{ (4.2.10.)}$$

Donde la función G sigue la distribución normal estándar según (Verbeek, 2004)

$$P_i = G(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \int_{-\infty}^{Z} \phi(t)dt = \Phi(Z) \text{ (4.2.17.)}$$

Donde $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$ es la función de densidad y $\Phi(Z)$ es la función de distribución normal estándar. Recordemos que la función de densidad en estadística se

refiere a la fórmula con la cual se calculan los valores de una variable aleatoria y la función de distribución normal estándar indica cómo se distribuyen esos valores.

(Gujarati & Porter, 2010) Explican que el modelo logit no es la única función de distribución acumulativa que se puede utilizar, también se puede hacer uso de una función de distribución acumulativa normal, de ahí que se le conoce como modelo probit o normit. Definen que, al tratarse de la probabilidad P_i que $Y_i = 1$ este se calcula por el área de la curva normal estándar de $-\infty$ a I_i, donde I_i se trata de una **variable latente** compuesta por un conjunto de regresores. Se les denomina **variable latente** a aquellas variables que no se pueden observar por sí mismas, sino que necesitan ser medidas utilizando otras variables, (Gujarati & Porter, 2010) Muestran un ejemplo donde se estudia la posibilidad de tener casa propia o no, la cual depende de un **índice de conveniencia** que está representado por el ingreso que perciben las familias debido a que el índice de conveniencia no es medible. Y establecen la siguiente igualdad.

$$I_i = Z + \mu = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu \text{ (4.2.18.)}$$

(Greene, 2012) Realiza el siguiente supuesto: μ es **normal** con media 0 y la varianza puede ser 1 o logística, y establece la siguiente relación.

$$Y_i = \begin{cases} 1 \; si \; I_i > 0 \\ 0 \; si \; I_i \leq 0 \end{cases} \text{ (4.2.19.)}$$

Además $\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}$ recibe el nombre de **función de índice**. Este tipo de funciones aparecen con frecuencia como un tipo de modelos con variable dependiente binomial.

> *"Un ejemplo que se cita muy a menudo es el de la decisión de hacer una compra importante: la teoría establece que el consumidor hace un cálculo beneficio marginal-coste marginal basándose en las utilidades que consigue si hace la compra o si no hace la compra y emplea el dinero en alguna otra cosa."* (Greene, 2012)

En este punto es importante asumir que el índice de conveniencia, el cual ha sido planteado como la variable latente, tiene ciertos **niveles críticos** o **umbrales del índice**. (Gujarati & Porter, 2010) Representan a los umbrales del índice como I^* y pueden tener los siguientes valores dependiendo de la variable latente.

$$Y_i = \begin{cases} 1 \; si \; I_i > I^* \\ 0 \; si \; I_i \leq I^* \end{cases} \text{ (4.2.20.)}$$

Usando el supuesto de normalidad tanto para la variable latente y para el umbral del índice se puede reescribir la expresión (4.2.17.) de una forma más extensa.

$$P_i = P(Y_i = 1) = P(I_i > I^*) = G(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \int_{-\infty}^{Z} \phi(t)dt = \Phi(Z)$$

(4.2.21.)

(Verbeek, 2004) Establece que en este tipo de modelos sobre la utilidad, también se puede utilizar la distribución logística y por ende el modelo logit, pero es más frecuente los modelos probit.

Entonces ¿Cuál es la distribución que se debe usar para los modelos de elección binaria? (Greene, 2012) Propone la siguiente respuesta recogida en la siguiente cita.

> *"Es natural preguntarse cuál de las dos distribuciones debe utilizarse. La distribución logística es similar a la distribución normal excepto por sus colas: son más altas en la distribución logística. Por tanto, las dos distribuciones tienden a dar probabilidades muy similares a los valores intermedios de* Z*. La distribución logística tiende a dar probabilidades mayores que la distribución normal al suceso* $Y = 0$ *cuando* Z *es muy pequeño y probabilidades menores que la distribución normal a* $Y = 0$ *cuando* Z *es muy grande."* (Greene, 2012)

(Wooldrige, 2009) Argumenta que si prevalecemos el supuesto de normalidad en el modelo especificado entonces, como economistas se tiende a favorecer el modelo probit, por lo que debería ser más famoso que el modelo logit. (Escobar M., Fernández M., & Bernardi, 2012) Comentan que si bien, es cierto que los resultados de ambos modelos cambian ligeramente, pero no existe una regla, principio o determinante que nos indique de forma tajante si elegir el modelo probit o modelo logit. Además plantean que sus resultados no son comparables entre sí.

(Colin C. & Trivedi, 2005) También comentan al respecto y proponen revisar tres aspectos: **las consideraciones teóricas**, **consideraciones empíricas** y **las regresoras endógenas**. Explican que, si tomamos en cuenta **las consideraciones teóricas** entonces la respuesta depende del **dgp (data-generating process)** el cual es desconocido, el problema radica en especificar la forma funcional de los estimadores. Si el dgp tiene $P = \Lambda(Z)$ entonces debemos usar el modelo logit, de forma similar si $P = \Phi(Z)$ entonces la

opción correcta es el modelo probit y en caso de usar la distribución incorrecta se podría obtener estimadores inconsistentes. No obstante, es posible que la incorrecta especificación de la función de distribución no conlleve a consecuencias demasiado graves. Si los regresores se distribuyen de modo que sus respectivas medias sean lineales en Z, entonces se demuestra que elegir la función incorrecta G afecta a todos los parámetros de la pendiente por igual, de modo que la relación de pendiente-parámetro es constante en los modelos logit o probit.

Las consideraciones teóricas indican que el modelo logit tiene una forma relativamente simple para la condición de primer orden y una distribución asintótica, de hecho, cuando Berkson propuso y posteriormente popularizó su uso, se valió de este argumento para que el modelo logit sea preferido ante el modelo probit, (Colin C. & Trivedi, 2005) También indican que la interpretación de la relación **log-odds**, los cuales corresponden a la forma funcional (4.2.16.), y el análisis discriminante son la principales atracciones del modelo logit. Por otro lado, tal como dijo (Wooldrige, 2009), los economistas prefieren al modelo probit porque toma en cuenta a las variables latentes aleatorias con distribución normal.

(Colin C. & Trivedi, 2005) Exponen algunas consideraciones empíricas, la más importante es que empíricamente tanto el modelo logit como el modelo probit se pueden utilizar para cualquier modelo, las probabilidades predichas tanto en el modelo logit como en el modelo probit son ligeramente diferentes y concordando con (Greene, 2012) La diferencia mayor, empíricamente hablando, está en sus colas.

Empíricamente, también se puede utilizar el **log-likelihood** para comparar los modelos logit y probit, este estadístico se traduce literalmente como **probabilidad de registro**, pero según la teoría econométrica se entiende como la **función de verosimilitud** y se calcula con el método de estimación **Maximum Likelihood Estimation,** traducido del inglés significa **Estimación por Máxima Verosimilitud** por lo tanto, se puede intuir que para estimar los modelos logit y probit se usa el método de estimación de máxima verosimilitud. Es posible que la función de verosimilitud sea ligeramente similar entre los modelos logit y probit en algunos modelos especificados.

Por último, ambos modelos se extienden para manejar las complicaciones que surgen en el análisis microeconométrico. Las regresoras endógenas se podrían tomar en cuenta para elegir qué modelo utilizar. Estas se acomodan usando métodos de estimación

similares a datos censurados y métodos de datos de panel. La presencia de tales complicaciones, conlleva a preferir el uso el modelo lineal de probabilidad, debido a que estos pueden aplicarse siempre que sus errores estándares se ajusten a la heterocedasticidad. Cabe recordar, que las regresoras endógenas son aquellas que presentan correlación con el término de error, es decir no cumplen el supuesto de exogeneidad.

Como conclusión, podríamos afirmar que no existe una regla de decisión determinante que nos indique cuál método utilizar. Podríamos apoyarnos en la distribución que siguen las probabilidades o en la función de verosimilitud, pero estas diferencias nos conducen a resultados, que en su mayoría de las veces, no son tan significativas. Teóricamente, el modelo logit es más sencillo de estimar que el modelo probit, sin embargo, empíricamente, sus resultados no son tan diferentes.

4.3. Estimación de los Modelos de Elección Binaria no Lineales.

4.3.1. Estimación de los estimadores según el método MV.

Debido a que los modelos logit y probit no son modelos lineales, sus estimadores no pueden ser estimados mediante MCO. Por lo tanto, se tiene que utilizar el método de máxima verosimilitud (MV). Una diferencia entre el método de MCO y MV es explicada por (Uriel & Aldás, 2005), quienes sostienen que, al ser modelos no lineales, el método de estimación MV hace uso de procedimientos iterativos, algo parecido a los métodos de corrección P-W o C-O iterativo.

(Bravo & Vásquez Javiera, 2008) Explican en qué consisten los estimadores de MV.

> *"El estimador Máximo Verosímil es otro método para estimar la relación que existe entre la o las variables explicativas y la variable dependiente, la idea de este estimador es que la variable dependiente al ser una variable aleatoria tiene asociada una función de probabilidad la que depende de ciertos parámetros, por ejemplo, en el caso de una distribución normal estos parámetros son la media y la varianza. Entonces asumiendo una cierta distribución de la variable se tiene que determinar los parámetros de esa distribución que hacen más probable la muestra que observamos."* (Bravo & Vásquez Javiera, 2008)

En palabras más simples, la estimación mediante MV se consigue asumiendo que la distribución de la variable dependiente es conocida y posteriormente, con dicha función se toman los estimadores que permiten aumentar la probabilidad de observar a la variable dependiente. La explicación anterior se puede describir como una explicación general, ya que el método MV no solo sirve para estimar modelos logit o probit sino también otros tipos de modelos.

Omitamos por un momento la estimación de modelos logit o probit, para explicar brevemente el método MV de forma general. (Gujarati & Porter, 2010) Suponen que tenemos un modelo econométrico cualquiera como $Y_i = \beta_1 + \beta_2 X_i + \mu_i$ y asumimos que tanto Y_i como μ_i siguen una distribución normal. Por lo que podemos denotar la función de densidad para cada valor de la variable dependiente como:

$$f(Y_i|\beta_1 + \beta_2 X_i, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}\frac{(Y_i-\beta_1-\beta_2 X_i)^2}{\sigma^2}] \quad (4.3.1.)$$

Debido al supuesto de independencia entre los valores de la variable dependiente, la función de densidad conjunta se expresa utilizando (4.3.1) de la siguiente forma:

$$f(Y_1, Y_2, \dots, Y_n|\beta_1 + \beta_2 X_i, \sigma^2) = f(Y_1|\beta_1 + \beta_2 X_i, \sigma^2) f(Y_2|\beta_1 + \beta_2 X_i, \sigma^2) \dots f(Y_n|\beta_1 + \beta_2 X_i, \sigma^2) = \prod_{i=1}^{n} f(Y_i|\beta_1 + \beta_2 X_i, \sigma^2) \quad (4.3.2)$$

$$\prod_{i=1}^{n} f(Y_i|\beta_1 + \beta_2 X_i, \sigma^2) = (\frac{1}{\sigma\sqrt{2\pi}})^n \exp\,[-\frac{1}{2}\frac{\sum_{i=1}^{n}(Y_1-\beta_1-\beta_2 X_i)^2}{\sigma^2}] \quad (4.3.3)$$

A la función (4.3.3) se le expresa concretamente como **función de verosimilitud** y tal como indica su nombre, nos permitirá estimar los estimadores de forma que la probabilidad de observar la variable dependiente sea la más alta. Entonces, el método MV calcula los estimadores maximizando la función de verosimilitud (4.3.3). Similar al método de estimación MCO, se espera que el método MV cumpla los supuestos como no multicolinealidad, homocedasticidad, etc.

El método MV aplicado a los modelos no lineales de elección binaria hace uso del supuesto de independencia, el cual se define como el supuesto que considera a cada observación como una realización individual de una variable aleatoria, y también usa a la función de densidad conjunta (función de verosimilitud), tal como ya se especificó anteriormente. Lo característico en esta aplicación del método MV es que debemos restringir la forma funcional, según (Verbeek, 2004). Es decir, debemos adecuar la forma

funcional de la función de verosimilitud en (4.3.3) acorde al modelo que deseamos estimar. (Greene, 2012) Expresa a la función de verosimilitud de un modelo con probabilidad de elección binaria.

$$\Pr(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | X_i) = \prod_{y_i=1} G(Z) \prod_{y_i=0} [1 - G(Z)]$$

(4.3.4)

(Bravo & Vásquez Javiera, 2008) Explican que en los modelos de elección binaria sólo tienen dos posibles valores, los cuales pueden ser 1 y 0 y cuya probabilidad que sea 1 o 0, depende de la función de distribución acumulada $G(Z)$.

$\Pr(Y_i = 1) = G(Z)$ (4.3.5)

$\Pr(Y_i = 0) = 1 - G(Z)$ (4.3.6)

La función de verosimilitud (4.3.4) se puede reescribir.

$L = \prod_{i=1}^{n} [G(Z)]^{y_1} [1 - G(Z)]^{1-y_1}$ (4.3.7)

Ahora se procede a tomar el logaritmo de (4.3.7)

$\ln L = \sum_{i=1}^{n} \{y_i \ln G(Z) + (1 - y_i) \ln[1 - G(Z)]\}$ (4.3.8)

La función (4.3.8) se le conoce como la función de log-verosimilitud y ya sea estimando el modelo probit o logit, si es una distribución simétrica se cumple que $1 - G(Z) = G(-Z)$. Para estimar correctamente los estimadores debemos maximizar la función log-verosimilitud con respecto a los estimadores. Para explicarlo matemáticamente recordemos que $Z = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki}$ y matricialmente se representa como $Z = X'\beta$, según (4.2.2.). Ahora podemos reemplazarlo en (4.3.8).

$\ln L = \sum_{i=1}^{n} \{y_i \ln G(X'\beta) + (1 - y_i) \ln[1 - G(X'\beta)]\}$ (4.3.9)

Con (4.3.9) veamos la condición de primera orden calculada derivando con respecto a la matriz de los estimadores.

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i g(X'\beta)}{G(X'\beta)} - \frac{(1-y_i) g(X'\beta)}{(1-G(X'\beta))}\right] X = 0$$ (4.3.10)

A (4.3.10) se le conoce como las **ecuaciones de verosimilitud** donde $g(X'\beta)$ es la función de densidad, que dependiendo de la distribución, será normal o logística. Según

(Greene, 2012) Al seleccionar una forma concreta para $G(X'\beta)$ se obtiene un modelo empírico y ya que los modelos logit y probit son modelos no lineales, las ecuaciones en (4.3.10) serán resueltas con un método iterativo que en la mayoría de programas estadísticos utilizan algoritmos, el procedimiento es parecido a los métodos iterativos de P-W y C-O.

Para estimar los estimadores del modelo logit primero debemos recordar su función de distribución logística acumulativa $P_i = G(Z) = \frac{\exp(Z)}{[1+\exp(Z)]} = \Lambda(Z)$ y utilizándolo en (4.3.10) podemos obtener sus respectivas ecuaciones de verosimilitud derivando con respecto a los estimadores, como se ve en (4.3.11).

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}(y_i - \Lambda(Z))X = 0 \text{ (4.3.11)}$$

(Greene, 2012) Impone una condición a X de (4.3.11). Si X tiene el término constante, podemos deducir que las condiciones de primer orden (4.3.11) implican que la media de las probabilidades estimadas coincida con la proporción de valores igual a 1 de la variable dependiente en la muestra; esta imposición también se cumple para el MPL, sin embargo, no se ha podido observar si en la práctica también se cumple en el modelo probit. (Colin C. & Trivedi, 2005) Agregan que si X tiene el término constante hace posible que se pueda derivar (4.3.11) y que sus residuos sumen 0, por lo tanto, es indispensable que el modelo especificado logit contenga el término constante. (Verbeek, 2004) También comenta sobre (4.3.11) exponiendo que si el modelo especificado tiene una variable ficticia como regresora, por ejemplo, la variable ***genero*** tiene el valor 1 para mujeres y 0 para varones, entonces la frecuencia estimada será igual a la frecuencia real para cada grupo de género.

Posteriormente, (Greene, 2012) Enseña la segunda derivada del modelo logit.

$$H = \frac{\partial \ln L}{\partial \beta \partial \beta'} = -\sum_i \Lambda_i(1 - \Lambda_i)XX' \text{ (4.3.12)}$$

(Greene, 2012) Explica que debido a que la variable dependiente no aparece en (4.3.12), los métodos de Newton y de tanteo pueden ser utilizados para obtener los resultados, incluso sin importar cuál método se utilice nos brindaran el mismo resultado. El hessiano en (4.3.12) siempre será una matriz definida negativa, por lo que la función de verosimilitud logarítmica es cóncava, al usar el método de Newton se converge al máximo de la función de verosimilitud logarítmica, por lo general, en pocas iteraciones.

Por otro lado, para estimar el modelo probit debemos recordar la función de distribución de probabilidad normal en (4.2.17.) $P_i = G(Z) = \int_{-\infty}^{Z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \int_{-\infty}^{Z} \phi(t)dt = \Phi(Z)$ donde $\Phi(Z)$ es la función normal estándar y $\phi(t)$ es la función de densidad. Matricialmente podemos escribir $Z = X'\beta$ según (4.2.2.), entonces con (4.2.17.) podemos expresar la función de verosimilitud logarítmica como.

$$\ln L = \sum_{y_i=0} \ln[1 - \Phi(X'\beta)] + \sum_{y_i=1} \ln[\Phi(X'\beta)] \text{ (4.3.13.)}$$

Ahora tenemos que maximizar (4.3.13.) para obtener las condiciones de primer orden derivando con respecto a β.

$$\frac{\partial \ln L}{\partial \beta} = \sum_{y_i=0} \frac{-\phi_i}{1-\Phi_i} X_i + \sum_{y_i=1} \frac{\phi_i}{\Phi_i} X_i = \sum_{y_i=0} \lambda_{0i} X_i + \sum_{y_i=1} \lambda_{1i} X_i$$

(4.3.14.)

(Greene, 2012) Sugiere usar $L = \sum_i \ln F(q_i X'\beta)$ si $q = 2y - 1$ para reducir (4.3.14.)

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} [\frac{q_i \phi(q_i X'\beta)}{\Phi(q_i X'\beta)}] X = \sum_{i=1}^{n} \lambda_i X = 0 \text{ (4.3.15.)}$$

A diferencia del sencillo cálculo de las segundas derivadas en el modelo logit, cuando se trata de calcular las segundas derivadas del modelo probit, estas resultan ser más complicadas de hallar. Con el uso de la simplificación de la variable $\lambda(y_i, X'\beta) = \lambda_i$ la segunda derivada puede ser obtenida usando el resultado para cualquier Z, $\frac{\partial \phi(Z)}{\partial Z} = -Z\phi(Z)$, el hessiano para el modelo probit se define como.

$$H = \frac{\partial \ln L}{\partial \beta \partial \beta'} = \sum_i -\lambda_i(\lambda_i - \mathrm{X}'\beta) XX' \text{ (4.3.16.)}$$

Por último, el método de estimación MV para modelos logit y probit también calculan los errores estándares de los estimadores en la matriz de covarianza asintótica. (Pérez L., 2012) Muestra la forma de cómo calcular la matriz de covarianza asintótica de los estimadores.

$$\widehat{Avar}(\hat{\beta}) = [I(\hat{\beta})]^{-1} = \{\sum_{i=1}^{n} \frac{[g(X\hat{\beta})]^2 X'X}{G(X\hat{\beta})[1-G(X\hat{\beta})]}\}^{-1} \text{ (4.3.17.)}$$

En (4.3.17.) debemos recordar que $G(X\hat{\beta})$ determina si estamos ante un modelo logit o probit. (Wooldrige, 2009) Compara la expresión (4.3.17.) con la forma de hallar

la matriz de covarianza de los estimadores mediante el método MCO, no obstante en (4.3.17.) no se toma en cuenta la varianza del término de error σ^2, e indica que (4.3.17.) representa la naturaleza no lineal de los modelos. De forma similar que la matriz de covarianza de los estimadores del modelo MCO, las raíces de la diagonal de (4.3.17.) son los errores estándares de los estimadores.

4.3.2. Los efectos marginales.

Si recordamos los estimadores obtenidos en el método de MCO, estos son capaces de medir el cambio en la variable dependiente frente a un cambio unitario en una regresora manteniendo el supuesto de linealidad. Sin embargo, en los modelos que no asumen el cumplimiento de la linealidad, como el logit y probit, sus estimadores no miden directamente el efecto de las regresoras sobre la probabilidad que la variable dependiente sea igual a 1. (Wooldrige, 2009) Establece que los estimadores pueden ser usados para observar el signo que tendrá el efecto marginal de la regresora sobre la probabilidad de éxito, pero no podemos fiarnos del todo en los estimadores. Para que quede más claro, (Gujarati & Porter, 2010) Definen las siguientes apreciaciones sobre lo que realmente quieren decir los estimadores en los modelos de regresión lineal, MPL, modelo logit y probit.

- En un MRLC, el estimador mide el cambio promedio de la variable dependiente ante un cambio unitario en el valor de la regresora, asumiendo ceteris paribus en las demás regresoras del modelo.
- En el MPL, al igual que en el MRLC, los estimadores miden directamente el cambio de probabilidad de éxito frente a un cambio unitario en el valor de las regresoras, asumiendo ceteris paribus en el resto de las regresoras.
- En un Modelo logit, los estimadores indican el cambio en el logaritmo de la probabilidad de ocurrencia ante un cambio unitario en la regresora, con el supuesto de ceteris paribus en las demás regresoras. No obstante, la tasa de cambio de probabilidad de éxito está determinada por $\beta_j P_i(1 - P_i)$, donde β_j es el coeficiente de una determinada regresora.
- En un Modelo probit, de forma parecida al modelo logit, la tasa de cambio de probabilidad está dada por $\beta_j f(Z_i)$ donde $f(Z_i)$ es la función de densidad de la variable normal estándar.

De los puntos anteriores, se concluye que en los modelos logit y probit, los estimadores se ven influenciados por las regresoras en el momento de calcular los cambios de probabilidad, a diferencia del MPL que sus estimadores no están influenciadas de las regresoras. En consecuencia, se debe calcular los efectos marginales de las regresoras en los modelos logit y probit.

(Colin C. & Trivedi, 2005) Presentan la fórmula para calcular el efecto marginal, sin importar si se trata de un modelo logit o probit.

$$\frac{\partial \Pr\,(Y=1|X)}{\partial X} = \frac{\partial G(Z)}{\partial Z} * \beta_j = g(X\beta) * \beta_j \text{ (4.3.18.)}$$

Donde $g(X\beta)$ es la función de densidad que puede ser logística (logit) o normal estándar (probit). Los efectos marginales en (4.3.18.) varían de individuo a individuo. Debido a que $g(Z) > 0$ para todo Z, entonces es válido asumir que el signo del efecto marginal de la regresora es el mismo al signo del estimador en el modelo estimado, según (Pérez L., 2012).

(Greene, 2012) Reescribe (4.3.18.) según la función de distribución usada. Si se utiliza la función normal estándar (probit), se obtiene:

$$\frac{\partial \Pr\,[Y=1|X]}{\partial X} = \phi(\beta' X)\beta \text{ (4.3.19.)}$$

Y si se utiliza a la función de distribución logística (logit), se obtiene:

$$\frac{\partial \Pr\,[Y=1|X]}{\partial X} = \frac{\partial \Lambda(\mathrm{X}'\beta)}{\partial (X'\beta)} = \frac{\exp\,(X'\beta)}{[1+\exp(X'\beta)]^2} = \Lambda(\mathrm{X}'\beta)[1 - \Lambda(\mathrm{X}'\beta)] \text{ (4.3.20.)}$$

4.4. Inferencia en los Modelos de Elección Binarios no Lineales.

Del mismo modo que en los modelos de regresión lineal clásico, en los modelos no lineales logit y probit, también se asume que cumpla los supuestos para obtener estimadores MELI, no obstante, es evidente que el supuesto de linealidad no se adopta en este tipo de modelos. Aunque, según (Uriel & Aldás, 2005) Conviene tomar en cuenta que el modelo logit tiene una relación lineal entre el logaritmo de los **odds ratio** y las variables regresoras.

Otra similitud entre los modelos de regresión lineal clásico y los modelos de probabilidad no lineal son las pruebas de hipótesis sobre su significancia global e individual y la bondad de ajuste de los modelos. Previamente a explicar los conceptos de

inferencia estadística en los modelos de probabilidad de elección, es necesario recalcar que, tanto los modelos logit como modelos probit utilizan los mismos estadísticos para determinar si el modelo estimado presenta significancia.

4.4.1. Prueba de hipótesis sobre la significancia global.

Si recordamos al modelo de regresión lineal clásico estimado mediante el método MCO, se ha determinado que la prueba de significancia global utiliza al estadístico F de Snedecor. En el caso de los modelos logit y probit, debido a su condición de ser modelos no lineales, no podemos utilizar al estadístico F para probar su significancia global.

La teoría econométrica prevé esta situación y sugiere la utilización del **Likelihood Ratio (LR) test** para determinar si el modelo estimado tiene significancia global, traducido del inglés significa **contraste de razón de verosimilitud (RV)**. (Wooldrige, 2009) Define al test LR como en la siguiente cita:

> *"La prueba RV está basada en el mismo concepto que la prueba F en un modelo lineal. La prueba F mide el incremento en la suma de los residuales cuadrados cuando las variables se desechan del modelo. La prueba RV está basada en la diferencia en las funciones de log-verosimilitud para los modelos restringidos y no restringidos."* (Wooldrige, 2009)

Para entender la cita anterior, veamos la fórmula con la cual se calcula LR.

$$LR = 2(\ln L_{nr} - \ln L_r) \quad (4.4.1.)$$

En (4.4.1.) tenemos que $\ln L_{nr}$ es el valor de la función de log-verosimilitud no restringido y $\ln L_r$ es la función de log-verosimilitud restringido.

Según (Greene, 2012), La teoría econométrica suscita a usar tres tipos de medida para concluir que el modelo está correctamente estimado y **son el test LR**, **contraste de Wald** y **el contraste del multiplicador de Langrange.** Ahora supongamos que hemos estimado un modelo logit o probit, θ representa un estimador de esos modelos y $H_0: c(\theta) = 0$ es la prueba de hipótesis que contrasta si la restricción sobre los estimadores $c(\theta)$ es válida o no. El test LR contrasta si $c(\theta)$ es válida, en el caso que sea válida entonces la diferencia en (4.4.1.) no debería ser grande. (Wooldrige, 2009) Complementa lo anterior comparando al LR con el coeficiente de determinación, si se omite una variable regresora importante en el modelo especificado, al momento de estimarlo, el LR será

menor al LR del modelo que no ha omitido ninguna variable regresora importante; entonces el modelo restringido sería el modelo que está omitiendo variables regresoras importantes, mientras que el modelo no restringido sería el modelo que no está omitiendo variables regresoras. Por consiguiente, si la restricción es inválida podemos asumir que el modelo no restringido tiene significancia global porque estamos tomando variables regresoras importantes.

Es necesario recalcar que el ejemplo de restricción sobre las variables regresoras es algo arbitraria, es decir, por lo general los programas estadísticos siguen la siguiente fórmula para el test de LR.

$$LR = 2(\ln L - \ln L_0) \text{ (4.4.2.)}$$

En (4.4.2.) tenemos que $\ln L$ es la función de log-verosimilitud del modelo original estimado y $\ln L_0$ es la función de log-verosimilitud en el modelo estimado solamente del término independiente. En este caso, $\ln L$ se ha calculado del modelo no restringido y $\ln L_0$ del modelo restringido y la restricción es que todos los estimadores del modelo no restringido, que es el modelo original, son iguales a 0; por lo que se está probando si el modelo no restringido tiene significancia global. La prueba de hipótesis $H_0: c(\theta) = 0$ pasaría a ser $H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0$ y de forma parecida al anterior ejemplo de restricción, si la diferencia en (4.4.2.) es grande se concluye que su restricción es inválida, entonces podemos concluir que los estimadores no son iguales a 0, por lo que efectivamente tenemos significancia global.

En (4.4.2.) la prueba de hipótesis sería.

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_k = 0 \text{ (4.4.3.)}$$

$$H_1: ningún\ \beta_k = 0$$

Podríamos decir que tanto (4.4.1.) cómo (4.4.2.) son los estadísticos calculados, y tienen la siguiente distribución.

$$LR \sim X_q^2 \text{ (4.4.4.)}$$

Donde q son los grados de libertad determinados por el número de regresoras en el modelo. Y la regla de decisión es la misma en la prueba F, si $LR > X_q^2$ entonces rechazamos la hipótesis nula y aceptamos la hipótesis alternativa, que en el caso (4.4.3.)

se concluye que el modelo tiene significancia global y si $LR < X_q^2$ entonces aceptamos la hipótesis nula y en (4.4.3.) se concluye que el modelo no tiene significancia global.

En conclusión, el test de LR utiliza (4.4.2.) para determinar si el modelo empleado tiene significancia global y (4.4.1.) para comprobar si no se ha omitido ninguna variable regresora importante. En ambos casos, si la restricción es inválida entonces la diferencia es más grande lo que permite rechazar la hipótesis nula.

4.4.2. Pseudo R^2.

Al igual que en el modelo de regresión lineal clásica, en los modelos logit y probit también se hace uso de una medida que determine cuánto es la bondad de ajuste de las regresoras con respecto a la variable dependiente. Sin embargo, en estos modelos no lineales, tanto la interpretación como el nombre con que se denomina a la medida de bondad de ajuste son distintos a los modelos de regresión clásicos. En los modelos logit y probit se les denomina como **pseudo R^2** y efectivamente es análogo a R^2.

Debido a su naturaleza de modelos de probabilidad, en los modelos logit o probit el pseudo R^2 está basado en la comparación de modelos, más específicamente las funciones de log-verosimilitud del modelo original y del modelo estimado solamente con el término de constante. Veamos la fórmula.

$$Pseudo\ R^2 = 1 - \frac{\ln L_F}{\ln L_0} \quad (4.4.5.)$$

(Acosta G., Andrada F. Julián, & Fernández M., 2009) Explican que en (4.4.5.) tenemos $\ln L_F$ que representa la función log-verosimilitud del modelo estimado y $\ln L_0$es la función log-verosimilitud del modelo estimado solamente con el término constante. Esta medida fue propuesta por McFadden en 1974, por ello a (4.4.5.) se le conoce como $McFadden\ R^2$ y la mayoría de programas estadísticos calculan de forma predeterminada el $McFadden\ R^2$, no obstante existen otras medidas de $Pseudo\ R^2$ pero son menos utilizadas.

(Greene, 2012) Determina que los valores posibles de (4.4.5.) están comprendidos entre 0 y 1. Similarmente al R^2, si $McFadden\ R^2$ se acerca a 1 entonces el modelo tiene una buena bondad de ajuste, mientras, si $McFadden\ R^2$ se acerca a 0 entonces el modelo no tiene una buena bondad de ajuste y se debería plantear cambiar la especificación del modelo. Sin lugar a dudas, el $McFadden\ R^2$ es útil para determinar la bondad de ajuste

en los modelos logit y probit, pero no es tan preciso para determinar cuánto es exactamente la bondad de ajuste, de hecho, es posible que no pase del 0.5 por lo que, si el $McFadden\ R^2$ se encuentra entre 0.2 y 0.4 podemos considerar que el modelo una buena bondad de ajuste.

4.4.3. El estadístico Z y la Test de Wald.

Otra similitud entre el MRLC y los modelos logit y probit, es que en ambos se tiene que determinar la significancia individual de sus estimadores. Se ha determinado que en los MRLC se haga uso del estadístico t calculado con una distribución según la tabla t de Student para contrastar la prueba de hipótesis de significancia individual de los estimadores, matemáticamente se expresa como $tc \sim t_{\frac{\alpha}{2},gl}$, donde $t_{\frac{\alpha}{2},gl}$ es el estadístico tabulado y tc es el estadístico calculado hallado mediante $tc = \frac{\widehat{\beta}_k}{\widehat{\sigma}_{\widehat{\beta}_k}}$.

En los modelos logit y probit, el estadístico usado para determinar la significancia individual de sus estimadores también es el estadístico t pero con una distribución normal estándar, es decir se utiliza la misma fórmula para calcular el estadístico calculado, pero en los modelos logit y probit no se encuentran distribuidas en la tabla t de Student, sino según la tabla normal estándar.

Para distinguirlo se emplea el estadístico Z calculado, hallado mediante.

$$Zc = \frac{\widehat{\beta}_k}{\widehat{\sigma}_{\widehat{\beta}_k}} \text{ (4.4.6.)}$$

La distribución normal estándar es usada para determinar el estadístico Z tabulado. Se plantean las mismas pruebas de hipótesis.

$H_0: \beta_k = 0$ (4.4.7.)

$$H_1: \beta_k \neq 0$$

Se sigue la misma regla de decisión, si el estadístico Z calculado es mayor al estadístico Z tabulado, entonces se rechaza la hipótesis nula y se concluye que el estimador es significativo y por lo tanto la variable que le acompaña debe estar incluido en el modelo. Por el contrario, si el estadístico Z calculado es menor al estadístico Z tabulado, se acepta la hipótesis nula y se concluye que el estimador no es significativo y por lo tanto se debe plantear si la variable que le acompaña debe estar incluido en el

modelo. Según (Gujarati & Porter, 2010) La utilización de la distribución normal estándar en vez de la t de Student se debe a que los errores estándares de los estimadores son asintóticos. No obstante, esta no es la única forma de contrastar si el estimador es significativo o no, en la mayoría de trabajos de investigación se opta por utilizar el **test de Wald**, el cual es muy efectivo en muestras pequeñas según (Baum, 2006).

Al igual que el test de LR, el test de Wald está basada en restricciones. Según (Greene, 2012) Si tenemos el conjunto de restricciones $R\beta = q$ donde β es la matriz vector de los estimadores y V es la matriz de varianza-covarianza, entonces el estadístico calculado de Wald es.

$$Wc = (R\hat{\beta} - q)'(RVR')^{-1}(R\hat{\beta} - q) \quad (4.4.8.)$$

Este cálculo no es necesario de aplicar de forma manual. En la mayoría de programas econométricos como en STATA, basta con introducir una instrucción en forma de un comando para ordenar al programa que calcule el estadístico de Wald. Para finalizar, el estadístico de Wald se encuentra distribuido según la tabla X_q^2, donde q son los grados de libertad determinado por el número de restricciones y mantienen la regla de decisión similar al test de LR. Posteriormente, se explicará cómo utilizarlo en STATA.

4.5. Ejemplo con STATA sobre la Estimación de un Modelo Logit con Datos de ENAHO

A continuación, se procede a mostrar el esquema de una investigación utilizando un modelo econométrico de elección binaria con el fin de complementar la explicación anterior y demostrar que, aunque los métodos de estimación sean distintos se deben seguir los mismos pasos para concluir de forma exitosa un trabajo de investigación con modelos econométricos.

El ejemplo que se presentará consiste en una réplica en STATA sobre el trabajo de investigación de (Aparicio, Jaramillo, & San Román , 2011) cuyo título es "Desarrollo de la infraestructura y reducción de la pobreza: el caso peruano"; y se utilizarán datos de la Encuesta Nacional de Hogares recolectados por el Instituto Nacional de Estadística e Informática en el año 2018 a nivel nacional. En la siguiente cita se expone brevemente el tema principal del trabajo de investigación.

> *"Este documento analiza el rol de la infraestructura en la reducción de la pobreza en los hogares del Perú, bajo una perspectiva dinámica y bajo un enfoque de*

activos. Para ello, se estiman modelos Logit para recoger el impacto de los distintos tipos de infraestructura sobre la probabilidad de ser pobre en el Perú." (Aparicio, Jaramillo, & San Román , 2011)

Como se puede apreciar en la cita, el objetivo del trabajo de investigación es especificar un modelo que estime los efectos que tienen los tipos de infraestructura sobre la probabilidad de ser pobre en el Perú, para cumplir el objetivo se ha optado por la construcción de un modelo logit y la estimación de los estimadores y sus respectivos efectos marginales. Esa es precisamente la importancia que tienen los modelos de probabilidad, ya que permiten calcular los efectos de las regresoras sobre la probabilidad de éxito que tiene la variable dependiente.

El objetivo de la réplica es explicar mediante un ejemplo práctico la realización de la especificación, estimación, evaluación e interpretación de los resultados y contrastar los resultados obtenidos en el año 2018 con los resultados del trabajo de investigación en 2011 y poder observar los principales cambios sobre las probabilidades de ser pobre en el Perú.

4.5.1. Problema de la investigación.

4.5.1.1. Planteamiento del problema.

(Aparicio, Jaramillo, & San Román , 2011) Presentan el problema de la investigación desde dos puntos de vista: **Acceso a infraestructura y pobreza en el Perú** y **Crecimiento económico, y lucha contra la pobreza en el Perú.** El primer punto de visto habla sobre el nivel de acceso a la infraestructura que ha tenido la población peruana y cómo influye sobre la pobreza, tomando en cuenta el género del jefe de hogar y la ubicación del hogar.

Algunos resultados que obtuvieron (Aparicio, Jaramillo, & San Román , 2011) Son: en el año 2007 la pobreza en las zonas rurales representó el 64.6% y en el 2010 fue el 54.2%, mientras que en las zonas urbanas en el año 2007 el 25.7% de la población se le consideró dentro de la pobreza y hacia el año 2010 pasó a ser 19.1%. En la zona rural en el año 2010, el 38.0% de los hogares tuvieron acceso a agua potable, el 10.4% obtuvieron acceso al desagüe, el 59.5% de los hogares obtuvieron electricidad y el 52.5% confirmaron tener acceso al teléfono. En simultáneo, en la zona urbana en el mismo año, el 87.5% de los hogares tuvieron acceso a agua potable, el 83.0% afirmaron tener acceso a desagüe, el 98.4% tuvieron acceso a electricidad y el 91.2% de los hogares tuvieron

acceso a teléfono. Estos cuatro servicios básicos son definidos como la infraestructura que será analizada según (Aparicio, Jaramillo, & San Román , 2011).

El segundo punto de vista relaciona el crecimiento económico del Perú y su repercusión en los esfuerzos por bajar los índices de pobreza. Para comprender este punto de vista debemos definir en primer lugar el término "pobreza". Muy por el contrario que a lo que la cultura popular considera a través de libros de autoayuda sobre qué es realmente la pobreza, esta no se trata de un modo de vida que las personas han escogido debido a su "mediocridad" o "conformismo". Es impreciso concluir que la pobreza es el resultado de un conformismo dominante en las personas, de hecho, el término pobreza tiene múltiples definiciones provistas desde varias perspectivas de la teoría económica.

Según (Chacaltana, 2006) la pobreza es heterogénea y dinámica y tiene determinantes tanto de corto plazo como de largo plazo, los primeros se deben a shocks temporales mientras que los segundos están definidos como los efectos demográficos, acceso a diferentes activos productivos y otros factores sobre la productividad. El crecimiento económico es necesario para el financiamiento de políticas estructurales que conlleven a disminuir los determinantes de largo plazo porque de esta forma se logrará la reducción de la pobreza de forma significativa.

4.5.1.2. Objetivo general y objetivos específicos.

(Aparicio, Jaramillo, & San Román , 2011) Establecen el siguiente objetivo general:

- **Objetivo general**
  - Analizar la contribución de los distintos tipos de infraestructura sobre la disminución de la pobre de los hogares del Perú.

Del mismo modo, (Aparicio, Jaramillo, & San Román , 2011) Han trazado los siguientes objetivos específicos.

- **Objetivos específicos.**
  - Discutir los canales a través de los cuales la infraestructura contribuye a reducir la pobreza en el Perú.
  - Identificar cuáles son los tipos de infraestructura que generan los mayores impactos sobre la disminución de la pobreza en el Perú.

- Identificar si existe un impacto diferenciado de la infraestructura sobre la disminución de la pobreza de acuerdo al sexo del jefe de hogar y la zona en donde se encuentra ubicado el hogar (urbano o rural).

4.5.1.3. Planteamiento de la pregunta.

¿Cuánto afecta el nivel de infraestructura a la probabilidad de que los hogares sean pobres en el Perú en el año 2018?

4.5.2. Identificar el marco teórico.

4.5.2.1. Marco teórico.

Desarrollo de la infraestructura y reducción de la pobreza: el caso peruano (Aparicio, Jaramillo, & San Román , 2011)

Es importante definir con exactitud el concepto del término infraestructura, cuáles son los tipos de infraestructuras y cuál es su relación con la pobreza. Según (Reinikka & Svensson, 1999) La infraestructura es el capital que brinda servicios necesarios para la operación de actividades privadas, en el sentido de la investigación, puede ser vista como el factor complementario al capital privado de los hogares.

Desde un enfoque de activos, podemos identificar y medir cómo la infraestructura de los servicios hacia las viviendas tiene un impacto sobre la reducción de la pobreza, para entenderlo debemos saber cuáles son las principales entradas y salidas de la pobreza. En palabras de (Chacaltana, 2006) la pobreza es heterogénea y dinámica, lo que implica que podemos concluir que la situación de pobreza de una persona no es la misma a la de otra persona y que constantemente las personas entran y salen de la pobreza. (Attanasio & Székely, 2001) Desarrollaron el enfoque de activos que permite analizar la pobreza desde una perspectiva multidimensional. Este enfoque explica que las salidas de la pobreza pueden deberse a la acumulación de activos, esta visión se ajusta a la pobreza por ingresos cuyo concepto indica que una persona es pobre si el ingreso del hogar no permite solventar el gasto necesario para satisfacer sus necesidades básicas. A continuación, se muestra una fórmula desarrollada por (Attanasio & Székely, 2001) Que expresa el ingreso familiar per cápita.

$$y_i = \frac{\left[\sum_{i=1}^{j}\sum_{a=1}^{l} A_{a,i}R_{a,i}P_a\right]+\sum_{i=1}^{k} T_i}{n} \quad (4.5.1.)$$

Dónde: y es el ingreso per cápita para cada hogar, i es la variable que representa a cada hogar, A es la variable que representa el stock del activo, a representa a los activos del hogar, R es la variable que representa a la tasa de uso del activo, P es el valor en el mercado de cada activo y T son las transferencias recibidas por cada hogar, la variable j es el número de individuos de cada hogar, l es el número de activos que posee cada hogar, k es el número de miembros de hogar que obtienen remesas y n es el tamaño del hogar del hogar.

(Attanasio & Székely, 2001) Especifican que la ecuación (4.5.1.) muestra componentes de corto y largo plazo de la pobreza. Si existen factores que afectan a T_i entonces estos serán efectos de corto plazo, mientras tanto, si A_i, R_i y P_i están afectados, sus efectos serán de largo plazo. Entonces, el objetivo fundamental sería concentrarse en los componentes A_i, R_i y P_i ya que así la reducción de la pobreza será más profunda y permanente. Los autores clasifican a los activos en tres categorías: **capital humano**, comprendido como las habilidades y conocimientos para producir un bien o servicio que permita generar ingresos; **capital físico**, son los valores monetarios de cualquier forma de activo financiero, propiedad o stock de capital usado en la producción, y **capital social**, está relacionado a un set de normas y redes sociales que facilitan la acción colectiva de los individuos. Por lo tanto, aquellas políticas que contribuyan a los activos del capital humano, físico y social, serán las más favorables para reducir la pobreza en el largo plazo, además, estas políticas deben estar acompañadas con medidas que eliminen las restricciones que impidan a los pobres acumular estos tipos de activos; algunas restricciones que enfrentan los pobres son el acceso al ingreso y al crédito y la incertidumbre generada por las asimetrías de la información.

Este estudio se centrará principalmente en el capital físico, según (Attanasio & Székely, 2001) Este tipo de capital, está subdividido en **capital físico privado**, referido a la tenencia de la vivienda y de bienes duraderos como refrigeradoras, teléfono, radio, etc., y el **capital físico público** el cual está relacionado con el acceso a distintos bienes y servicios públicos dentro o fuera del hogar como el agua, desagüe, telecomunicaciones y electricidad. El capital físico público comprende un factor importante en este análisis, ya que son considerados como activos físicos que permiten que el hogar genere ingresos. La tenencia del teléfono público y/o cualquier otro electrodoméstico puede ser vista tanto como capital físico privado y público. El acceso de estos tipos de infraestructuras puede incrementar el valor de la tasa de R_i y contribuye a mejorar el capital humano de los

hogares. Por último, el acceso a la infraestructura permite la disminución de los gastos de los hogares y aumentar un consumo corriente del hogar o una mayor compra de cantidad de activos que generen ingresos.

Tal como se ha mencionado, los impactos de la infraestructura son a largo plazo debido a que permiten acumular activos que generan ingresos al hogar, sin embargo, es posible que la infraestructura pueda tener efectos a corto plazo pero estos efectos dependen de la decisión del jefe de hogar en utilizar los ahorros para incrementar los ingresos o en decidir utilizar los ahorros para adquirir mejoras en los activos. Es más complicado de lo que aparenta, la decisión en utilizar el ahorro en incrementar el consumo o en adquirir activos depende de factores que a menudo son difíciles de cuantificar como la idiosincrasia de los miembros del hogar, entre otros. La información recaudada sobre los hogares del Perú nos puede ofrecer datos sobre las siguientes infraestructuras: agua, desagüe, electricidad y teléfono.

- **Agua.**

Se considera al hogar que cuenta con acceso a servicios de agua de potable mediante una red pública dentro o fuera de la vivienda.

- **Desagüe**.

Se toman en cuenta a aquellos hogares que tienen acceso a servicios de desagüe mediante una red pública dentro o fuera de la vivienda. Es importante mencionar, que en algunas zonas rurales es frecuente encontrar viviendas con pozos sépticos, no obstante, aquellas viviendas no serán tomadas en cuenta debido a que los pozos sépticos representan una amenaza al bienestar de los miembros del hogar por ser considerados como focos de infecciones.

- **Electricidad**.

En este estudio se tomará información de aquellas viviendas con acceso a electricidad del tipo alumbrado provista desde la red pública de energía eléctrica. El uso de generadores no supone que el hogar sea considerado.

- **Teléfono**.

Si los hogares tienen acceso a los servicios de telefonía fijo y/o móvil entonces serán tomados en cuenta para este estudio.

Cada servicio (agua, desagüe, electricidad y teléfono) tiene una forma de impactar sobre la reducción en la pobreza. En el caso del acceso al agua potable y desagüe, estos servicios sugieren que permiten consolidar el capital humano de los pobres, ya que estos incrementan la productividad de sus trabajadores y además contribuyen a la disminución de costos sobre la compra de agua de cisternas o bidones, este ahorro es importante en los hogares considerados como pobres. En cuanto al servicio de electricidad, a este se le considera directamente como una fuente primordial de energía y por ello constituye ser un activo y/o insumo relevante para la producción en zonas rurales, entonces el acceso al servicio de electricidad le permite aumentar sus ingresos a las personas y mejora el capital social de los hogares. Por último, el acceso al servicio de telecomunicaciones está relacionado al incremento en el número de clientes, debido a que la tenencia de telefonía permite el incremento en la tasa de los activos que posee el hogar.

4.5.3. Especificación del modelo econométrico.

(Aparicio, Jaramillo, & San Román , 2011) Han tomado en cuenta que el uso de los modelos econométricos debe estar justificado en medir los impactos de la infraestructura sobre la pobreza en el corto y largo plazo. Del mismo modo, los impactos de la infraestructura deben ser medidos y comprobados según el lugar de residencia y el género del jefe de hogar. (Aparicio, Jaramillo, & San Román , 2011) Especifican dos modelos econométricos, uno de corte transversal para analizar los efectos a corto plazo y el segundo será de datos de panel para recoger los componentes de largo plazo, en este caso se analizará solamente el modelo de corte transversal.

Revisando la teoría anteriormente explicada se puede expresar lo siguiente.

$$C = f(A_H, A_F, A_P, A_S, T, \psi, X) \quad (4.5.2.)$$

Donde C representa el consumo que mide el bienestar del hogar, A_H representa todos los tipos de capital humano, A_F recoge los impactos de todos los tipos de capital físico, A_P representa todos los tipos de capital físico público, A_S es la variable que recoge los efectos de los tipos de capital social, T es la variable que recoge las transferencias que recogió el hogar, ψ son los shocks que enfrenta el hogar y X son las características del jefe de hogar, los miembros del hogar. Entonces, el análisis pretende analizar principalmente la relación entre A_P y C.

A partir de (4.5.2.) podemos construir el siguiente modelo econométrico para medir los efectos de la infraestructura sobre el nivel de la pobreza.

$$Y_i = \begin{cases} 1 \; si \; el \; individuo \; (i) \; es \; pobre \\ 0 \; de \; otro \; modo \end{cases} \quad (4.5.3.)$$

$$Y_i = \beta_1 + \beta_2 A_{Hi} + \beta_3 A_{Fi} + \beta_4 A_{Pi} + \beta_5 A_{Si} + \beta_6 T_i + \beta_7 \psi_i + \beta_8 X + \mu_i \quad (4.5.4.)$$

$$E(Y_i/R) = Pr(Y_i = 1) = G(\beta_1 + \beta_2 A_{Hi} + \beta_3 A_{Fi} + \beta_4 A_{Pi} + \beta_5 A_{Si} + \beta_6 T_i + \beta_7 \psi_i + \beta_8 X) \quad (4.5.5.)$$

Donde Y_i representa la pobreza, R representa a todas las regresoras, μ_i es el término de error y debido a su distribución acumulada logística, el modelo se estimará mediante un modelo Logit, $G(.)$ es la función de distribución acumulada logística. Este modelo será estimado siguiendo el método de estimación MV, el cual contiene las siguientes regresoras.

- A_{Pi} es el vector que recoge los tipos de capital público.
  - Agua potable.
  - Desagüe.
  - Electricidad.
  - Teléfono.
- A_{Hi} es el vector que recoge los tipos de capital humano en el hogar.
  - Nivel educativo del jefe de hogar (primaria completa, secundaria completa y superior completa).
- A_{Fi} es el vector que recoge los tipos de capital físico privado.
  - Títulos de propiedad de la vivienda, cocina, auto, camión y número de habitaciones de la vivienda.
- A_{Si} es el vector que toma los efectos de los tipos de capital social.
  - Pertenencia a alguna asociación productiva.
- X_i es el vector que recoge las características del jefe de hogar, miembros del hogar y hogar, que influyen sobre la capacidad de generar ingresos.
  - Número de miembros del hogar.
  - Número de perceptores de ingresos en el hogar.
  - La edad.
  - Edad al cuadrado del jefe de hogar.
  - Lengua materna del jefe de hogar (si es lengua nativa).

- Sector de trabajo del jefe de hogar (comercio)
- Zona donde se ubica el hogar (Rural o Urbano)

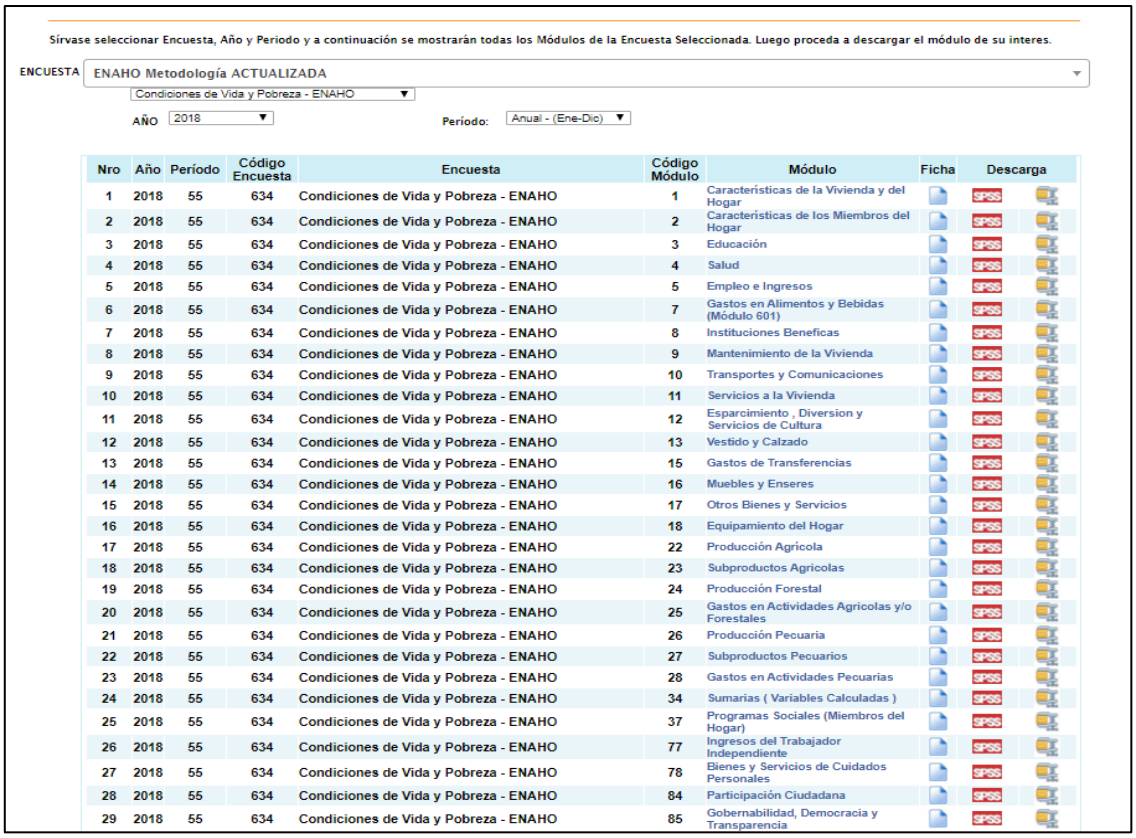
- T_i es el vector que indica las transferencias.
  - Transferencias varias y transferencias de jubilación.
- ψ_i es el vector que representa los shocks que tienen que enfrentar los hogares.
  - Shocks varios y desastres naturales.

4.5.4. Acceso a la base de datos.

En esta sección se mostrarán los pasos para construir la base de datos que se usará para la estimación del modelo Logit del modelo (4.5.4.), desde la consolidación de la base de datos hasta la creación de variables relevantes. Para una mejor comprensión se dividirá en dos partes, la primera contendrá la explicación sobre la construcción de una base de datos unificada consolidada con los módulos necesarios para obtener la información requerida y la segunda parte hablará sobre la creación de las variables regresoras y variable dependiente para la especificación y estimación del modelo.

4.5.4.1. Construcción de la base de datos consolidada.

Para empezar, debemos ingresar a las Consultas por Encuesta de la Encuesta Nacional de Hogares brindado por el Instituto Nacional de Estadística e Informática en el siguiente link: http://iinei.inei.gob.pe/microdatos/.



Sírvase seleccionar Encuesta, Año y Periodo y a continuación se mostrarán todas los Módulos de la Encuesta Seleccionada. Luego proceda a descargar el módulo de su interes.

ENCUESTA: ENAHO Metodología ACTUALIZADA — Condiciones de Vida y Pobreza - ENAHO

AÑO: 2018 — Período: Anual - (Ene-Dic)

Nro	Año	Período	Código Encuesta	Encuesta	Código Módulo	Módulo	Ficha	Descarga
1	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	1	Características de la Vivienda y del Hogar		
2	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	2	Características de los Miembros del Hogar		
3	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	3	Educación		
4	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	4	Salud		
5	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	5	Empleo e Ingresos		
6	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	7	Gastos en Alimentos y Bebidas (Módulo 601)		
7	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	8	Instituciones Beneficas		
8	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	9	Mantenimiento de la Vivienda		
9	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	10	Transportes y Comunicaciones		
10	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	11	Servicios a la Vivienda		
11	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	12	Esparcimiento , Diversion y Servicios de Cultura		
12	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	13	Vestido y Calzado		
13	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	15	Gastos de Transferencias		
14	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	16	Muebles y Enseres		
15	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	17	Otros Bienes y Servicios		
16	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	18	Equipamiento del Hogar		
17	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	22	Producción Agricola		
18	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	23	Subproductos Agricolas		
19	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	24	Producción Forestal		
20	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	25	Gastos en Actividades Agricolas y/o Forestales		
21	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	26	Producción Pecuaria		
22	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	27	Subproductos Pecuarios		
23	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	28	Gastos en Actividades Pecuarias		
24	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	34	Sumarias (Variables Calculadas)		
25	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	37	Programas Sociales (Miembros del Hogar)		
26	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	77	Ingresos del Trabajador Independiente		
27	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	78	Bienes y Servicios de Cuidados Personales		
28	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	84	Participación Ciudadana		
29	2018	55	634	Condiciones de Vida y Pobreza - ENAHO	85	Gobernabilidad, Democracia y Transparencia		

Figura 3.2. Consulta por Encuesta de ENAHO.

Después de ingresar en el URL, descargamos los siguientes módulos.

- Características de la vivienda y del hogar.
- Equipamiento del hogar.
- Participación ciudadana.
- Características de los miembros del hogar y Educación.
- Empleo e ingresos.
- Sumaria.

Con el fin de ser ordenados y no perder el hilo a la hora de construir la base de datos se recomienda hacer el uso de una carpeta exclusivamente para el modelo que se pretende construir. Para este ejemplo se ha considerado la siguiente carpeta “Aplicación t” y en esta se encuentran las siguientes carpetas: “1. Input t”, “2. Temp t”, “3. Output t” y “4. Dofile t”. Veámoslo.

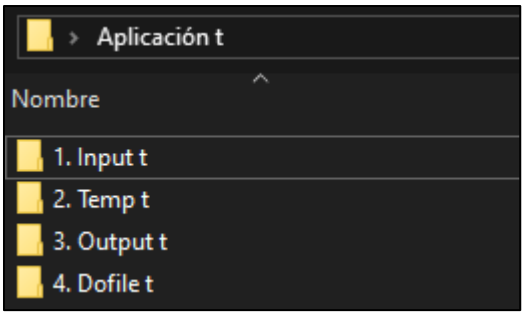


Figura 4.2. Carpeta "Aplicación t"

La carpeta "1. Input t" se usará para albergar los archivos descargados de STATA, la carpeta "2. Temp t" se usará para guardar las bases transformadas de ENAHO, "3. Output t" corresponde a las bases de datos finales y "4. Dofile t" contendrá los Dofile que se crea conveniente. Una vez descargados los archivos de base de datos anteriormente mencionados obtendremos los siguientes archivos de STATA.

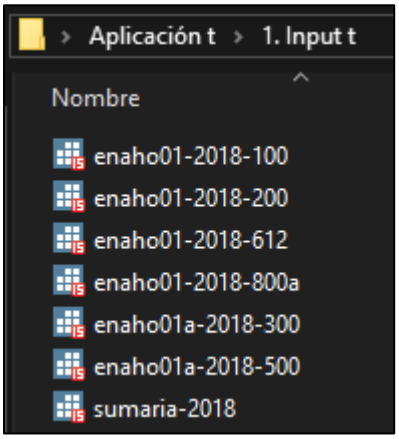


Figura 4.3. Carpeta "Input t"

Ahora entramos a STATA y configuramos el cambio del directorio. El directorio se refiere a la dirección donde se guardan de forma predeterminada cada archivo, Dofile, archivo log, gráficos, etc. De forma predeterminada está configurada para que el directorio se encuentre en donde está instalado el programa, y puede ser configurado en "Archivo" y posteriormente "Cambiar directorio de trabajo…".

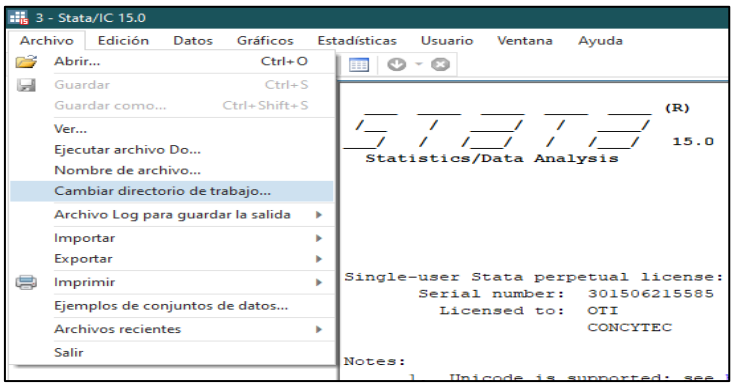


Figura 4.4. "Cambiar directorio de trabajo…"

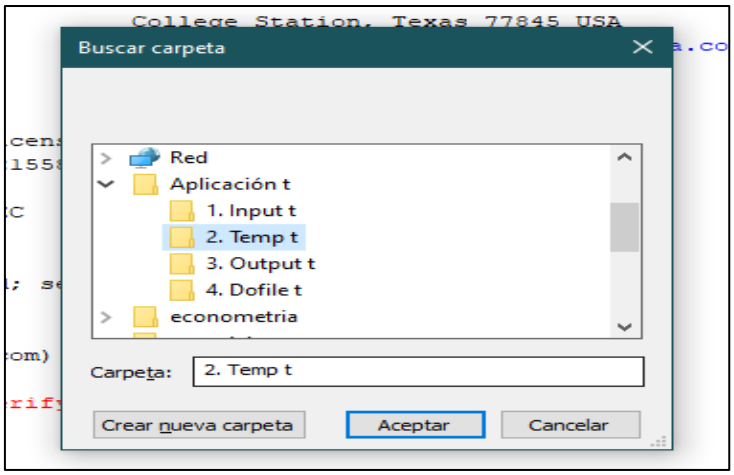


Figura 4.5. Selección de la carpeta "2. Temp t" como directorio de trabajo (1).

Aparecerá la siguiente instrucción en la pantalla de resultados que indica que el cambio de dirección ha sido realizado con éxito como se ve la figura 4.6.

```
. cd "C:\Users\USUARIO\Desktop\Aplicación t\2. Temp t"
C:\Users\USUARIO\Desktop\Aplicación t\2. Temp t
```

Figura 4.6. Selección de la carpeta "2. Temp t" como directorio de trabajo (2).

Ahora abriremos el archivo "enaho01-2010-100.dta" que representa al módulo "Características de la vivienda y del hogar" y permite tener información sobre variables relacionadas a los servicios básicos (agua, desagüe, electricidad y teléfono), estrato, número de habitaciones, título de propiedad de la vivienda y presencia de la cocina, el dominio y el factor de expansión. Con el comando **gen** crearemos la variable ***t*** que indica el año, en este caso será 2018 y con el comando **keep** seleccionaremos de forma rápida las variables que nos interesan de la base de datos.

```
. gen t=2018
. keep t conglome vivienda hogar dominio estrato p104 p104a p110 p111 p1121 p1141 p1142 p106a p1138 factor07
```

Figura 4.7. Generación de variables sobre las características de la vivienda y del hogar (1).

Terminamos el uso de este archivo con el comando **save** y el comando **replace**. Y el nombre con el que guardaremos el archivo será "hogar2018t.dta"

```
. save hogar2018t,replace
```

Figura 4.8. Generación de variables sobre las características de la vivienda y del hogar (2).

Proseguimos con el archivo “enaho01-2018-612.dta” descargado del módulo “Equipamiento del hogar”. De este archivo obtendremos información si las viviendas cuentan con camión o auto y otras variables. Empezamos generando la variable ***t*** que indique el año 2018.

```
. gen t=2018
```

Figura 4.9. Generación de una variable que indica el año 2018.

Para saber cuáles son las variables que contienen la información sobre camión y auto realizaremos una tabla con el comando **tab** y las variables ***p612n*** y ***p612***, la primera variable indica el tipo de artefacto y la segunda variable muestra si el hogar los tiene o no, esto se complementará con la opción **missing.**

```
. tab p612n p612, missing
```

equipamiento del hogar	□su hogar tiene:? si	no	.	Total
radio	20,611	16,116	735	37,462
tv a color	26,868	9,973	621	37,462
tv blanco y negro	1,869	34,858	735	37,462
equipo de sonido	10,741	26,030	691	37,462
dvd	12,048	24,679	735	37,462
video grabadora	223	36,504	735	37,462
computadora/laptop	10,466	26,339	657	37,462
plancha electrica	17,709	19,122	631	37,462
licuadora	20,485	16,335	642	37,462
cocina a gas	30,997	5,844	621	37,462
cocina a kerosene	80	36,643	739	37,462
refrigeradora/congela	16,463	20,354	645	37,462
lavadora de ropa	8,195	28,587	680	37,462
horno microondas	5,256	31,503	703	37,462
m□quina de coser	2,943	33,780	739	37,462
bicicleta	5,923	30,809	730	37,462
auto, camioneta	3,640	33,083	739	37,462
motocicleta	5,082	31,641	739	37,462
triciclo	433	36,290	739	37,462
mototaxi	2,389	34,334	739	37,462
cami□n	222	36,501	739	37,462
otro	11,211	25,512	739	37,462
otro	5,038	31,685	739	37,462
otro	2,309	34,414	739	37,462
otro	1,127	35,596	739	37,462
otro	532	36,191	739	37,462
Total	222,860	732,723	18,429	974,012

Figura 4.10. Tabla sobre las variables ***p612n*** y ***p612*** (1).

Para reducir el número de observaciones utilizaremos el comando **tab** y la variable ***p612n***, generando la variable ***p612a*** con la opción **gen.**

```
. tab p612n, gen(p612a)
```

equipamiento del hogar	Freq.	Percent	Cum.
radio	37,462	3.85	3.85
tv a color	37,462	3.85	7.69
tv blanco y negro	37,462	3.85	11.54
equipo de sonido	37,462	3.85	15.38
dvd	37,462	3.85	19.23
video grabadora	37,462	3.85	23.08
computadora/laptop	37,462	3.85	26.92
plancha electrica	37,462	3.85	30.77
licuadora	37,462	3.85	34.62
cocina a gas	37,462	3.85	38.46
cocina a kerosene	37,462	3.85	42.31
refrigeradora/congeladora	37,462	3.85	46.15
lavadora de ropa	37,462	3.85	50.00
horno microondas	37,462	3.85	53.85
m□quina de coser	37,462	3.85	57.69
bicicleta	37,462	3.85	61.54
auto, camioneta	37,462	3.85	65.38
motocicleta	37,462	3.85	69.23
triciclo	37,462	3.85	73.08
mototaxi	37,462	3.85	76.92
cami□n	37,462	3.85	80.77
otro	37,462	3.85	84.62
otro	37,462	3.85	88.46
otro	37,462	3.85	92.31
otro	37,462	3.85	96.15
otro	37,462	3.85	100.00
Total	974,012	100.00	

Figura 4.11. Tabla sobre las variables ***p612n*** y ***p612*** (2).

Para seleccionar solamente las variables ***p612a17*** y ***p612a21***, que muestran información sobre las variables ***auto*** y ***camión***, haremos uso del comando **keep**.

```
. keep if p612a17==1 | p612a21==1
(899,088 observations deleted)
```

Figura 4.12. Comando **keep**.

Ahora haremos dos bases de datos, donde cada uno concentrará información de cada variable (***p612a17*** y ***p612a21***) y los uniremos con el comando **merge**. Empezamos con el comando **preserve** y con el comando **keep** seleccionamos las variables que nos interesan, en esta primera base de datos son ***conglome***, ***vivienda***, ***hogar***, ***p612*** y ***p612a17*** cuando sea igual a 1.

```
                . preserve
. keep if p612a17==1
(37,462 observations deleted)

.
. keep t conglome vivienda hogar p612
```

Figura 4.13. Selección de las variables que nos interesan.

Se puede renombrar a una variable con el comando **rename**, renombramos la variable ***p612*** por ***p612_auto***.

```
. rename p612 p612_auto
```

Figura 4.14. Cambiando el nombre de la variable ***p612***.

Y guardamos la primera base con el comando **save** con el nombre "transporte2018_1t" y la opción **replace**.

```
. save transporte2018_1t, replace
```

Figura 4.15. Guardando la primera base de datos.

Para restaurar la base de datos anterior a los cambios que hicimos con el comando **keep** usaremos el comando **restore** y la opción **preserve**.

```
. restore, preserve
```

Figura 4.16. Restaurando la base de datos original.

Y con el comando **keep** seleccionamos las variables que nos interesan.

```
. keep if p612a21==1
(37,462 observations deleted)

.
. keep conglome vivienda hogar p612
```

Figura 4.17. Seleccionando variables.

Para distinguirnos de la base de datos que muestra información sobre los hogares con auto, se debe renombrar la variable ***p612*** por ***p612_camion*.**

```
. rename p612 p612_camion
```

Figura 4.18. Renombrando la variable ***p612***.

Y guardaremos esta base de datos con el nombre “transporte2018t_2” con el comando **save** y la opción **replace**.

```
. save transporte2018t_2, replace
```

Figura 4.19. Guardando la segunda base de datos.

Restauremos la base de datos con el comando **restore** y posteriormente con el comando **use** indicaremos a STATA que utilice el archivo “transporte2018_1t” con la opción **clear**.

```
. restore

. use transporte2018_1t,clear
```

Figura 4.20. Usando el archivo “transporte2018_1t”.

Ahora uniremos esta base de datos con la base de datos “transporte2018t_2” con el comando **merge**.

```
. merge 1:1 conglome vivienda hogar using transporte2018t_2
(label p612 already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            37,462  (_merge==3)
    -----------------------------------------
```

Figura 4.21. Comando **merge**.

Eliminemos la variable ***merge*** con el comando **drop** y guardemos la base de datos con el nombre “transporte2018t.dta” con el uso del comando **save** y la opción **replace**.

```
. drop _merge

. save transporte2018t.dta, replace
```

Figura 4.22. Guardando la base de datos “transporte2018t.dta”.

Como ya no son necesarios las bases de datos “transporte2018_1t” y “transporte2018t_2” pueden ser eliminados con el comando **erase**.

```
. erase transporte2018_1t.dta

. erase transporte2018t_2.dta
```

Figura 4.23. Guardando la base de datos “transporte2018t.dta”.

El siguiente módulo con el que trabajaremos es “Participación ciudadana” cuyo archivo de STATA es “enaho01-2018-800a” y tomaremos a aquellas variables que nos brindan información sobre las relaciones de pertenencia de algún miembro del hogar con alguna asociación productiva. Empezamos generando la variable ***t*** que representa el año 2018 y la variable ***as*** que suma las variables ***p801_4***, ***p801_5***, ***p801_6***, ***p801_7*** y ***p801_8***.

```
. gen t=2018

. gen as=p801_4+p801_5+p801_6+p801_7+p801_8
(1,553 missing values generated)
```

Figura 4.24. Creando las variables ***t*** y ***as***.

Estas variables representan si algún miembro del hogar pertenece a las siguientes asociaciones respectivamente: asociación vecinal/ junta vecinal, ronda campesina, asociación de regantes, asociación profesional y asociación de trabajadores o sindicato. Ahora crearemos la variable dummy ***asociación*** que será igual a “0” si ***as*** es 0 y “1” si ***as*** es mayor de 0, con los comandos **gen** y **replace**.

```
. gen asociacion=0 if as==0
(8,397 missing values generated)

.
. replace asociacion=1 if as>0
(8,397 real changes made)
```

Figura 4.25. Creando la variable ***asociacion***.

Con el comando **label** y el componente **define** creamos etiquetas a cada valor de la variable anteriormente creada.

```
. label define a 0 "No Pertenece" 1 "Pertenece"
```

Figura 4.26. Otorgando etiquetas a los valores de la variable creada.

Finalizamos seleccionando solamente las variables ***t***, ***conglome***, ***vivienda***, ***hogar*** y ***asociación*** y guardamos la base de datos con el nombre “asociacion2018t”

```
. keep t conglome vivienda hogar asociacion

. save asociacion2018t.dta, replace
file asociacion2018t.dta saved
```

Figura 4.27. Otorgando etiquetas a los valores de la variable creada.

Los siguientes módulos con los que trabajaremos son “Características del miembro del hogar” y “Educación” con los archivos de STATA “enaho01-2018-200” y “enaho01a-2018-300” respectivamente. Empezamos abriendo el archivo “enaho01-2018-200”, generamos la variable ***t*** que representa el año 2018 y mantenemos solamente a las variables ***t, conglome***, ***vivienda***, ***hogar***, ***codperso***, ***p207***, ***p208a*** y los valores 1 de la variable ***p203*** para elegir solamente a los jefes de hogar.

```
. gen t=2018


.
. keep if p203==1
(102,195 observations deleted)

. keep t conglome vivienda hogar codperso p207 p208a
```

Figura 4.28. Otorgando etiquetas a los valores de la variable creada.

Guardamos la base de datos con el nombre “jefe_hogar2018t”.

```
. save jefe_hogar2018t.dta, replace
```

Figura 4.29. Guardando la base de datos “jefe_hogar2018t”.

Se procede a abrir el archivo STATA correspondiente al módulo “Educación” cuyo nombre es “enaho01a-2018-300” utilizando el comando **use** y su opción **clear** con el fin de consolidar la base de datos con el archivo “jefe_hogar2018” usando el comando **merge**.

```
. merge n:n conglome vivienda codperso using jefe_hogar2018t.dta
(label p207 already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                        89,299
        from master                    89,299  (_merge==1)
        from using                          0  (_merge==2)

    matched                            37,462  (_merge==3)
    -----------------------------------------
```

Figura 4.30. Uniendo las bases de datos “enaho01a-2018-300” y “jefe_hogar2018”.

Mantenemos la variable ***merge*** solamente cuando vale 3 y posteriormente lo eliminamos con el comando **drop** y mantenemos las variables ***t***, ***conglome***, ***vivienda***, ***hogar***, ***p207***, ***p208a***, ***p300a*** y ***p301a***.

```
. keep if _merge==3
(89,299 observations deleted)

.
. drop _merge
. keep t conglome vivienda hogar p207 p208a p300a p301a
```

Figura 4.31. Manteniendo variables.

Guardamos la base de datos creada con el nombre "datos_generales2018t" y eliminamos la base de datos "jefe_hogar2018t".

```
. save datos_generales2018t.dta, replace

. erase jefe_hogar2018t.dta
```

Figura 4.32. Guardando la base de datos "datos_generales2018t".

Ahora utilizaremos el archivo de STATA "enaho01a-2018-500" descargado del módulo "Empleo e ingresos" para recoger información sobre las transferencias, para esto utilizaremos las variables que empiezan con ***p556***. Si cada variable tiene un valor igual a 2 o un valor perdido, entonces lo reemplazamos con 0 con el comando **replace**.

```
. replace p5561a=0 if p5561a==2|p5561a==.
(99,553 real changes made)

.
. replace p5562a=0 if p5562a==2|p5562a==.
(97,181 real changes made)

.
. replace p5563a=0 if p5563a==2|p5563a==.
(93,717 real changes made)

.
. replace p5564a=0 if p5564a==2|p5564a==.
(96,287 real changes made)

.
. replace p5565a=0 if p5565a==2|p5565a==.
(98,681 real changes made)

.
. replace p5566a=0 if p5566a==2|p5566a==.
(94,877 real changes made)

.
. replace p5567a=0 if p5567a==2|p5567a==.
(95,533 real changes made)
 . replace p5568a=0 if p5568a==2|p5568a==.
 (97,531 real changes made)
```

Figura 4.33. Comando **replace**.

Creamos la variable ***tr_1*** cuya función es la suma de todas las variables que empiezan con ***p556.***

```
. gen tr_1=p5561a+p5562a+p5563a+p5564a+p5565a+p5566a+p5567a+p5568a
```

Figura 4.34. Generando la variable ***tr***.

Con el comando **recode** y la opción **gen** crearemos la variable ***transferencia***. La variable ***transferencia*** será igual a 0 si, la variable ***tr*** es igual a "0", y tendrá la etiqueta "El hogar no recibió transferencia de ningún tipo", por el contrario si la variable ***tr*** es igual a 1 o un valor perdido, entonces la variable ***transferencia*** tendrá un valor igual a 1 y la etiqueta "El hogar recibió transferencias de distinto tipo".

```
. recode tr_1 (0=0 "El hogar no recibió transferencias de ningún tipo") (1/.=1 "El hogar recibió transfer
> encias de distinto tipo"), gen(transferencias)
(2390 differences between tr_1 and transferencias)
```

Figura 4.35. Generando la variable ***transferencias***.

Ahora renombramos la variable ***p5564a*** por ***transferencias_jub*** y con el comando **label** y su componente **variable** le pondremos la etiqueta "El hogar ha recibo transferencias de la jubilación".

```
. rename p5564a transferencias_jub

.
. label variable transferencias_jub "El hogar recibió transferencias de la jubilación"
```

Figura 4.36. Creando la variable ***transferencias_jub***.

Ahora nos quedamos solamente con los jefes de hogar con la variable ***p203*** igual a 1.

```
. keep if p203==1
(62,184 observations deleted)
```

Figura 4.37. Seleccionando a los jefes de hogar.

Y mantenemos a las variables ***conglome***, ***vivienda***, ***hogar***, ***codperso***, ***transferencias*** y ***transferencias_jub*.**

```
. keep conglome vivienda hogar codperso transferencias transferencias_jub
```

Figura 4.38. Manteniendo variables importantes.

Guardamos la base de datos con el nombre "transferencias2018t"

```
. save transferencias2018t.dta, replace
```

Figura 4.39. Guardando el archivo con el comando **save**.

Después, abrimos el archivo “sumaria-2018”, que corresponde a la información del módulo “Sumaria”. Este módulo es muy importante en la mayoría de trabajos de investigación porque permite obtener información sobre variables calculadas ligadas al gasto e ingreso de las familias. Generamos la variable ***t*** que indica el año 2018, renombramos a las variables ***inghog2d*** y ***gashog2d*** por ***ingreso_total*** y ***gasto_total***, respectivamente. Después mantenemos las variables ***conglome***, ***vivienda***, ***hogar***, ***totmieho***, ***pobreza***, ***línea***, ***linpe***, ***ingreso_total*** y ***gasto_total.***

```
. gen t=2018

.
. rename inghog2d ingreso_total

.
. rename gashog2d gasto_total

.
. keep conglome vivienda hogar totmieho pobreza linea linpe ingreso_total gasto_total
```

Figura 4.40. Utilizando el archivo “sumaria-2018”.

Guardamos esta base de datos con el nombre “sumaria2018t”

```
. save sumaria208t.dta,replace
```

Figura 4.41. Guardando el archivo “sumaria-2018t”.

Ahora uniremos todos los archivos guardados con el comando **merge**. Cada base de datos con que uniremos a la base de datos “sumaria208t” seguirán la misma instrucción: mantener los valores de cada variable ***merge*** si son iguales a 3 y eliminaremos la variable ***merge***. Veámoslo.

```
. merge 1:1 conglome vivienda hogar using datos_generales2018t.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            37,462  (_merge==3)
    -----------------------------------------

.
. keep if _merge==3
(0 observations deleted)

.
. drop _merge
```

Figura 4.42. Agregando el archivo “datos_generales2018t” a la base de datos.

```
. merge 1:1 conglome vivienda hogar using hogar2018t.dta
(label dominio already defined)
(label estrato already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                        10,238
        from master                         0  (_merge==1)
        from using                     10,238  (_merge==2)

    matched                            37,462  (_merge==3)
    -----------------------------------------

. keep if _merge==3
(10,238 observations deleted)

. drop _merge
```

Figura 4.43. Agregando el archivo “hogar2018t” a la base de datos.

```
. merge 1:1 conglome vivienda hogar using asociacion2018t.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            37,462  (_merge==3)
    -----------------------------------------

.
. keep if _merge==3
(0 observations deleted)

.
. drop _merge
```

Figura 4.44. Agregando el archivo “asociacion2018t” a la base de datos.

```
. merge 1:n conglome vivienda hogar using transporte2018t.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            37,462  (_merge==3)
    -----------------------------------------

.
. keep if _merge==3
(0 observations deleted)

.
. drop _merge
```

Figura 4.45. Agregando el archivo “transporte2018t” a la base de datos.

```
. merge 1:n conglome vivienda hogar using transferencias2018t.dta

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                            37,462  (_merge==3)
    -----------------------------------------

.
. keep if _merge==3
(0 observations deleted)

.
. drop _merge
```

Figura 4.46. Agregando el archivo "transferencias2018t" a la base de datos.

Para finalizar esta primera parte de la sección guardamos la base de datos consolidada con el nombre de "consolidado2018t", si revisamos el número de observaciones de esta base de datos podemos ver que son 37462. La importancia que tiene esta base de datos radica en que contiene todas las variables que se usarán para crear las variables necesarias para estimar el modelo (4.5.5.).

```
. save consolidado2018t.dta
file consolidado2018t.dta saved
```

Figura 4.47. Guardando el archivo "consolidado2018t".

4.5.4.2. Creación de las variables regresoras y de la variable dependiente.

En esta segunda sección se explicarán los procesos para generar las variables necesarias que se utilizarán en la estimación del modelo. Previamente debemos introducir el comando **preserve,** ya que habrán modificaciones a la base de datos y será necesario restaurar la base de datos original.

```
. preserve
```

Figura 4.48. Comando **preserve**.

- **Variable dependiente: *niv_ pobreza.***

Como se ha definido la variable dependiente es ***niv_pobreza***, es una variable dummy que toma el valor de 1 cuando la variable ***pobreza*** es igual a 1 o 2 y toma el valor de 0 cuando la variable ***pobreza*** es igual a 3.

```
. recode pobreza (1/2=1 "Pobre") (3=0 "No pobre"), gen(niv_pobreza)
(36393 differences between pobreza and niv_pobreza)

.
. label variable niv_pobreza "Nivel de Pobreza"
```

Figura 4.49. Variable dependiente ***niv_pobreza***.

- **Variables independientes: Acceso a servicios básicos.**
  - **Agua.**

La información sobre el tipo de acceso al agua de un hogar se encuentra en la variable ***p110***, por ello, a partir de dicha variable se creará la variable ***agua*** que tendrá el valor igual a 1 cuando la variable ***p110*** sea igual a 1 o 2 mientras será igual a 0 cuando ***p110*** sea igual a 3, 4, 5, 6, 7 o 8. Los comandos usados serán **recode** y la opción **gen** para crear la variable ***agua*** con las recodificaciones especificadas anteriormente, del mismo modo con el comando **label** y su componente **variable** se le otorgará la etiqueta a la variable ***agua***. Muy parecida a la figura 4.49. De hecho, estos comandos serán usados frecuentemente en casi todas las variables regresoras. Después de haber creado la variable ***agua***, ya no será necesaria la variable ***p110***, entonces la eliminaremos para evitar confusiones y llenar la base de datos con variables que no utilizaremos para la estimación.

```
. recode p110 (1/2=1 "Red Publica") (3/8=0 "Otros"), gen(agua)
(7651 differences between p110 and agua)

.
. label variable agua "Agua potable por red publica"

.
```

Figura 4.50. Variable independiente ***agua***.

  - Desagüe

Del mismo modo, la variable ***desague*** se creará con la variable ***p111*** y se recodificará sus valores de la siguiente forma: si ***p111*** es igual a 1 o 2 entonces la variable ***desague*** será igual a 1 y si ***p111*** es igual a 3, 4, 5, 6, 7, 8 o 9 entonces la variable ***desague***

será igual a 0. Con el comando **label** y el componente **variable** se le dará la etiqueta a la variable ***desague***.

```
. recode p111 (1/2=1 "Red Publica") (3/9=0 "Otros"), gen(desague)
(15925 differences between pllla and desague)

.
. label variable desague "Servicio de desague por red publica"

.
. drop p111
```

Figura 4.51. Variable independiente ***desague***.

- **Electricidad**.

Con la variable ***p1121*** se hará la variable ***electricidad*** renombrándola y otorgándole etiquetas a la variable y a sus valores. En este caso ya no es necesario recodificar sus valores porque ya es una variable Dummy.

```
. rename p1121 electricidad

.
. label variable electricidad "Acceso al servicio de electricidad"

.
. label define a 0 "Otros" 1 "Electricidad"

.
. label values electricidad a
```

Figura 4.52. Variable independiente ***electricidad***.

- **Teléfono.**

Se debe crear la variable ***x*** que representa la suma de las variables ***p1141*** y ***p1142***. Con la variable ***x*** crearemos la variable ***telefono*** que tendrá valores igual a 1 cuando la variable ***x*** sea igual a 1 o 2, e igual a 0 cuando la variable ***x*** sea igual a 0. Le otorgamos una etiqueta y borramos las variables ***p1141***, ***p1142*** y ***x***.

```
. gen x=p1141+p1142

.
. recode x (1/2=1 "Telefonia fija o movil") (0=0 "Ninguna"), gen(telefono)
(5196 differences between x and telefono)

.
. label variable telefono "Acceso a telefonia fija o móvil"

.
. drop p1141 p1142 x
```

Figura 4.53. Variable independiente ***telefono***.

- **Variables independientes: Capital humano.**
  - **Primaria Completa.**

Con la variable ***p301a*** crearemos a la variable ***primaria***, cuyos valores registran información de los jefes de hogar sobre su condición de tener la primaria completa, tendrá dos valores: 0 cuando ***p310a*** sea igual a 1, 2, 3, 6, 7, 8, 9, 10, 11, 12 o esté vacío y 1 cuando ***p301a*** sea igual a 4 o 5. Después le daremos una etiqueta a la variable.

```
. recode p301a (1/3 6/.=0 "Otro") (4/5=1 "Máximo Primaria Completa"), gen(primaria)
(37462 differences between p301a and primaria)

.
. label variable primaria "Primaria completa"
```

Figura 4.54. Variable independiente ***primaria***.

  - **Secundaria Completa.**

Otra vez utilizaremos la variable ***p301a*** para crear la variable ***secundaria*** y sus valores serán iguales a 0 cuando ***p301a*** sea igual a 1, 2, 3, 4, 5, 8, 10, 11, 12 o esté vacío y 1 cuando ***p301a*** sea igual a 6, 7, o 9. Posteriormente, daremos su respectiva etiqueta.

```
. recode p301a (1/5 8 10/.=0 "Otro") (6 7 9=1 "Máximo Secundaria Completa"), gen(secundaria)
(37462 differences between p301a and secundaria)

.
. label variable secundaria "Secundaria completa"
```

Figura 4.55. Variable independiente ***secundaria***.

  - **Superior Completo.**

Una vez más utilizaremos la variable ***p301a*** para crear la variable ***superior*** y tendrá los siguientes valores: 0 cuando ***p301a*** sea igual a 1, 2, 3, 4, 5, 6, 7, 9, 12 o esté vacío y 1 cuando ***p301a*** sea igual a 8, 10 u 11. Les daremos una etiqueta y eliminaremos la variable ***p301a***.

```
. recode p301a (1/7 9 . 12=0 "Otro") (8 10 11=1 "Máximo Superior Completa"), gen(superior)
(37462 differences between p301a and superior)

.
. label variable superior "Educacion superior completa"

.
. drop p301a
```

Figura 4.56. Variable independiente ***superior***.

- **Variables independientes: Capital físico.**
  - **Título de propiedad.**

Con la variable ***p106a*** generamos la variable ***propiedad*** y será una variable Dummy con valor igual a 1 cuando ***p106a*** sea igual a 1 y 0 cuando ***p106a*** sea igual a 2 o esté vacio, le damos una etiqueta y eliminamos la variable ***p106a***.

```
. recode p106a (1=1 "Con titulo de propiedad") (2/.=0 "Sin titulo de propiedad"), gen(propiedad)
(24064 differences between p106a and propiedad)

.
. label variable propiedad "Titulo de propiedad"

.
. drop p106a
```

Figura 4.57. Variable independiente ***propiedad***.

- **Cocina**.

Crearemos la variable ***cocina*** con la variable ***p1138*** y tendrá valores igual a 0 cuando ***p1138*** sea igual a 0 y 1 cuando ***p1138*** sea igual a 1 o esté vacío, le daremos etiqueta a la variable ***cocina*** y borraremos la variable ***p1138***.

```
. recode p1138 (0=0 "Cocina") (1 .=1 "No cuenta con cocina"), gen(cocina)
(0 differences between p1138 and cocina)

.
. label variable cocina "Cuenta con cocina"

.
. drop p1138
```

Figura 4.58. Variable independiente ***cocina***.

- **Auto propio**.

La variable ***auto*** se generará con la variable ***p612_auto*** y tendrá valores iguales a 1 cuando ***p612_auto*** sea igual a 1 y 0 cuando ***p612_auto*** sea igual a 2 o esté vacío. Le otorgaremos una etiqueta y eliminaremos la variable ***p612_auto***.

```
. recode p612_auto (1=1 "Cuenta con auto propio") (2 .=0 "No cuenta con auto propio"), gen(auto)
(33822 differences between p612_auto and auto)

.
. label variable auto "Cuenta con auto propio"

.
. drop p612_auto
```

Figura 4.59. Variable independiente ***auto***.

- **Camión propio.**

Con la variable ***p612_camion*** se creará la variable ***camion*** y tendrá valores iguales a 1 cuando ***p612_camion*** sea igual a 1 y 0 cuando ***p612_camion*** sea igual a 2 o esté vacío. Le otorgamos una etiqueta y eliminaremos la variable ***p612_camion***.

```
. recode p612_camion (1=1 "Cuenta con camion") (2 .=0 "No cuenta con camion"), gen(camion)
(37240 differences between p612_camion and camion)

.
. label variable camion "Cuenta con camion"

.
. drop p612_camion
```

Figura 4.60. Variable independiente ***camion***.

- **Número de habitaciones.**

Para crear la variable ***habitaciones*** se hará uso de la variable ***p104***, no obstante, a diferencia de las anteriores variables, la variable ***habitaciones*** es una variable cuantitativa, por ello no es necesario recodificar sus valores. Las únicas transformaciones que haremos será el nombre de la variable ***p104*** por ***habitaciones*** y reemplazamos con 0 si ***p104*** tiene algún dato faltante.

```
. rename p104 habitaciones

.
. replace habitaciones=0 if habitaciones==.
(409 real changes made)

.
. label variable habitaciones "Numero de habitaciones"
```

Figura 4.61. Variable independiente ***habitaciones***.

- **Variables independientes: Capital social**.
  - **Asociaciones**.

Simplemente le pondremos la etiqueta “Pertenencia a una asociación productiva” a la variable ***asociacion***.

```
. label variable asociacion "Pertenencia a una asociacion productiva"
```

Figura 4.62. Variable independiente ***habitaciones***.

- **Variables independientes: Características del hogar o jefe de hogar**
  - **Total de miembros del hogar.**

Al igual que la variable ***asociacion***, simplemente renombraremos la variable ***totmieho*** por ***personas*** y le agregaremos la etiqueta **“Numero de personas en el hogar”.** La variable ***totmieho*** es encontrada en el archivo descargado del módulo “Sumaria” y es muy importante debido a que, en la mayoría de investigaciones es usada para calcular ingresos per cápita, gastos per cápita, entre otros.

```
. rename totmieho personas

.
. label variable personas "Numero de personas en el hogar"
```

Figura 4.63. Variable independiente ***personas***.

- **Edad del jefe de hogar y edad al cuadrado del jefe de hogar**.

Ambas variables se construirán con la variable ***p208***. Simplemente renombramos a ***p208*** por ***edad*** y después generamos el cuadrado de la variable ***edad*** con el comando **gen** y les daremos etiquetas a cada variable.

```
. rename p208 edad

.
. gen edad2=edad^2

.
. label variable edad "Edad del jefe de hogar"

.
. label variable edad2 "Edad del jefe de hogar al cuadrado"
```

Figura 4.64. Variables independientes ***edad*** y ***cuadrado de la edad***.

- **Lengua nativa**.

La variable ***lengua_nativa*** contiene información sobre la lengua nativa del jefe de hogar, se creará con la variable ***p300a*** y tendrá valores iguales a 1 cuando ***p300a*** sea 1, 2 o 3 y tendrá valores iguales a 0 cuando ***p300a*** sea 4, 5, 6, 7, 8 o si es un dato faltante. Le colocaremos una etiqueta a la variable ***lengua_nativa*** y eliminaremos la variable ***p300a***.

```
. recode p300a (1/3=1 "Lengua nativa") (4/.=0 "Otros"), gen(lengua_nativa)
(28096 differences between p300a and lengua_nativa)

.
. label variable lengua_nativa "El jefe de hogar habla lengua nativa como lengua materna"
```

Figura 4.65. Variable independiente ***lengua_nativa***.

- **Urbanismo**.

La variable ***rural*** contendrá información sobre la procedencia rural de la vivienda. Se usara la variable ***estrato*** para construir la variable ***rural*** y tendrá valores igual a 1 si

estrato es igual a 6,7 u 8 y los demás valores serán iguales a 0 si ***estrato*** es igual a 1, 2, 3, 4 o 5.

```
. recode estrato (6/8=1 "Rural") (1/5=0 "Urbano"), gen(rural)
(37462 differences between estrato and rural)

.
. label variable rural "Area de procedencia rural"

.
. drop estrato
```

Figura 4.66. Variable independiente ***rural***.

- **Variables independientes: transferencias**.
  - **Transferencias**.

La variable ***transferencias*** ya está creada, por ello solamente le agregaremos la etiqueta **"transferencias totales anuales al hogar".**

```
. label variable transferencias "Transferencias totales anuales al hogar"
```

Figura 4.67. Variable independiente ***transferencias***.

- **Variables independientes: Sexo del jefe de hogar**.
  - **Sexo del jefe de hogar.**

La variable ***sexo*** ya está creada en la variable ***p207*** entonces renombramos la variable ***p207*** por ***sexo*** y le agregaremos la etiqueta **"Sexo del jefe de hogar".**

```
. rename p207 sexo

.
. label variable sexo "Sexo del jefe de hogar"
```

Figura 4.68. Variable independiente ***sexo***.

- **Variables independientes: Dominio geográfico**.
  - **Lima**

Para crear la variable ***lima*** tomaremos la variable ***dominio*** y según sus valores, calcularemos los valores de la variable ***lima***. Si ***dominio*** tiene valores iguales a 1, 2 o 3 entonces la variable ***lima*** será igual a 1, si ***dominio*** tiene valores iguales a 4, 5 o 6 entonces la variable ***lima*** será igual a 2, si ***dominio*** tiene valores igual a 7 entonces la variable ***lima*** tendrá valores iguales a 3 y si la variable ***dominio*** es igual a 8 entonces la variable ***lima*** será igual a 0.

```
. recode dominio (1/3=1 "Costa") (4/6=2 "Sierra") (7=3 "Selva") (8=0 "Lima Metropolitana"), gen(lima)
(32681 differences between dominio and lima)

.
. label variable lima "Dominio geográfico"

.
. drop dominio
```

Figura 4.69. Variable independiente ***lima***.

Estas han sido todas las variables que utilizaremos, entonces eliminaremos las variables ***vivienda***, ***hogar***, ***linpe***, ***línea***, ***p104*** y ***t***. Guardaremos la base de datos con el nombre "data_final2018t"

```
. drop vivienda hogar linpe linea p104 t

. save data_final2018t.dta,replace
```

Figura 4.70. Guardando el archivo "data_final2018t".

En la siguiente tabla se muestra un breve resumen de la información sobre las variables creadas.

Tipo de variable en el modelo.	Vector de cada activo o infraestructura.	Elemento de cada vector según el activo o infraestructura.	Variable.	Valores y/o etiquetas.	Tipo de variable.
Variable dependiente.	Nivel de pobreza.	Pobreza.	***niv_pobreza***	0. No pobre. 1. Pobre.	Variable dicotómica.
Variables explicativas.	Acceso a servicios básicos.	Agua.	***agua***	0. Otros. 1. Red Pública.	Variable dicotómica.
		Desagüe.	***desague***	0. Otros. 1. Red Pública.	Variable dicotómica.
		Electricidad.	***electricidad***	0. Otros. 1. Electricidad.	Variable dicotómica.
		Teléfono.	***teléfono***	0. Ninguna. 1. Telefonía fija o móvil.	Variable dicotómica.
	Capital humano.	Primaria completa.	***primaria***	0. Otro. 1. Máximo Primaria Completa.	Variable dicotómica.
		Secundaria completa.	***secundaria***	0. Otro. 1. Máximo Secundaria Completa.	Variable dicotómica.
		Superior completo.	***superior***	0. Otro. 1. Máximo Superior Completa.	Variable dicotómica.

	Título de propiedad.	***propiedad***	0. Sin título de propiedad. 1. Con título de propiedad.	Variable dicotómica.
	Cocina.	***cocina***	0. Cocina. 1. No cuenta con cocina.	Variable dicotómica.
Capital físico.	Auto propio.	***auto***	0. No cuenta con auto propio. 1. Cuenta con auto propio.	Variable dicotómica.
	Camión.	***camion***	0. No cuenta con camión. 1. Cuenta con camión.	Variable dicotómica.
	Número de habitaciones.	***habitaciones***		Variable discreta.
Capital social.	Asociaciones.	***asociacion***	0. No pertenece. 1. Pertenece.	Variable dicotómica.
	Total de miembros del hogar.	***personas***		Variable discreta.
Características del hogar o del jefe de hogar.	Edad del jefe de hogar.	***edad***		Variable discreta.
	Edad del jefe de hogar al cuadrado.	***edad2***		Variable discreta.
	Lengua nativa.	***lengua_nativa***	0. Otros. 1. Lengua nativa.	Variable dicotómica.
	Urbanismo.	***rural***	0. Urbano. 1. Rural.	Variable dicotómica.
Transferencias.	Transferencias por jubilación.	***transferencias_jub***	0. El hogar no recibió transferencias de ningún tipo. 1. El hogar recibió transferencias de distinto tipo	Variable dicotómica.
Sexo.	Sexo del jefe de hogar.	***sexo***	1. Varón 2. Mujer	Variable dicotómica.
Dominio.	Dominio geográfico.	***lima***	0. Lima Metropolitana. 1. Costa. 2. Sierra. 3. Selva	Variable multinomial.

Tabla 4.1. Información sobre las variables que se usarán en el modelo especificado.

4.5.5. Estimación de los coeficientes de regresión.

Previamente a la estimación de los modelos de Logit, veamos algunos estadísticos descriptivos representados en tablas y gráficos para tener una mayor visión sobre el tema investigado. A continuación, se presenta una tabla sobre la tasa de pobreza en el año 2018 a nivel nacional.

```
. tab niv_pobreza [aw=factor07*personas]

  Nivel de
   Pobreza |      Freq.     Percent        Cum.
-----------+-----------------------------------
  No pobre | 29,810.602       79.58       79.58
     Pobre | 7,651.3983       20.42      100.00
-----------+-----------------------------------
     Total |     37,462      100.00
```

Figura 4.71. Tasa de pobreza en el año 2018 a nivel nacional.

En la figura 4.71. Se puede ver una tabla que muestra la tasa de la pobreza a nivel nacional en el año 2018. Esta tabla ha sido construida con el comando **tab** y utilizando el factor de expansión **aw** que representa la **ponderación analítica**. Básicamente, se le ha ordenado a STATA que multiplique a las variables ***factor07*** y ***personas*** que representan el factor de expansión y el total de miembros de los hogares, respectivamente. De esta forma podemos obtener una tasa que se ajuste mejor a la población. Podemos ver que la tasa de la pobreza es del 20.42% en el año 2018, esta es una cifra alentadora si tomamos en cuenta que, según (Aparicio, Jaramillo, & San Román , 2011) La tasa de la pobreza a nivel nacional en el año 2010 ha sido 27.64%. Se deduce que la pobreza en casi la última década se ha reducido casi en 7 puntos porcentuales.

A continuación, veremos gráficas sobre el porcentaje del total de hogares que tienen acceso a los servicios de agua, desagüe, electricidad y teléfono, tomando en cuenta los quintiles de gastos e ingresos en el año 2018 a nivel nacional.

Veamos la gráfica para el servicio del agua. Primero introducimos el comando **preserve.** En segundo lugar, con el comando **xtile**, la ponderación **pw**, la variable ***ingreso_total*** y la opción **n(5)** crearemos la variable ***q_ingreso***, la cual representa a los 5 quintiles. En economía, el término quintil se refiere a una variable que distribuye a otra variable en 5 grupos, otros términos familiarizados a la distribución son percentiles, deciles, etc.; es muy usado para ordenar a regiones según el quintil donde se encuentren, de tal forma que los quintiles superiores tienen valores mayores que los quintiles inferiores.

```
. xtile q_ingreso=ingreso_total [pw=factor07], n(5)
```

Figura 4.72. Creación de la variable ***q_ingreso*** que muestra los quintiles según la distribución del ingreso de los hogares.

Para que quede claro veamos una tabla que muestra los valores de la variable ***q_ingreso***.

```
. tab q_ingreso

5 quantiles |
         of |
ingreso_tot |
        al |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |      9,430       25.17       25.17
         2 |      8,269       22.07       47.25
         3 |      7,315       19.53       66.77
         4 |      6,554       17.50       84.27
         5 |      5,894       15.73      100.00
-----------+-----------------------------------
     Total |     37,462      100.00
```

Figura 4.73. Tabla de la variable ***q_ingreso***.

En la figura 4.73. Se ve que hemos creado 5 grupos según el ingreso de las variables, se puede interpretar de la siguiente manera: en el quintil Q5 se encuentra el 15.73% de la muestra, entonces los hogares que se encuentren en este quintil obtendrían mayores ingresos que los quintiles que están por debajo. Por otro lado, el quintil Q1 abarca el 25.17% de la muestra y los hogares que se encuentren en este quintil sería el grupo más pobre de todos. Después de crear la variable ***q_ingreso***, generamos la variable ***agua_1*** siendo igual a la variable ***agua*** multiplicada por 100.

```
. gen agua_1=100*agua
```

Figura 4.74. Creación de la variable ***agua_1***.

El motivo de la creación de esta variable es obtener porcentajes. El siguiente paso es "colapsar" la base de datos, es decir, con el comando **collapse** mantendremos una variable que represente un estadístico descriptivo de otra variable. En este caso calcularemos el porcentaje del total de hogares que tengan acceso al servicio básico de agua, según la distribución de los quintiles e ingresos de las familias tomando en cuenta la ponderación **pw**. Para ello, la siguiente figura muestra la instrucción.

```
. collapse (mean) ingreso=agua_1 [pw=factor07], by(q_ingreso)
```

Figura 4.75. Colapso de la base de datos.

En la figura 4.75. Se puede apreciar que entre paréntesis hemos colocado el estadístico descriptivo deseado, para el ejemplo se ha requerido el promedio, por eso es que se ha colocado **(mean)**, si se hubiera requerido la desviación estándar entonces colocaremos **(sd)**, la mediana **(median)**, los valores máximos **(max)** y así entre otros. El componente de la instrucción que se coloca en el paréntesis, indica a STATA que genere una base de datos con el estadístico descriptivo deseado tomando a la variable ***agua_1***, el

cual debe ser contenido en la variable ***ingreso*** distribuido para cada quintil de ingresos, ya que la opción **by()** lo está indicando. Generando así los porcentajes totales de los hogares con acceso a agua. Además, se ha ordenado que tome en cuenta a la ponderación **pw** para la creación de dicha base de datos.

Con el comando **tabstat** y la opción **by()** se crea una tabla que muestran los porcentajes calculados distribuidos según los quintiles.

```
. tabstat ingreso, by( q_ingreso)

Summary for variables: ingreso
     by categories of: q_ingreso (5 quantiles of ingreso_total )

q_ingreso |      mean
----------+----------
        1 |  76.32054
        2 |  84.30296
        3 |  88.69212
        4 |  92.28581
        5 |   96.3444
----------+----------
    Total |  87.58916
------------------------
```

Figura 4.76. Tabla sobre la variable **ingreso** (1).

Para interpretar el Q5 podríamos decir: el 96.34% del total de hogares en el Q5 tienen acceso a agua potable, mientras que el 76% del total de hogares en el Q1 tienen acceso a agua potable.

```
. rename q_ingreso a

.
. label variable ingreso "Ingreso"

.
. label define aa 1 "Quintil 1" 2 "Quintil 2" 3 "Quintil 3" 4 "Quintil 4" 5 "Quintil 5"

.
. label values a aa
```

Figura 4.77. Agregando etiquetas y cambiando nombres.

A la variable ***q_ingreso***, que muestra información sobre los quintiles, se le ha cambiado su nombre por ***a***, posteriormente a la variable ***ingreso*** se le ha puesto la etiqueta **"Ingreso"** y se le añadido etiquetas a cada valor de la variable ***a*** con la lista de etiquetas ***aa***. Si realizamos la misma tabla podremos ver sus etiquetas.

```
. tabstat ingreso, by( a)

Summary for variables: ingreso
     by categories of: a (5 quantiles of ingreso_total )

        a |      mean
----------+----------
Quintil 1 |  76.32054
Quintil 2 |  84.30296
Quintil 3 |  88.69212
Quintil 4 |  92.28581
Quintil 5 |   96.3444
----------+----------
    Total |  87.58916
------------------------
```

Figura 4.78. Tabla sobre la variable ***ingreso*** (2).

La única diferencia entre las figuras 4.76. Y 4.78. Es que en la segunda figura se le ha agregado etiquetas a cada valor de la variable ***a*** que muestra los quintiles. La interpretación se mantiene. Ahora guardamos la base de datos con el nombre "b_1" y restauramos la base de datos original con los comandos **save** y **restore** respectivamente.

```
. save b_1,replace
. restore,preserve
```

Figura 4.79. Guardada la base de datos colapsada.

El segundo para para observar dicha gráfica es replicar el proceso anterior para crear una variable que sea el porcentaje del total de familias que tienen acceso a agua potable según los quintiles del gasto. Los comandos que se utilizaran son.

```
. xtile q_gasto=gasto_total [pw=factor07], n(5)

.
. gen agua_2=100*agua

.
. collapse (mean) gasto=agua_2 [pw=factor07], by(q_gasto)

.
. rename q_gasto a

.
. label variable gasto "Gasto"

.
. label define aa 1 "Quintil 1" 2 "Quintil 2" 3 "Quintil 3" 4 "Quintil 4" 5 "Quintil 5"

.
. label values a aa
```

Figura 4.80. Creando la variable ***gasto***.

Para ver los porcentajes calculados lo haremos utilizando el comando **tabsat**.

```
. tabstat gasto,by(a)

Summary for variables: gasto
     by categories of: a (5 quantiles of gasto_total )

        a |      mean
----------+----------
Quintil 1 |  75.55467
Quintil 2 |  84.54222
Quintil 3 |  88.97844
Quintil 4 |  92.08242
Quintil 5 |  96.79263
----------+----------
    Total |  87.59008
------------------------
```

Figura 4.81. Tabla de la variable ***gasto*** según la variable ***a***.

Guardamos la base de datos con el nombre "b_2" con el comando **save**.

```
. save b_2,replace
```

Figura 4.82. Guardando la base de datos "b_2".

Posteriormente, la uniremos con la base de datos "b_1" con el comando **merge** y podremos guardar la base con el nombre "agua". Esta base de datos será utilizada para realizar la gráfica sobre el porcentaje del total de hogares que tienen acceso a agua potable según su ubicación en los quintiles de ingreso y gasto.

```
. merge 1:1 a using b_1
(label aa already defined)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                 5  (_merge==3)
    -----------------------------------------

.
. save agua,replace
```

Figura 4.83. Guardando la base de datos "agua".

Para generar un gráfico en donde se pueda apreciar los quintiles del ingreso y el gasto haremos uso del comando **graph**, su componente **bar**, las variables ***gasto*** e ***ingreso***, ambas ya tienen los porcentajes de las familias según los quintiles de ingreso y gasto que tienen acceso al servicio de agua potable. Algunas opciones que se utilizarán para

```
. graph bar gasto ingreso, over(a) saving(agua.gph, replace) blabel(bar, format(%4.1f)) title("Agua Potab
> le")legend(lab(1 "Gasto") lab(2 "Ingreso"))
```

complementar a la gráfica de barras son: **over()** la cual muestra una categoría según la variable que seleccionemos en el paréntesis, **saving** guarda el gráfico otorgándole un nombre y un formato de imagen que para este ejemplo será "Agua Potable" el nombre y .gph el formato usado para guardar la imagen del gráfico, **blabel()** agrega un formato a las barras del gráfico, **title()** añade un título a la gráfica y **legend()** muestra una leyenda según las variables utilizadas ***ingreso*** y ***gasto***. Veamos la sintaxis del comando.

Figura 4.84. Generando la gráfica de barras "agua".

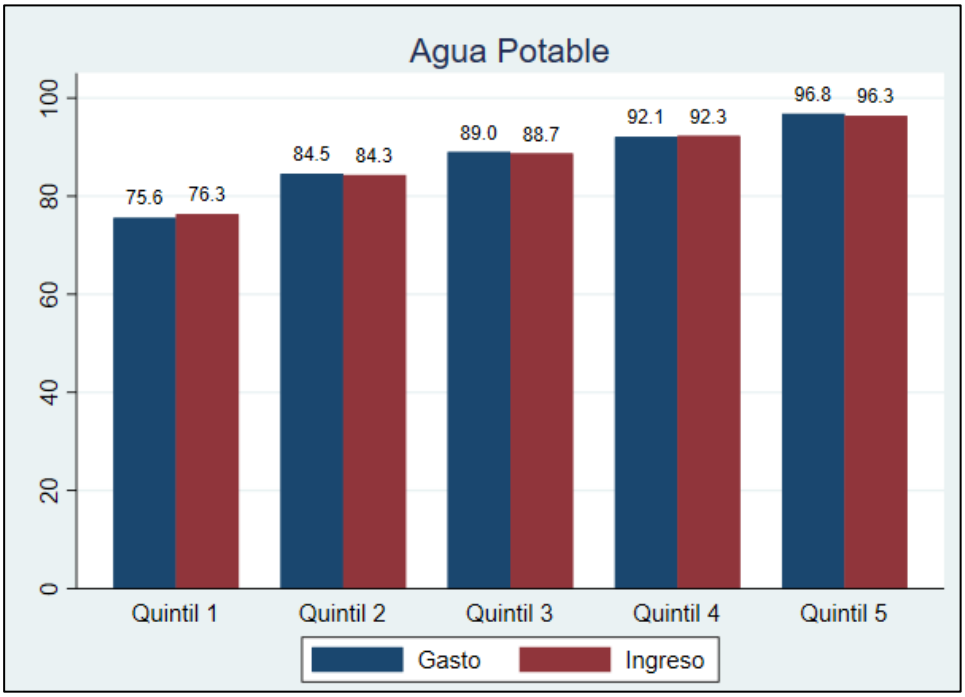


Figura 4.85. Gráfica de barras "Agua Potable".

Si repetimos el proceso para los servicios básicos de desagüe, electricidad y teléfono, podemos obtener los siguientes gráficos sobre el porcentaje del total de hogares que tienen acceso a los servicios básicos de desagüe, electricidad y teléfono según los quintiles de ingreso y gasto.

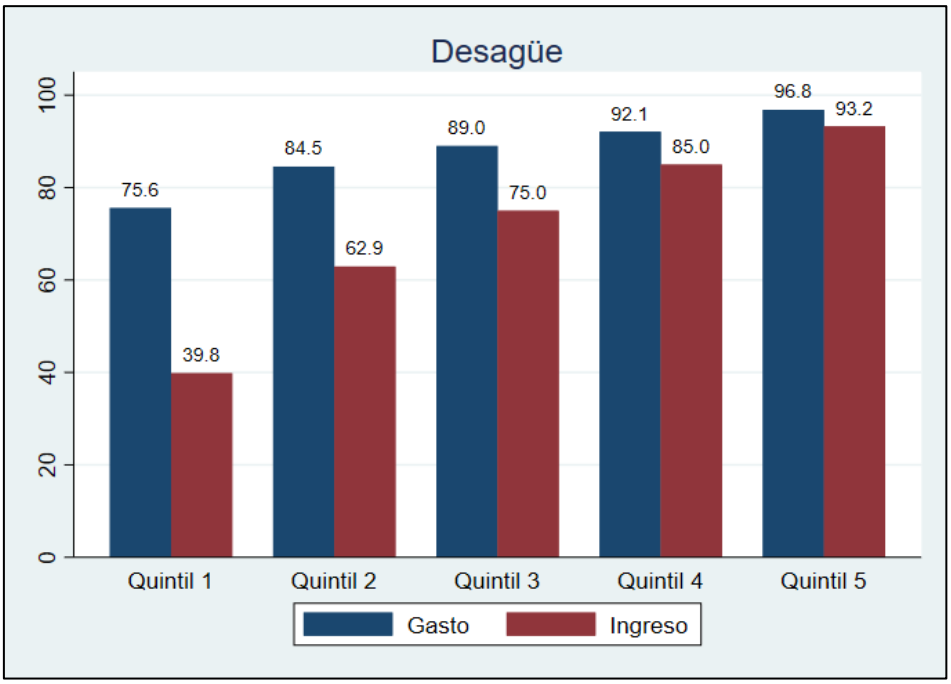


Figura 4.86. Gráfica de barras "Desagüe".

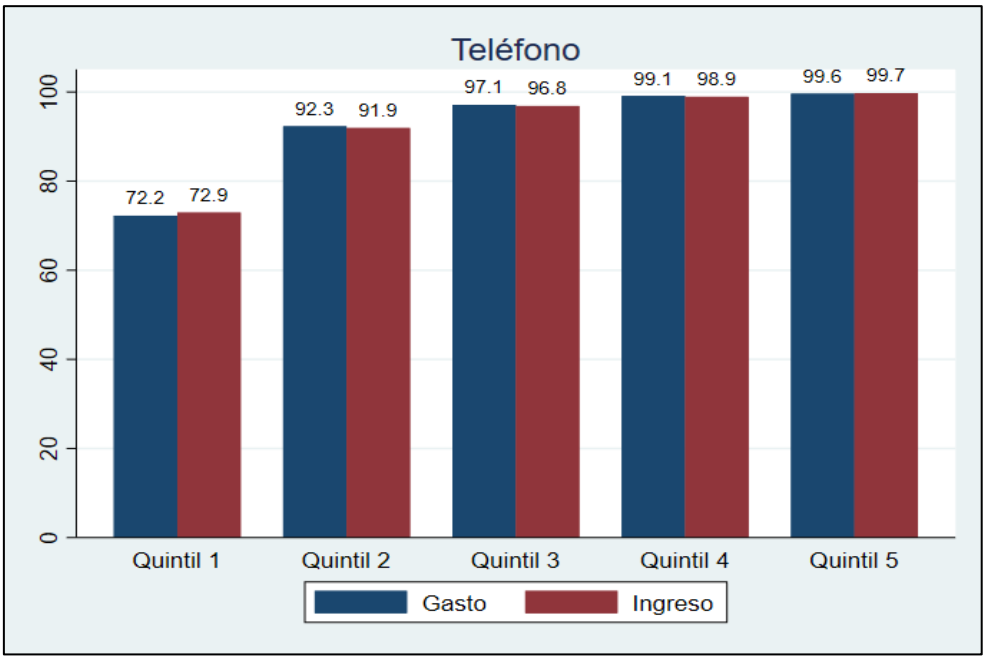


Figura 4.87. Gráfica de barras "Teléfono".

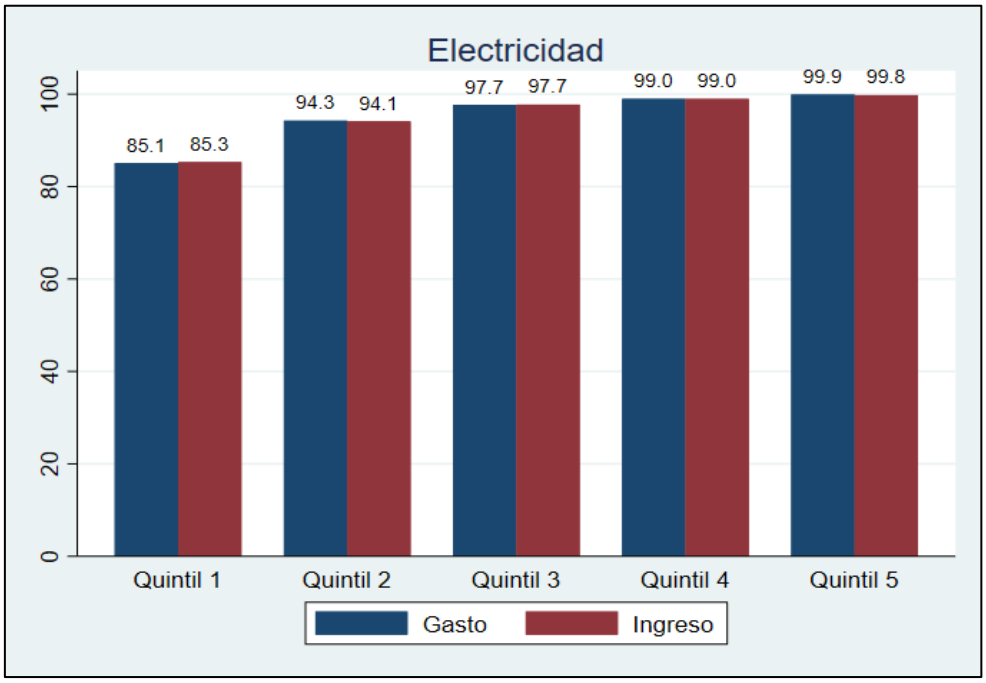


Figura 4.88. Gráfica de barras "Electricidad".

La disminución de la pobreza en 7 puntos porcentuales se traduce en el aumento del total de hogares con accesibilidad a los servicios básicos de agua potable, desagüe, electricidad y telefonía. Estas mejoras son más apreciables en los Q1 tanto de ingresos y gastos de las familias.

(Aparicio, Jaramillo, & San Román , 2011) Indican que el 47.7% de las familias tuvieron acceso al servicio básico de agua potable según el quintil Q1 del ingreso y 41.0% del quintil Q1 del gasto en el año 2010. Mientras tanto, según la figura 4.85., para el año 2018 el 76.3% del total de hogares tienen acceso a agua en el quintil Q1 del ingreso y el 75.6% del total de hogares tienen acceso a agua potable en el quintil Q1 del gasto. Podemos ver que, en el quintil Q1 el porcentaje ahora es casi el doble de lo que era hace casi una década, no obstante, la brecha aún es palpable en los quintiles Q1 y quintiles Q5 pese a que se ha logrado reducir considerablemente.

A continuación, veamos una gráfica de barras que muestre el porcentaje del total de hogares que se encuentran en situación de pobreza distribuido según el área de residencia, para lograrlo debemos generar la variable ***pobreza_temp*** que sea el producto

```
. gen pobreza_temp=100*niv_pobreza

.
. gr bar pobreza_temp, over(rural) saving(residencia.gph,replace) blabel(bar, format(%4.1f)) title("Pobreza en los Hogares del Perú se
> gún lugar de residencia")legend(lab(1 "Urbano") lab(2 "Rural")) asyvars bargap(10)
```

de la variable ***niv_pobreza*** por 100 para lograr calcular los porcentajes. La gráfica de barras la realizaremos con el comando **gr** y el componente **bar**.

En este caso, hemos utilizado casi todas las opciones que fueron utilizadas en las

Figura 4.89. Generando la gráfica de barras de la pobreza según el área de residencia.

grupo de la variable dentro de la opción, **over()** como la variable que va en el eje Y. Es necesario colocar a la opción **over()** si queremos trabajar con **asyvars**.

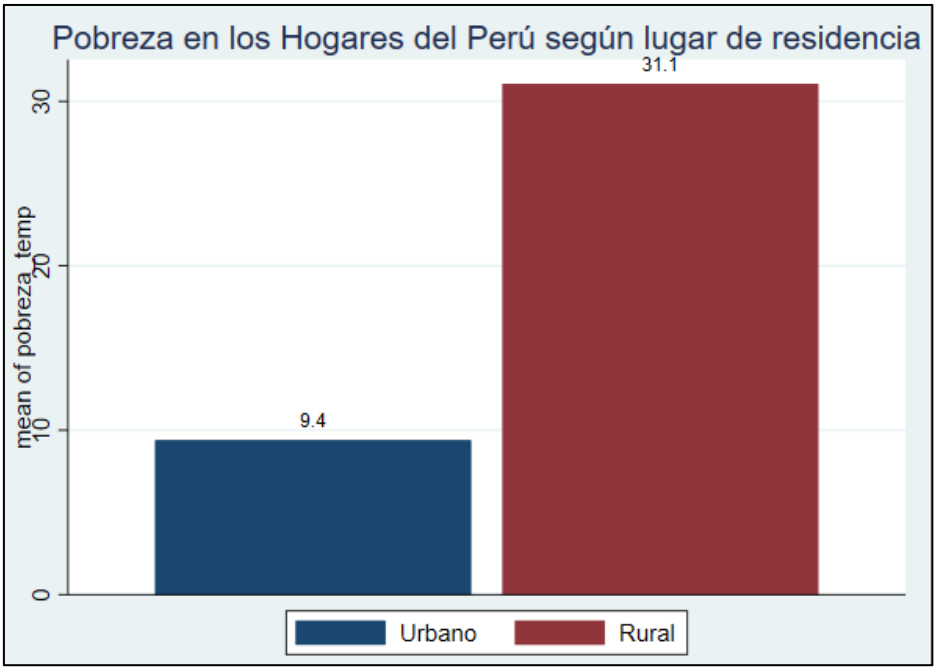


Figura 4.90. Gráfica de barras de la pobreza según el área de residencia.

en el porcentaje de pobres en los hogares según el área de residencia del hogar. En 2018, los hogares pobres ubicados en las áreas urbanas representaron el 9.4% y en 2010 la cifra fue de 19.1%. Por otro lado, los hogares pobres ubicados en las zonas rurales fueron el 31.1% del total de hogares en 2018 y en el año 2010 la cifra fue de 54.2%. Pese a las mejoras, aún se puede apreciar que las zonas rurales mantienen más hogares pobres que las zonas urbanas.

La reducción de la pobreza en las áreas de residencia ha sido causada por un efecto de la accesibilidad de los servicios básicos de los hogares en cada área, la siguiente tabla muestra el porcentaje de hogares del total con acceso a los servicios básicos según el área de residencia.

Empezamos con el comando **preserve** ya que necesitaremos restaurar la base de datos después de las modificaciones que le haremos, como segundo paso generamos una variable llamada ***servicio*** que contenga solamente valores igual a 1, luego renombramos a la variable ***agua*** por ***serv*** y mantendremos las variables ***rural***, ***servicio*** y ***serv*** con el comando **keep.** Finalmente, guardamos la base de datos con el nombre "c_1t" y restauramos la base de datos original con **restore** y su opción **preserve**.

```
. preserve
. gen servicio=1

.
. rename agua serv

.
. keep rural servicio serv

.
. save c_1t, replace
```

Figura 4.91. Guardando la base de datos "c_1t".

El mismo proceso haremos con el resto de servicios básicos para el desagüe, electricidad y teléfono.

```
. restore, preserve

. gen servicio=2

.
. rename desague serv

.
. keep rural servicio serv

.
. save c_2t, replace
```

Figura 4.92. Guardando la base de datos "c_2t".

```
. restore, preserve

.
. gen servicio=3

.
. rename electricidad serv

.
. keep rural servicio serv

.
. save c_3, replace
```

Figura 4.93. Guardando la base de datos “c_3”.

```
. restore, preserve

.
. gen servicio=4

.
. rename telefono serv

.
. keep rural servicio serv

.
. save c_4t, replace
```

Figura 4.94. Guardando la base de datos “c_4t”.

Con el comando **append** podremos agregar los valores de las bases de datos “c_1t”, “c_2t” y “c_3” lo cual agregará los valores de cada base de datos que contiene información sobre agua potable, desagüe y electricidad respectivamente a las variables que ya están en la base de datos “c_4t”.

```
. append using c_1t c_2t c_3
(label agua already defined)
(label rural already defined)
(label desague already defined)
(label rural already defined)
(label a already defined)
(label rural already defined)
```

Figura 4.95. Guardando la base de datos “c_4t”.

Después de agregar las bases de datos, ahora le agregaremos etiquetas a la variable ***servicio*** con el comando **label define** y **label values**, posteriormente generamos la variable ***serv_1*** el cual es el producto de la variable ***serv*** por 100 para calcular los porcentajes. Con el comando **table** y las opciones **c()** y **format()** crearemos la tabla usando además las variables ***servicio*** y ***rural***.

```
. label define servicio 1 "Agua Potable" 2 "Desague" 3 "Electricidad" 4 "Telefono"

.
. label values servicio servicio

.
. gen serv_1=serv*100

. table servicio rural, c(mean serv_1) format(%4.1f)
```

	Area de procedencia rural	
servicio	Urbano	Rural
Agua Potable	93.4	67.6
Desague	86.2	21.7
Electricidad	98.9	80.8
Telefono	95.7	80.1

Figura 4.96. Generando la tabla sobre el porcentaje del total de hogares con acceso a los servicios básicos según el área de residencia.

En la figura 4.96., el comando **label define** crea una lista de etiquetas bajo el nombre ***servicio***, guarda la lista definida en la memoria, y el comando **label values** hace uso de tal lista de etiquetas para agregarle las etiquetas correspondientes a los valores de la variable ***servicio***.

La tabla, que se puede ver en la figura, representa los porcentajes de los hogares con accesibilidad a los servicios básicos según el área de residencia. Siendo el servicio básico con una mayor cobertura es el servicio teléfono en el área urbana y el servicio básico de electricidad cuenta con mayor cobertura en el área rural.

Según (Aparicio, Jaramillo, & San Román , 2011) En 2010 el 52.5% de los hogares en el área rural han tenido acceso a teléfono y en 2018 la cifra aumento en 80.1%. Mientras en las zonas urbanas ha alcanzado en el año 2018 el 95.7%, lo cual representa un aumento de 4 puntos porcentuales con respecto al año 2010. No obstante, el servicio con una mejora en la cobertura menor ha sido el servicio de desagüe; en 2010, el 10.4% de los hogares en las zonas rurales han tenido acceso al servicio de desagüe mientras que

en 2018 apenas ha logrado aumentar 11 puntos porcentuales en la misma zona rural; del mismo modo, en las zonas urbanas ha sido 83.0% en el año 2010 y para el año 2018 ha logrado aumentar 3 puntos porcentuales. A continuación, replicaremos el mismo proceso, pero ahora tomaremos en cuenta el sexo de los jefes de hogar para observar el porcentaje del total de hogares.

Para generar una gráfica de barras que represente el porcentaje del total de hogares pobres según el sexo del jefe de hogar, utilizaremos la variable generada ***pobreza_temp*** y ordenamos la misma sintaxis del comando que se ve en la figura 4.89., pero esta vez utilizaremos la variable ***sexo*** en lugar de la variable ***rural*** en la opción **over()** y acorde a los valores de la variable ***sexo*** configuramos las opciones **title()** y **legend()** para que muestren el título y las etiquetas correspondientes, respectivamente. Por último, guardamos al grafico generado con el nombre "sexo" con la opción **saving()**

```
. gr bar pobreza_temp, over(sexo) saving(sexo.gph,replace) blabel(bar, format(%4.1f)) title("Pobreza en l
> os Hogares del Perú según sexo del jefe de hogar") legend(lab(1 "Hombre") lab(2 "Mujer")) asyvars barga
> p(10)
```

Figura 4.97. Generando la gráfica de barras de la pobreza según el área de residencia.

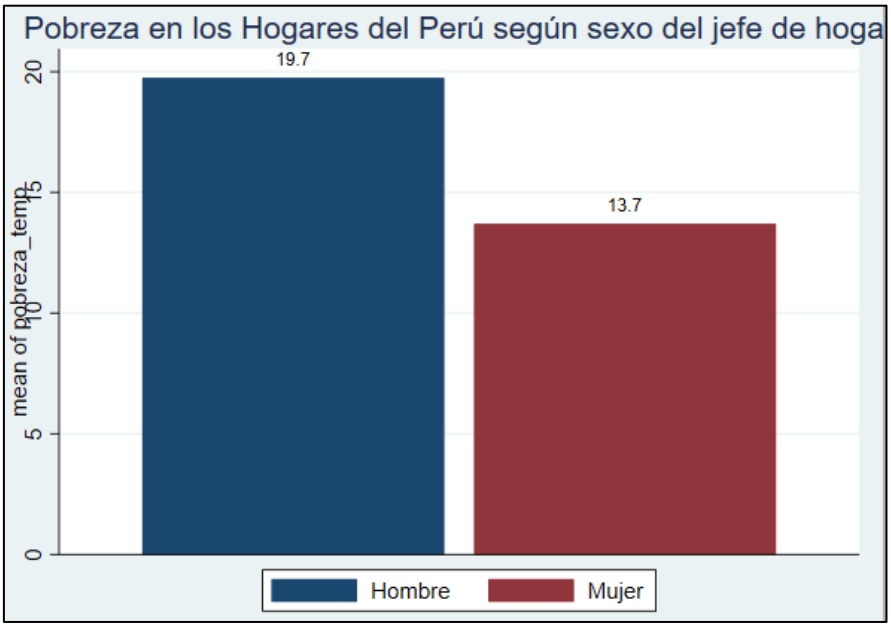


Figura 4.98. Gráfica de barras de la pobreza según el sexo del jefe de hogar.

La disminución de la pobreza entre los años 2018 y 2010 se refleja en el porcentaje de hogares pobres. En los hogares pobres con jefe de hogar con sexo femenino han logrado la reducción de 11.6 puntos porcentuales con respecto al año 2010, mientras el porcentaje de los hogares pobres con jefe de hogar con sexo masculino se ha reducido 13.2 puntos porcentuales.

Para replicar la tabla que se ve en la figura 4.96. Se utiliza el mismo procedimiento, la diferencia es que, en vez de mantener la variable ***rural***, mantendremos la variable ***sexo*** y los demás pasos se realizarán sin efectuar cambios. Para distinguir de las bases de datos usadas para la tabla en la figura 4.90. Guardaremos cada base de datos utilizando “d_1t”, “d_2t”, “d_3” y “d_4t” respectivamente para cada servicio básico.

```
. preserve

.
. gen servicio=1

.
. rename agua serv

.
. keep sexo servicio serv

.
. save d_1t, replace
```

Figura 4.99. Guardando la base de datos “d_1t”.

```
. restore, preserve

.
. gen servicio=2

.
. rename desague serv

.
. keep sexo servicio serv

.
. save d_2t, replace
```

Figura 4.100. Guardando la base de datos “d_2t”.

```
. restore, preserve

.
. gen servicio=3

.
. rename electricidad serv

.
. keep sexo servicio serv

.
. save d_3, replace
```

Figura 4.101. Guardando la base de datos “d_3”.

```
. restore, preserve

.
. gen servicio=4

.
. rename telefono serv

.
. keep sexo servicio serv

.
. save d_4t, replace
```

Figura 4.102. Guardando la base de datos “d_4t”.

```
. append using d_1t d_2t d_3
(label agua already defined)
(label p207 already defined)
(label p207 already defined)
(label desague already defined)
(label p207 already defined)
(label a already defined)

.
. label define servicio 1 "Agua Potable" 2 "Desague" 3 "Electricidad" 4 "Telefono"

.
. label values servicio servicio

.
. label define sexo 1 "Hombre" 2 "Mujer"

.
. label values sexo sexo
```

Figura 4.103. Agregando las bases de datos “d_1t”, “d_2t” y “d_3” a la base de datos “d_4t”.

```
. table servicio sexo, c(mean serv_1) format(%4.1f)

                 |  Sexo del jefe
                 |     de hogar
        servicio |  Hombre    Mujer
-----------------+----------------------
   Agua Potable  |    82.1     85.8
        Desague  |    58.1     66.6
   Electricidad  |    90.9     93.7
       Telefono  |    90.6     86.5
```

Figura 4.104. Tabla de porcentajes de los hogares con acceso a los servicios básicos según el sexo del jefe de hogar.

Los efectos en los datos porcentuales que se ven en la gráfica en la figura 4.98. Se pueden visualizar en la tabla de la figura 4.104.

En todos los servicios básicos a excepción del teléfono, el hogar es más propenso a disfrutar del acceso del servicio básico si el jefe de hogar tiene sexo femenino. Después

de estas tablas y gráficos podemos concluir que ha existido una reducción de hogares pobres en el Perú durante los años 2010 y 2018 y esta reducción se aprecia revisando el nivel de acceso que tienen los hogares a los distintos tipos de servicios básicos, tal como mencionó la teoría propuesta por (Aparicio, Jaramillo, & San Román , 2011).

Con estos datos ya podemos hacernos una idea de cómo serán los resultados del modelo especificado, y ya que es posible que los efectos puedan ser distintos según el sexo del jefe de hogar y el área de residencia del hogar después de ejecutar la estimación del modelo especificado (4.5.4.) se realizará el mismo modelo especificado en (4.5.4.), pero tomando, en cuenta cuando el sexo del jefe de hogar es femenino y masculino y cuando el área de residencia del hogar es rural y urbano.

Para agrupar a las distintas variables acorde al tipo de activo al cual pertenecen según la tabla 4.1., se utilizará el comando **global**, el cual es muy útil cuando tenemos muchas variables y queremos agruparlas para evitar que los comandos sean demasiado extensos y engorrosos. Su sintaxis es la siguiente, el primer término que le sigue al comando **global** es el nombre del grupo y colocamos los nombres de las variables que queremos que conformen ese grupo entre comillas. En la siguiente figura se muestra.

```
. global y niv_pobreza

. global x_servicios "agua desague electricidad telefono"

. global x_cap_humano "primaria secundaria superior"

. global x_cap_fisico "propiedad cocina auto camion habitaciones"

. global x_cap_social asociacion

. global x_caracteristicas "personas edad edad2 lengua_nativa rural"

. global x_transferencias transferencias_jub
```

Figura 4.105. Creación de macros globales.

Utilizando la terminología correcta de STATA sobre el comando **global**, el nombre del grupo que crea este comando es **macro global**, en algunos macros globales se puede ver que no se ha omitido colocar a la variable entre comillas, esta sintaxis es válida cuando hacemos macros globales utilizando solamente una variable.

Un método para saber cuáles variables pueden ser seleccionadas para el modelo Logit que se pretende estimar, es utilizando el algoritmo **Stepwise**, que en términos simples se trata de un algoritmo que indica cuales son las variables significativas utilizando un nivel de significancia acorde a un modelo predeterminado. En STATA se puede utilizar con el comando **stepwise** y las opciones **pe()** y el comando que representa al tipo de modelo que queremos estimar, como se trata de un modelo Logit entonces el comando será **logit**, en cuanto a la opción **pe()** este mide la significancia para agregar la variable al modelo. En la siguiente figura se puede ver los resultados obtenidos con el comando **stepwise** y del comando **logit**. Solo mostraremos los resultados del comando **stepwise** ya que los resultados del comando **logit** se analizarán después.

```
. stepwise, pe(0.05):logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $
> x_transferencias
                      begin with empty model
p = 0.0000 <  0.0500  adding  desague
p = 0.0000 <  0.0500  adding  personas
p = 0.0000 <  0.0500  adding  habitaciones
p = 0.0000 <  0.0500  adding  rural
p = 0.0000 <  0.0500  adding  telefono
p = 0.0000 <  0.0500  adding  superior
p = 0.0000 <  0.0500  adding  secundaria
p = 0.0000 <  0.0500  adding  lengua_nativa
p = 0.0000 <  0.0500  adding  auto
p = 0.0000 <  0.0500  adding  primaria
p = 0.0000 <  0.0500  adding  propiedad
p = 0.0000 <  0.0500  adding  transferencias_jub
p = 0.0005 <  0.0500  adding  camion
p = 0.0021 <  0.0500  adding  edad
p = 0.0000 <  0.0500  adding  edad2
p = 0.0005 <  0.0500  adding  cocina
p = 0.0033 <  0.0500  adding  agua
p = 0.0006 <  0.0500  adding  electricidad
```

Figura 4.106. Resultados del comando **stepwise**.

En primer lugar, para utilizar los macros globales se debe anteponer a cada macro el símbolo “$” para que el programa STATA reconozca el uso de los macros globales. Después, podemos ver una lista de variables que conforman a los macros globales cuyos valor-p son menores al nivel de significancia del 5%, por lo que según el comando **stepwise**, deberíamos seleccionar solamente a las variables de la lista para estimar el modelo especificado. En efecto, el programa indica que la variable debe ser agregada con el componente “adding”.

A continuación, veamos los resultados del modelo Logit utilizando el comando **logit**.

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias

Iteration 0:   log likelihood = -17668.195
Iteration 1:   log likelihood = -14277.257
Iteration 2:   log likelihood = -13716.241
Iteration 3:   log likelihood = -13682.857
Iteration 4:   log likelihood = -13682.543
Iteration 5:   log likelihood = -13682.543

Logistic regression                             Number of obs     =     37,462
                                                LR chi2(19)       =    7971.30
                                                Prob > chi2       =     0.0000
Log likelihood = -13682.543                     Pseudo R2         =     0.2256
```

Figura 4.107. Resultados de la estimación del modelo Logit (1).

Como se dijo en la sección que expone el modelo Logit, estos tipos de modelos se resuelven mediante iteraciones de la función de log-verosimilitud y serán tantas como sea necesaria hasta que STATA considere que ya no se puede seguir maximizando la función de log-verosimilitud. Se puede ver que la función de log-verosimilitud (Log likelihood) es -13682.542, la cual ha sido calculado en la quinta iteración (Iteration 5). Aparentemente, la cuarta y quinta iteración es la misma, pero en realidad, la quinta iteración es mayor a la cuarta iteración, no obstante, la diferencia entre ambas es tan ínfima, que a simple vista se podría pensar que se trata de la misma.

Estas iteraciones son importantes para estimar los estimadores utilizando el modelo Logit, según (Escobar M., Fernández M., & Bernardi, 2012) En la primera iteración todos los estimadores a excepción del intercepto son iguales a 0 y según las iteraciones de las funciones log-verosimilitud vayan aumentando los estimadores son más verosímiles. Al lado derecho de las iteraciones se pueden observar algunos estadísticos, de arriba hacia abajo son: el número de observaciones (Number of obs), la razón de verosimilitud (LR chi2(19)), el valor-p de la razón de verosimilitud (Prob>chi2) y el pseudo R^2 ($pseudo\, R^2$). La razón de verosimilitud es equivalente al estadístico F calculado y ya que su valor-p es menor al 5% de significancia podemos deducir que el modelo tiene significancia global. Tenemos un $pseudo\, R^2 = 0.2256$, podemos interpretarlo como el porcentaje de la varianza de la variable dependiente que es explicado por el modelo especificado, no obstante, su uso no es tan recomendado debido a que no suele ser tan preciso como en los modelos de regresión lineales. Ahora veamos los coeficientes estimados.

niv_pobreza	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.1571427	.0414704	3.79	0.000	.0758622	.2384232
desague	-.4223579	.0419784	-10.06	0.000	-.5046341	-.3400816
electricidad	-.174614	.0501915	-3.48	0.001	-.2729875	-.0762405
telefono	-.5533436	.0471211	-11.74	0.000	-.6456993	-.4609879
primaria	-.3898558	.0385292	-10.12	0.000	-.4653717	-.3143398
secundaria	-.8104147	.0472125	-17.17	0.000	-.9029495	-.7178798
superior	-1.93279	.0931229	-20.76	0.000	-2.115308	-1.750273
propiedad	-.3123753	.0410453	-7.61	0.000	-.3928226	-.231928
cocina	-.4478482	.1335809	-3.35	0.001	-.7096619	-.1860345
auto	-1.363958	.1073245	-12.71	0.000	-1.57431	-1.153606
camion	-.9852788	.2933507	-3.36	0.001	-1.560236	-.4103219
habitaciones	-.2309376	.0118527	-19.48	0.000	-.2541685	-.2077066
asociacion	.0311951	.0384817	0.81	0.418	-.0442276	.1066178
personas	.3910248	.0089836	43.53	0.000	.3734172	.4086324
edad	-.0559453	.0060901	-9.19	0.000	-.0678817	-.0440088
edad2	.0004853	.0000555	8.74	0.000	.0003764	.0005941
lengua_nativa	.4281439	.0324876	13.18	0.000	.3644694	.4918183
rural	.5507213	.0409653	13.44	0.000	.4704308	.6310118
transferencias_jub	-1.033465	.1300681	-7.95	0.000	-1.288394	-.7785359
_cons	.1220407	.1692762	0.72	0.471	-.2097346	.4538159

Figura 4.108. Resultados de la estimación del modelo Logit (2).

Por el momento omitamos la interpretación de los estimadores debido a que en el Modelo Logit se busca interpretar los **odds ratio**, **efectos marginales** y **elasticidades**. Para determinar si un estimador es significativo o no, se hace uso del estadístico *Z* el cual sigue una distribución normal estándar, a diferencia de los MRLC, que para determinar la significancia individual de los estimadores usan el estadístico *t* con una distribución de *t* de Student. Con un nivel de significancia del 5% podemos ver que todos los estimadores son significativos individualmente a excepción del estimador que acompaña a la variable ***asociacion*** y el intercepto. Para corroborarlo podemos usar el contraste de Wald con el comando **test**.

```
. test asociacion

 ( 1)  [niv_pobreza]asociacion = 0

           chi2(  1) =    0.66
         Prob > chi2 =    0.4176
```

Figura 4.109. Contraste de significancia de Wald de la variable **asociacion**.

STATA realiza el contraste de Wald, el cual se distribuye siguiendo la distribución del X^2. Podemos ver que el valor-p es mayor al nivel de significancia del 5%, por tanto aceptamos la hipótesis nula y se concluye que el estimador no es significativo.

La siguiente tabla resume las anteriores figuras.

Variable dependiente: pobreza (variable dicotómica) Muestra: Encuesta Nacional de Hogares del Perú 2018 Función de distribución acumulada asumida logística (Modelo Logit)			
Variable	**Muestra Completa**	**Variable**	**Muestra Completa**
Constante	0.1220407 (0.1692762)	**Capital Social**	
Infraestructura		Asociaciones	0.0311951* (0.0384817)
Agua Potable	0.1571427 (0.0414704)	**Características del hogar o del jefe de hogar**	
Desagüe	-0.4223579 (0.0419784)	Miembros	0.3910248 (0.0089836)
Electricidad	-0.174614 (0.0501915)	Edad	-0.0559453 (0.0060901)
Teléfono	-0.5533436 (0.0471211)	Edad2	0.0004853 (0.0000555)
Capital Humano		Lengua indígena	0.4281439 (0.0324876)
Primaria completa	-0.3898558 (0.0385292)	Rural	0.5507213 (0.0409653)
Secundaria completa	-0.8104147 (0.0472125)	**Transferencias**	
Superior completa	-1.93279 (0.0931229)	Transf. Jubilación	-1.033465 (0.1300681)
Capital Física			
Título de propiedad	-0.3123753 (0 .0410453)	**N° observaciones**	37462
Cocina	-0.4478482 (0.1335809)	**LR chi2**	7971.30
Auto	-1.363958 (0.1073245)	**Prob>chi2**	0.0000
Camión	-0.9852788 (0.2933507)	**Log likelihood**	-13682.543
Habitaciones	-0.2309376 (0.0118527)	**Pseudo R2**	0.2256

Tabla 4.2. Determinantes de la pobreza bajo un enfoque de activos (estimador de Máxima Verosimilitud).

Aunque los modelos de probabilidad lineal y el modelo probit no han sido seleccionados para estimar los estimadores del modelo (4.5.4.) brevemente se explicará cómo realizar sus estimaciones en el programa STATA. Como el MPL se trata del modelo de regresión lineal clásico con una variable dependiente binomial, se puede utilizar al comando **reg** y la opción **robust** para calcular estimadores que no estén afectados por heterocedasticidad.

```
. reg $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias,robust

Linear regression                               Number of obs     =     37,462
                                                F(19, 37442)      =     457.76
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1925
                                                Root MSE          =     .34543
```

niv_pobreza	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
agua	.0225929	.006971	3.24	0.001	.0089296	.0362562
desague	-.0634568	.0059226	-10.71	0.000	-.0750653	-.0518482
electricidad	-.0646347	.0098002	-6.60	0.000	-.0838433	-.0454261
telefono	-.1040736	.0083086	-12.53	0.000	-.1203586	-.0877886
primaria	-.0606266	.0060492	-10.02	0.000	-.0724832	-.0487701
secundaria	-.1094604	.0063203	-17.32	0.000	-.1218484	-.0970725
superior	-.130089	.0062919	-20.68	0.000	-.1424212	-.1177567
propiedad	-.0251359	.0038652	-6.50	0.000	-.0327117	-.0175601
cocina	-.0241018	.0090568	-2.66	0.008	-.0418533	-.0063503
auto	-.0595442	.0036804	-16.18	0.000	-.0667578	-.0523305
camion	-.0874647	.0156227	-5.60	0.000	-.1180856	-.0568437
habitaciones	-.0232994	.0011861	-19.64	0.000	-.0256242	-.0209746
asociacion	.0037535	.0042162	0.89	0.373	-.0045102	.0120173
personas	.0518877	.0011443	45.34	0.000	.0496448	.0541306
edad	-.0059179	.0007407	-7.99	0.000	-.0073698	-.0044661
edad2	.000048	6.86e-06	7.00	0.000	.0000346	.0000615
lengua_nativa	.0531582	.0047188	11.27	0.000	.0439092	.0624071
rural	.0777914	.0055736	13.96	0.000	.066867	.0887157
transferencias_jub	-.0131748	.0056998	-2.31	0.021	-.0243465	-.0020031
_cons	.4511699	.0221242	20.39	0.000	.4078059	.4945339

Figura 4.110. Estimación mediante el Modelo de Probabilidad Lineal.

Olvidemos que es probable que los estimadores de MPL sean menos idóneos que los estimadores calculados mediante el Modelo Logit para ejemplificar cómo podríamos interpretarlos. Tomemos el caso de la variable ***agua***, podemos ver que su estimador es 0.0225929 entonces lo multiplicamos por 100 e interpretamos de la siguiente manera: "Si el hogar tiene acceso al servicio básico de agua entonces la probabilidad que el hogar sea pobre aumenta en 2.26 puntos porcentuales", tomemos ahora a una variable cuantitativa como la variable ***personas*** cuyo estimador es 0.0518877 su interpretación es: "Si el número de miembros en el hogar aumenta en una personas más, entonces la probabilidad que el hogar sea pobre aumenta en 5.19 puntos porcentuales". En el caso de los MPL no es necesario calcular los **odds ratios** ni los **efectos marginales**. Sin embargo, en los MPL, y al igual que en los Modelos Logit y Probit, podemos calcular el valor estimado en un punto específico, por ejemplo queremos calcular el valor estimado cuando el número de habitaciones en el hogar es igual a 3 manteniendo constante las demás

variables, entonces el comando **margins** y sus opciones **predict(xb)** y **at()** nos facilitara el cálculo, rehagamos la regresión sin utilizar los macros globales.

```
. reg niv_pobreza agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion habitaciones asociacion personas e
> dad edad2 lengua_nativa rural transferencias_jub, robust

. margins, predict(xb) at(habitaciones=3 )

Predictive margins                              Number of obs     =     37,462
Model VCE    : Robust

Expression   : Linear prediction, predict(xb)
at           : habitaciones    =           3

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .1859618   .0018502   100.51   0.000     .1823353    .1895882
------------------------------------------------------------------------------
```

Figura 4.111. Valor estimado cuando ***habitaciones***=3.

Como se puede ver en la figura 4.111. Dentro de los paréntesis de la opción **at()** colocamos el punto específico que deseamos estimar, este punto específico solo admite el uso del símbolo "=". Otra observación que se aprecia es el término "Expression", este término indica el tipo de predicción que hará el comando; al tratarse del MPL es lógico que indique que se trata de una predicción lineal.

El número que se ve en la columna representa el valor promedio de la variable dependiente. Recordemos que en el modelo de probabilidad lineal el promedio condicional de la variable dependiente dado la variable independiente es la probabilidad que la variable dependiente sea igual a 1. Revisemos (4.2.6.).

$E(Y_i|X_i) = \Pr(Y_i = 1|X_i)$ (4.2.6.)

Entonces su interpretación es la siguiente: "Si en toda la muestra, los hogares tuvieran 3 habitaciones, la probabilidad que un hogar sea pobre es del 19%.

Es posible que queramos estimar puntos específicos usando más de una variable explicativa, supongamos que ahora queremos calcular el valor estimado cuando un hogar tiene 3 habitaciones y no tiene acceso al servicio básico de electricidad.

```
. margins, predict(xb) at(habitaciones=3 electricidad=0)

Predictive margins                              Number of obs     =     37,462
Model VCE    : Robust

Expression   : Linear prediction, predict(xb)
at           : electricidad    =           0
               habitaciones    =           3

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .2452496   .0091762    26.73   0.000     .2272641    .2632352
------------------------------------------------------------------------------
```

Figura 4.112. Valor estimado cuando ***habitaciones***=3 y ***electricidad***=0.

Su interpretación es: "Si en toda la muestra, los hogares tienen tres habitaciones y además carece del servicio eléctrico entonces la probabilidad que el hogar sea pobre es 24%. La variable ***electricidad*** se trata de una variable dicotómica y por tanto, tiene dos posibles valores: cuando el hogar no tiene acceso a electricidad "0" y cuando el hogar tiene acceso a electricidad "1". En este ejemplo, hemos seleccionado la probabilidad para los hogares que carecen de acceso a electricidad. En la siguiente figura, utilizaremos la variable ***superior*** para estimar sus puntos específicos en sus dos únicos valores, manteniendo el número de habitaciones en un hogar igual a 3.

```
. margins, predict(xb) at(habitaciones=2 superior=0 superior=1)

Predictive margins                              Number of obs     =     37,462
Model VCE    : Robust

Expression   : Linear prediction, predict(xb)

1._at        : superior        =           0
               habitaciones    =           2

2._at        : superior        =           1
               habitaciones    =           2

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .2312772   .0027633    83.69   0.000     .2258609    .2366934
          2  |   .1011882   .0056269    17.98   0.000     .0901593     .112217
------------------------------------------------------------------------------
```

Figura 4.113. Valor estimado cuando ***habitaciones***=3, ***superior***=0 y ***superior***=1.

En la figura 4.113., se observan dos valores esperados de la variable dependiente, los cuales representan la probabilidad que tiene la variable dependiente sea igual a 1 cuando tiene el hogar tiene tres habitaciones y dado cada valor de la variable ***superior***.

Es evidente que la probabilidad que el hogar sea pobre es menor cuando el jefe de hogar tiene superior completa que cuando el jefe de hogar no la tiene.

En cuanto al Modelo Probit, este puede ser estimado utilizando el comando **probit**.

```
. probit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias

Iteration 0:   log likelihood = -17668.195
Iteration 1:   log likelihood = -13889.104
Iteration 2:   log likelihood = -13636.246
Iteration 3:   log likelihood = -13631.722
Iteration 4:   log likelihood = -13631.721

Probit regression                               Number of obs     =     37,462
                                                LR chi2(19)       =    8072.95
                                                Prob > chi2       =     0.0000
Log likelihood = -13631.721                     Pseudo R2         =     0.2285
```

niv_pobreza	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.0872849	.0241427	3.62	0.000	.039966	.1346038
desague	-.239778	.0237794	-10.08	0.000	-.2863848	-.1931712
electricidad	-.1058651	.0296583	-3.57	0.000	-.1639943	-.0477358
telefono	-.3324985	.0274464	-12.11	0.000	-.3862925	-.2787044
primaria	-.2285118	.0221575	-10.31	0.000	-.2719396	-.185084
secundaria	-.4725496	.0265272	-17.81	0.000	-.5245419	-.4205573
superior	-1.003689	.044791	-22.41	0.000	-1.091477	-.9158999
propiedad	-.1672378	.0223971	-7.47	0.000	-.2111353	-.1233402
cocina	-.2383238	.0709085	-3.36	0.001	-.3773019	-.0993457
auto	-.6937738	.051961	-13.35	0.000	-.7956156	-.5919321
camion	-.5460028	.1528332	-3.57	0.000	-.8455503	-.2464554
habitaciones	-.132351	.0065398	-20.24	0.000	-.1451686	-.1195333
asociacion	.0131065	.0216829	0.60	0.546	-.0293912	.0556042
personas	.2260902	.0050052	45.17	0.000	.2162801	.2359003
edad	-.030363	.0034934	-8.69	0.000	-.03721	-.023516
edad2	.0002614	.0000319	8.18	0.000	.0001988	.0003239
lengua_nativa	.2482986	.01853	13.40	0.000	.2119805	.2846168
rural	.3134609	.0230113	13.62	0.000	.2683596	.3585622
transferencias_jub	-.4877126	.0624155	-7.81	0.000	-.6100447	-.3653804
_cons	.0184095	.0971977	0.19	0.850	-.1720945	.2089136

Figura 4.114. Resultados de la estimación del modelo Probit.

Una característica similar entre los modelos Logit y Probit es que sus estimadores no interesan tanto para interpretarlos, ya que para que estos modelos tengan un sentido interpretativo se utilizan los **odds ratio**, **efectos marginales** y **elasticidades** en ambos, pero podemos tomar los signos de sus estimadores para tener una idea de cuál será el impacto de los efectos marginales de las variables independientes, debido a que los estimadores y los efectos marginales tienen el mismo signo.

En STATA es posible realizar varias estimaciones de distintos modelos, almacenar sus estimadores en la memoria del programa y mostrarlos en una tabla.

Veámoslo ejemplificado en la estimación del modelo (4.5.4.) para toda la muestra, cada área de residencia del hogar (urbano y rural) y para cada sexo del jefe de hogar (hombre y mujer). Como la estimación del modelo (4.5.4.) ya se ha realizado en las figuras anteriores solo mostraremos los comandos utilizados.

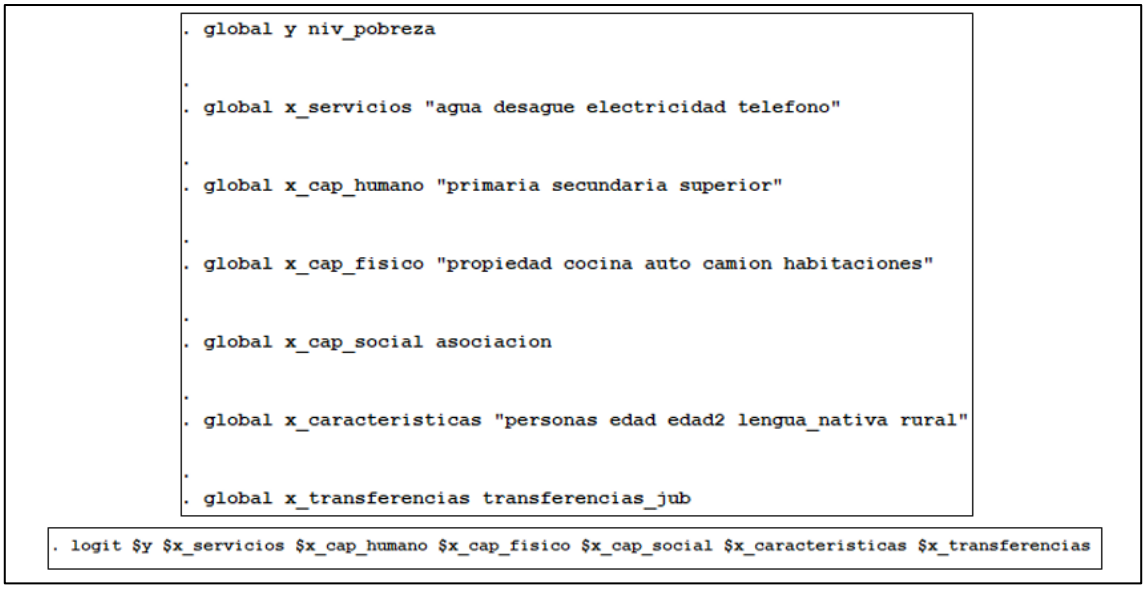


Figura 4.115. Resultados de la estimación del modelo Logit usando toda la muestra.

Inmediatamente después de la estimación debemos utilizar el comando **estimates** y su componente **store**. Este comando guarda en la memoria del programa los resultados de los estimadores después de haber estimado cualquier modelo. Para distinguirlos del resto de modelos que realizaremos después, les colocaremos el nombre "Muestra".

```
. estimates store Muestra
```

Figura 4.116. Guardando los estimadores del modelo Logit usando toda la muestra.

Debemos tener cuidado con el nombre que le asignaremos, ya que STATA no permite cambiarles de nombre y podría generar confusiones.

(Aparicio, Jaramillo, & San Román , 2011) Indican que para estimar al modelo Logit para los hogares en áreas de residencia urbanas, debemos excluir a la variable ***camion*** del grupo de activos de capital físico y a la variable ***rural*** del grupo de características. Configuremos los macros globales creados.

```
. global x_cap_fisico "propiedad cocina auto habitaciones"

. global x_caracteristicas "personas edad edad2 lengua_nativa"
```

Figura 4.117. Configuración de los macros globales creados (1).

Ahora podemos realizar la regresión del Modelo Logit, añadiendo la condicional **if** para ordenar a STATA que, solo tome en cuenta cuando el hogar esté en un área urbana (***rural***=0).

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas if (rural==0)

Iteration 0:   log likelihood = -7022.6244
Iteration 1:   log likelihood = -5972.3962
Iteration 2:   log likelihood = -5597.6941
Iteration 3:   log likelihood = -5575.7284
Iteration 4:   log likelihood = -5575.4647
Iteration 5:   log likelihood = -5575.4645

Logistic regression                             Number of obs     =     22,550
                                                LR chi2(16)       =    2894.32
                                                Prob > chi2       =     0.0000
Log likelihood = -5575.4645                     Pseudo R2         =     0.2061

------------------------------------------------------------------------------
  niv_pobreza |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        agua |  -.1241529   .0886518    -1.40   0.161    -.2979073    .0496014
     desague |  -.4202527   .0668674    -6.28   0.000    -.5513103   -.2891951
electricidad |  -.6858997    .159331    -4.30   0.000    -.9981826   -.3736167
    telefono |  -.9870773   .0994624    -9.92   0.000     -1.18202   -.7921345
     primaria |  -.5563514   .0677298    -8.21   0.000    -.6890995   -.4236034
   secundaria |  -1.065562   .0732294   -14.55   0.000    -1.209089   -.9220352
     superior |    -2.1733   .1193003   -18.22   0.000    -2.407124   -1.939476
    propiedad |   -.135927   .0560206    -2.43   0.015    -.2457253   -.0261287
       cocina |   -.717825   .2253019    -3.19   0.001    -1.159409   -.2762414
         auto |  -1.405516   .1502055    -9.36   0.000    -1.699913   -1.111118
 habitaciones |  -.3419974   .0197818   -17.29   0.000     -.380769   -.3032258
   asociacion |  -.0022441   .0629503    -0.04   0.972    -.1256245    .1211362
     personas |   .4257533   .0141068    30.18   0.000     .3981046    .4534021
         edad |  -.0229716   .0100983    -2.27   0.023     -.042764   -.0031792
        edad2 |   .0001223   .0000924     1.32   0.186    -.0000587    .0003033
lengua_nativa |   .5468416   .0570529     9.58   0.000     .4350199    .6586632
        _cons |   .8179418   .3012598     2.72   0.007     .2274834      1.4084
------------------------------------------------------------------------------
```

Figura 4.118. Resultados de la estimación del modelo Logit para los hogares en zonas urbanas.

Y utilizamos el comando **estimates store** con el nombre "Urbano" para guardar los resultados de sus estimadores.

```
. estimates store Urbano
```

Figura 4.119. Guardando los estimadores del modelo Logit para hogares en zonas urbanas.

Para estimar el modelo Logit para los hogares que están en áreas de residencia rurales, (Aparicio, Jaramillo, & San Román , 2011) Excluyen al servicio básico de ***agua***.

```
. global x_servicios "desague electricidad telefono"
```

Configuremos el macro global creado, estimemos el modelo y guardemos los estimadores.

Figura 4.120. Configuración de los macros globales creados (2).

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias if (rural==1)

Iteration 0:   log likelihood = -9239.4407
Iteration 1:   log likelihood = -8105.1608
Iteration 2:   log likelihood = -8059.2547
Iteration 3:   log likelihood = -8058.6158
Iteration 4:   log likelihood = -8058.6152

Logistic regression                             Number of obs     =     14,912
                                                LR chi2(16)       =    2361.65
                                                Prob > chi2       =     0.0000
Log likelihood = -8058.6152                     Pseudo R2         =     0.1278
```

niv_pobreza	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
desague	-.2737188	.053491	-5.12	0.000	-.3785593	-.1688783
electricidad	-.1254058	.0501463	-2.50	0.012	-.2236907	-.027121
telefono	-.4702646	.0526592	-8.93	0.000	-.5734747	-.3670545
primaria	-.3427228	.046459	-7.38	0.000	-.4337807	-.2516649
secundaria	-.6259749	.0626249	-10.00	0.000	-.7487175	-.5032323
superior	-1.556485	.1601931	-9.72	0.000	-1.870458	-1.242512
propiedad	-.4997487	.0639231	-7.82	0.000	-.6250356	-.3744617
cocina	-.3514712	.168853	-2.08	0.037	-.6824171	-.0205253
auto	-1.314505	.1548195	-8.49	0.000	-1.617946	-1.011064
habitaciones	-.162121	.0147489	-10.99	0.000	-.1910284	-.1332136
asociacion	.0283034	.0491955	0.58	0.565	-.0681179	.1247248
personas	.3740736	.0118542	31.56	0.000	.3508399	.3973073
edad	-.0643863	.0076715	-8.39	0.000	-.0794222	-.0493505
edad2	.0005755	.0000693	8.30	0.000	.0004396	.0007113
lengua_nativa	.3498006	.0393721	8.88	0.000	.2726328	.4269685
transferencias_jub	-.8863649	.2146809	-4.13	0.000	-1.307132	-.4655982
_cons	.6959238	.2100385	3.31	0.001	.2842559	1.107592

Figura 4.121. Resultados de la estimación del modelo Logit para los hogares en zonas rurales.

```
. estimates store Rural
```

Figura 4.122. Guardando los estimadores del modelo Logit para hogares en zonas rurales.

La teoría propuesta por (Aparicio, Jaramillo, & San Román , 2011) Señala que para estimar los hogares donde el jefe de hogar es masculino, no debemos excluir ninguna variable, entonces configuremos los macros globales de la misma manera que los hemos configurado para el modelo Logit estimado, usando toda la muestra y guardamos sus estimadores con el nombre “Hombre”.

```
. global x_servicios "agua desague electricidad telefono"

.
. global x_cap_fisico "propiedad cocina auto camion habitaciones"

.
. global x_caracteristicas "personas edad edad2 lengua_nativa rural"

.
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias if (sexo==1)
```

Figura 4.123. Configuración de los macros globales creados (3).

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias if (sexo==1)

Iteration 0:   log likelihood = -13305.885
Iteration 1:   log likelihood = -10657.466
Iteration 2:   log likelihood = -10242.763
Iteration 3:   log likelihood = -10220.949
Iteration 4:   log likelihood = -10220.784
Iteration 5:   log likelihood = -10220.784

Logistic regression                             Number of obs     =     26,786
                                                LR chi2(19)       =    6170.20
                                                Prob > chi2       =     0.0000
Log likelihood = -10220.784                     Pseudo R2         =     0.2319

------------------------------------------------------------------------------
  niv_pobreza |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         agua |   .1900101   .0477008     3.98   0.000     .0965182     .283502
      desague |  -.4186635   .0481628    -8.69   0.000    -.5130608    -.3242662
  electricidad |  -.0811529   .0578378    -1.40   0.161     -.194513     .0322071
     telefono |  -.5528908   .0563002    -9.82   0.000    -.6632371    -.4425446
     primaria |  -.3741244   .0443362    -8.44   0.000    -.4610217    -.2872271
   secundaria |   -.828951    .053939   -15.37   0.000    -.9346695    -.7232326
     superior |  -1.932609   .1039227   -18.60   0.000    -2.136294    -1.728925
   propiedad |  -.3209859    .047827    -6.71   0.000    -.4147251    -.2272466
      cocina |  -.5790512   .1576338    -3.67   0.000    -.8880078    -.2700946
        auto |  -1.354651   .1123063   -12.06   0.000    -1.574768    -1.134535
      camion |  -.9731071   .2950771    -3.30   0.001    -1.551448    -.3947666
 habitaciones |  -.2240369   .0133236   -16.82   0.000    -.2501507    -.1979231
   asociacion |   .0143683    .043146     0.33   0.739    -.0701964     .098933
     personas |   .3833745   .0103669    36.98   0.000     .3630556     .4036933
         edad |  -.0672975   .0071754    -9.38   0.000     -.081361    -.0532341
        edad2 |   .0006028   .0000665     9.06   0.000     .0004724     .0007331
  lengua_nativa |   .4665065   .0370397    12.59   0.000     .3939099     .539103
        rural |   .5221119   .0475868    10.97   0.000     .4288435     .6153802
transferencias_jub |  -1.094925   .1395421    -7.85   0.000    -1.368423    -.8214279
        _cons |    .352885   .1948316     1.81   0.070    -.0289779     .7347479
------------------------------------------------------------------------------
```

Figura 4.124. Resultados de la estimación del modelo Logit para los hogares con jefe de hogar masculino.

```
. estimates store Hombre
```

Figura 4.125. Guardando los estimadores del modelo Logit para hogares con jefe de hogar masculino.

En cuanto a la estimación del modelo cuando el jefe de hogar es femenino, (Aparicio, Jaramillo, & San Román , 2011) Indican que debemos excluir a la variable ***camion*** del grupo de activos de capital físico y a la variable ***rural*** del grupo de características.

```
. global x_cap_fisico "propiedad cocina auto habitaciones"

.
. global x_caracteristicas "personas edad edad2 lengua_nativa"
```

Figura 4.126. Configuración de los macros globales creados (4).

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias if (sexo==2)

Iteration 0:   log likelihood = -4263.7221
Iteration 1:   log likelihood = -3589.0227
Iteration 2:   log likelihood = -3428.6284
Iteration 3:   log likelihood = -3414.5729
Iteration 4:   log likelihood =  -3414.337
Iteration 5:   log likelihood = -3414.3369

Logistic regression                             Number of obs     =     10,676
                                                LR chi2(17)       =    1698.77
                                                Prob > chi2       =     0.0000
Log likelihood = -3414.3369                     Pseudo R2         =     0.1992
```

niv_pobreza	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.0181309	.0843822	0.21	0.830	-.1472553	.183517
desague	-.575466	.0771612	-7.46	0.000	-.7266992	-.4242328
electricidad	-.4930737	.1013394	-4.87	0.000	-.6916953	-.2944521
telefono	-.6226731	.0879071	-7.08	0.000	-.7949679	-.4503783
primaria	-.7240071	.0859647	-8.42	0.000	-.8924948	-.5555194
secundaria	-1.103744	.1072234	-10.29	0.000	-1.313898	-.8935897
superior	-2.293625	.2171892	-10.56	0.000	-2.719308	-1.867942
propiedad	-.2959907	.0800911	-3.70	0.000	-.4529663	-.1390151
cocina	-.2210137	.2528631	-0.87	0.382	-.7166163	.2745889
auto	-1.741688	.3725461	-4.68	0.000	-2.471865	-1.011511
habitaciones	-.2772601	.0264961	-10.46	0.000	-.3291916	-.2253287
asociacion	.0258858	.0863274	0.30	0.764	-.1433128	.1950843
personas	.3626603	.0190435	19.04	0.000	.3253357	.3999848
edad	-.0377303	.0121652	-3.10	0.002	-.0615737	-.013887
edad2	.0002866	.0001062	2.70	0.007	.0000785	.0004948
lengua_nativa	.3260008	.0678107	4.81	0.000	.1930943	.4589072
transferencias_jub	-1.334335	.3937313	-3.39	0.001	-2.106034	-.5626355
_cons	.752823	.3515219	2.14	0.032	.0638527	1.441793

Figura 4.127. Resultados de la estimación del modelo Logit para los hogares con jefe de hogar femenino.

```
. estimates store Mujer
```

Figura 4.128. Guardando los estimadores del modelo Logit para hogares con jefe de hogar femenino.

Para construir una tabla que muestre los distintos estimadores que hemos guardado en todas las estimaciones, la instrucción **estimates table** le ordena a STATA que elabore dicha tabla. Después de la instrucción colocamos cada nombre con que hemos guardado los estimadores.

```
. estimates table Muestra Urbano Rural Hombre Mujer
```

Variable	Muestra	Urbano	Rural	Hombre	Mujer
agua	.15714273	-.12415291		.19001012	.01813085
desague	-.42235785	-.42025268	-.27371881	-.4186635	-.57546603
electricidad	-.17461399	-.68589967	-.12540584	-.08115294	-.4930737
telefono	-.55334356	-.98707727	-.47026463	-.55289084	-.62267313
primaria	-.38985578	-.55635142	-.34272282	-.3741244	-.7240071
secundaria	-.81041468	-1.0655621	-.6259749	-.82895102	-1.1037437
superior	-1.9327903	-2.1733	-1.5564849	-1.9326092	-2.2936254
propiedad	-.31237528	-.135927	-.49974865	-.32098586	-.2959907
cocina	-.44784821	-.717825	-.35147117	-.57905123	-.22101372
auto	-1.3639581	-1.4055156	-1.314505	-1.3546515	-1.7416877
camion	-.98527876			-.97310706	
habitaciones	-.23093759	-.3419974	-.16212099	-.22403687	-.27726012
asociacion	.0311951	-.00224413	.02830344	.01436827	.02588578
personas	.39102481	.42575333	.37407359	.38337446	.36266028
edad	-.05594527	-.0229716	-.06438635	-.06729755	-.03773034
edad2	.00048528	.00012227	.00057545	.00060277	.00028663
lengua_nat~a	.42814387	.54684159	.34980064	.46650647	.32600076
rural	.55072132			.52211185	
transferen~b	-1.0334648		-.88636494	-1.0949254	-1.3343347
_cons	.12204066	.81794182	.6959238	.35288501	.75282298

Figura 4.129. Estimadores de los modelos “Muestra”, “Urbano”, “Rural”, “Hombre” y “Mujer”.

Pese a que los estimadores de los modelos Logit no suelen ser interpretados, podemos utilizar sus signos para tener una idea de lo que nos espera cuando estimemos los efectos marginales. Por ejemplo, en las 5 estimaciones la variable ***superior*** tiene un estimador con signo negativo, entonces podemos inferir que si el jefe de hogar tiene educación superior completa, la probabilidad que el hogar sea pobre es menor a los hogares que no tiene un jefe de hogar con educación superior completa.

Algo parecido podríamos hacer para comparar los distintos resultados de los estimadores, si el modelo es estimado mediante distintos métodos, por ejemplo, hagamos una tabla para comparar los estimadores, sus respectivos errores estándares, estadísticos *Z* calculados y valores-p, además de la función de log-verosimilitud y el pseudo coeficiente de determinación de cada modelo estimado (Logit y Probit). Para ello utilizaremos las opciones **stats()**, **se**, **t** y **p**. La opción **stats()** es utilizada por lo general, para mostrar los coeficientes de determinación de los modelos y otros estadísticos exclusivos de cada modelo, en el caso de los Modelos de Probabilidad no Lineal utilizamos **r2_p** y **ll** para ordenar a STATA que muestre el pseudo coeficiente de determinación y la función de log-verosimilitud. Por otro lado, las opciones **se**, **t** y **p**

indican que agreguen los errores estándares, el estadístico calculado (*t* o *z*) y el valor-p de cada estimador.

```
. estimates table Logit Probit, stats(ll r2_p) se t p
```

Variable	Logit	Probit
agua	.15714273	.0872849
	.04147041	.02414272
	3.79	3.62
	0.0002	0.0003
desague	-.42235785	-.23977799
	.04197844	.02377943
	-10.06	-10.08
	0.0000	0.0000
electricidad	-.17461399	-.10586509
	.05019148	.02965832
	-3.48	-3.57
	0.0005	0.0004
telefono	-.55334356	-.33249849
	.04712113	.02744645
	-11.74	-12.11
	0.0000	0.0000
primaria	-.38985578	-.22851177
	.03852924	.02215746
	-10.12	-10.31
	0.0000	0.0000

Figura 4.130. Estimadores de los modelos "Logit" y "Probit" (1).

secundaria	-.81041468	-.47254956
	.04721253	.02652717
	-17.17	-17.81
	0.0000	0.0000
superior	-1.9327903	-1.0036887
	.0931229	.04479103
	-20.76	-22.41
	0.0000	0.0000
propiedad	-.31237528	-.16723775
	.0410453	.02239714
	-7.61	-7.47
	0.0000	0.0000
cocina	-.44784821	-.23832376
	.13358086	.07090849
	-3.35	-3.36
	0.0008	0.0008
auto	-1.3639581	-.69377383
	.10732451	.05196104
	-12.71	-13.35
	0.0000	0.0000
camion	-.98527876	-.54600283
	.29335075	.15283315
	-3.36	-3.57
	0.0008	0.0004
habitaciones	-.23093759	-.13235096
	.01185274	.00653975
	-19.48	-20.24
	0.0000	0.0000

Figura 4.131. Estimadores de los modelos "Logit" y "Probit" (2).

asociacion	.0311951	.01310652
	.03848168	.02168289
	0.81	0.60
	0.4176	0.5455
personas	.39102481	.22609018
	.00898363	.00500524
	43.53	45.17
	0.0000	0.0000
edad	-.05594527	-.03036301
	.00609012	.00349345
	-9.19	-8.69
	0.0000	0.0000
edad2	.00048528	.00026135
	.00005554	.00003193
	8.74	8.18
	0.0000	0.0000
lengua_nat~a	.42814387	.24829864
	.03248757	.01853001
	13.18	13.40
	0.0000	0.0000
rural	.55072132	.31346092
	.04096528	.02301127
	13.44	13.62
	0.0000	0.0000
transferen~b	-1.0334648	-.48771257
	.13006815	.06241551
	-7.95	-7.81
	0.0000	0.0000

Figura 4.132. Estimadores de los modelos "Logit" y "Probit" (3).

_cons	.12204066	.01840954
	.16927619	.09719774
	0.72	0.19
	0.4709	0.8498
ll	-13682.543	-13631.721
r2_p	.22558345	.2284599

legend: b/se/t/p

Figura 4.133. Estimadores de los modelos "Logit" y "Probit" (4).

Como se puede ver, el comando **estimates store** no solo guarda el resultado de los estimadores de un modelo, también es capaz de guardar otros resultados, permitiendo así, una comparación entre los distintos métodos de estimación que estemos usando con el fin de ayudarnos a escoger el que creamos más conveniente.

4.5.6. Evaluación del cumplimiento de los supuestos.

Aunque el método de estimación de los Modelos de Probabilidad no Lineales Logit no sea igual a los Modelos Lineales, este no deja de ser un modelo que ajusta una variable dependiente en función a un conjunto de variables explicativas y como tal, presenta un término de error que representa el aspecto estocástico en el modelo. Por lo tanto, debemos analizar la capacidad de ajuste del modelo, no solo para saber si el modelo especificado realmente está correctamente estimado, sino también para comparar distintos métodos de estimación.

El comando **fitstat** es un comando de postestimación que muestra información sobre medidas de capacidad de ajuste. Es posible que el comando **fitstat** no se encuentre instalado en el programa, podremos saberlo si aparece el siguiente error.

```
. fitstat
command fitstat is unrecognized
r(199);
```

Figura 4.134. Comando **fitstat** .

Si ese fuese el caso entonces debemos usar el comando **search** seguido del comando **fitsatat** para que STATA muestre una ventana de búsqueda sobre el comando que hemos seleccionado.

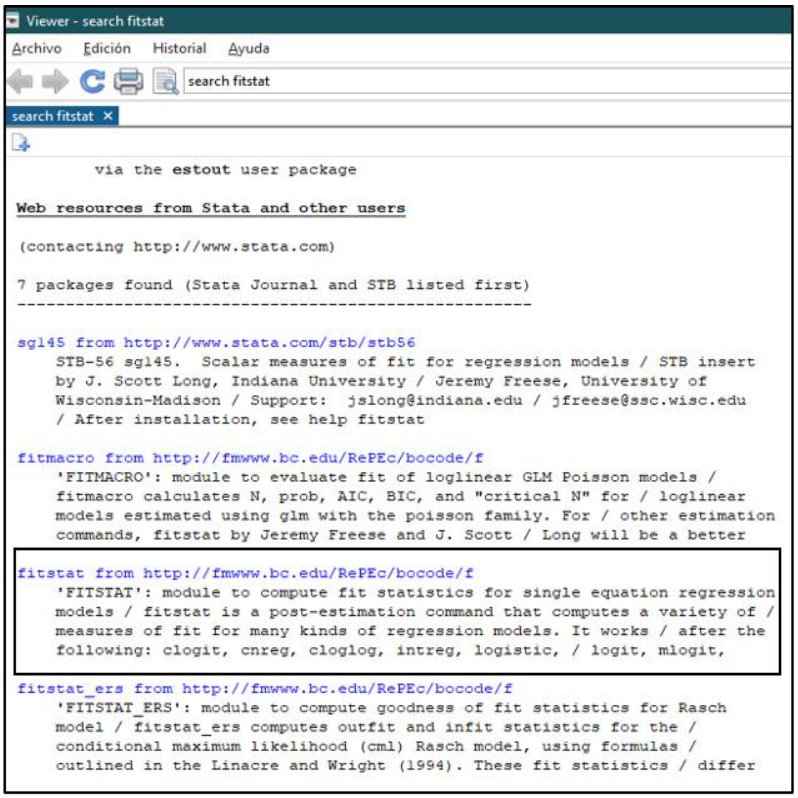


Figura 4.135. Comando **search**.

Si hacemos clic en la tercera búsqueda aparece una ventana donde, podemos ver una descripción sobre lo que es el comando, los autores y el vínculo “clic here to install” para instalarlo.

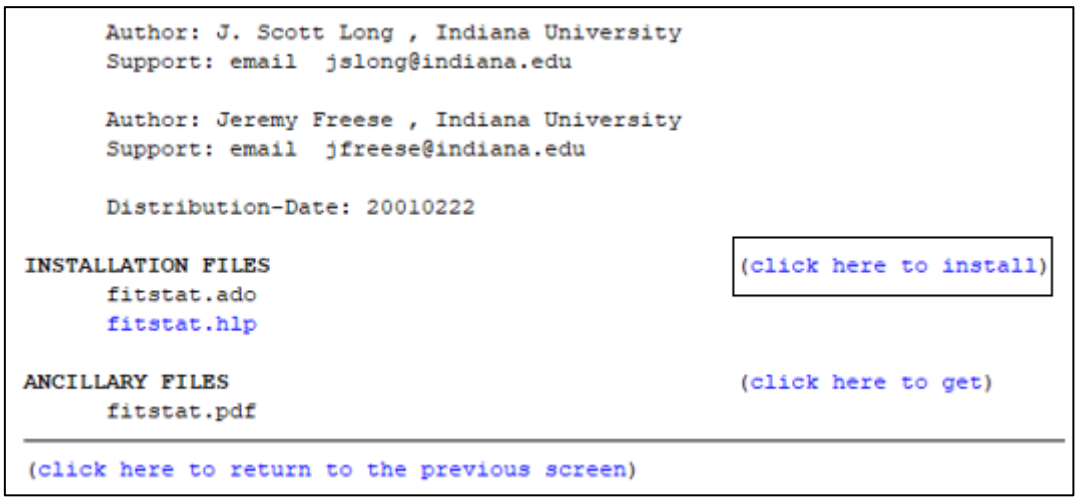


Figura 4.136. Instalando el comando **fitstat** (1).

La instalación estará completada cuando aparezca el siguiente mensaje “isntallation complete”.

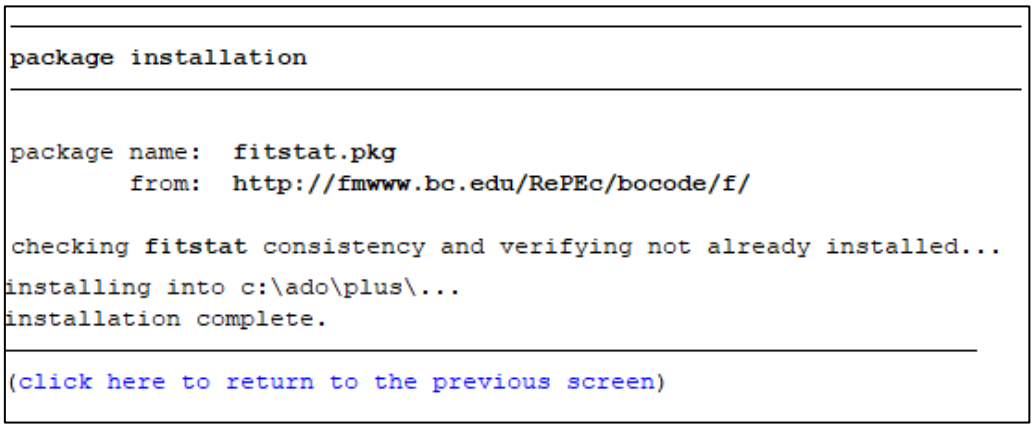


Figura 4.137. Instalando el comando **fitstat** (2).

Ahora podemos ordenar la instrucción a STATA después de estimar el modelo para toda la muestra.

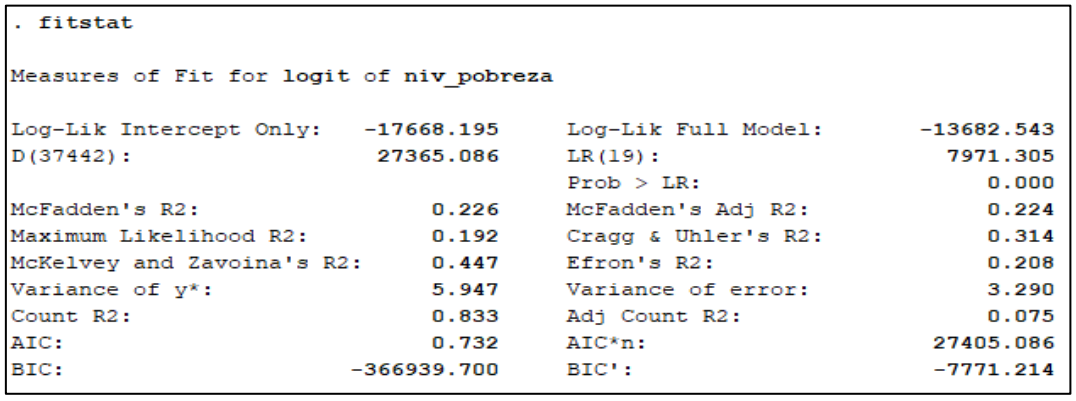


Figura 4.138. Ejecución del comando **fitstat**.

En la primera fila aparecen las funciones de log-verosimilitud, a la derecha está la función de log-verosimilitud del modelo "Log-Lik Full Model" y a la izquierda está la función de log-verosimilitud del modelo que solamente incluye al intercepto "Log-Lik Intercept Only". (Escobar M., Fernández M., & Bernardi, 2012) Definen a estas funciones como importantes para entender la estimación, ya que se pueden entender como la probabilidad que los datos de la muestra hayan sido generados por el modelo. Si "Log-Lik Full Model" es ampliamente mayor que "Log-Lik Intercept Only", entonces podemos interpretar que las variables tengan realmente un efecto sobre la variable dependiente. El término "LR(19)" es el test de razón de verosimilitud y entre sus paréntesis están los grados de libertad que el modelo usa y debajo de "LR(19)" se encuentra su valor-p, podemos ver que este último es menor a una significancia del 5% por lo que se concluye que el modelo tiene relevancia global.

En la siguiente fila se aprecia las medidas sobre bondad de ajuste más importantes de los modelos de probabilidad no lineales, se trata del $Pseudo\ R^2$ y $Pseudo\ Adj\ R^2$ también llamados $McFadden\ R^2$ "McFadden's R2" y $McFadden\ Adj\ R^2$ "McFadden's Adj R2", respectivamente. A continuación, se presenta la fórmula del $Pseudo\ R^2\ ajustado$.

$$Pseudo\ Adj\ R^2 = 1 - \frac{\ln L_F - (k+1)}{\ln L_0} \quad (4.5.6.)$$

La importancia del $Pseudo\ Adj\ R^2$ es que corrige al $Pseudo\ R^2$ su naturaleza de incrementar artificialmente al añadir nuevas variables. Podemos interpretar $Pseudo\ R^2$ como la bondad de ajuste que tiene el modelo al momento de explicar a la variable dependiente, como ya se mencionó anteriormente.

En las siguientes filas podemos ver otras medidas de $Pseudo\ R^2$ menos usadas y poco frecuentes. No obstante, de entre todas esas medidas, $Count\ R^2$ "Count R2" y $Adj\ Count\ R^2$ "Count Adj R2" merecen nuestra atención, ya que están basadas en la comparación de los datos observados en la muestra y los datos estimados por el modelo que hemos especificado, pero hablaremos después de esas medidas.

Para finalizar, en las dos últimas filas se pueden ver los **criterios** de **información,** cuya función es netamente comparar los resultados de varios modelos, incluida la estimación del mismo modelo usando distintas muestras. Se tratan del "AIC" (Akaike

Information Criteria) y "BIC" (Bayesian Information Criteria), respectivamente. Se calculan siguiendo la siguiente fórmula.

$$AIC = \frac{-2\ln L_F + 2(k+1)}{n} = \frac{-2(-13682{,}543) + 2(20)}{37462} = 0.731 \text{ (4.5.7.)}$$

$$BIC = -2\ln L_F - (n-k-1)\ln n = -366939.70 \text{ (4.5.8.)}$$

AIC y BIC tienen dos variantes causadas por discrepancias entre los autores sobre los detalles en sus fórmulas, estas son AIC*n "AIC*n" y BIC' "BIC'" respectivamente.

$$AIC * n = -2\ln L_F + 2(k+1) = 27405.086 \text{ (4.5.9.)}$$

$$BIC' = -LR + k\ln n = -7771.214 \text{ (4.5.10.)}$$

Las interpretaciones de los criterios de información, consiste en tomar en cuenta al modelo con los criterios de información menores, es decir el modelo con el criterio AIC menor es el mejor ajustado y el modelo con el criterio BIC más negativo se le considera un mejor ajuste. (Escobar M., Fernández M., & Bernardi, 2012) Recomiendan usar el criterio BIC para comparar distintos modelos Logit. Cabe señalar, que estos criterios de información no suponen una forma estricta de decidir cuál modelo es el idóneo para el tema investigado. El marco teórico, el cumplimiento de los signos esperados, las significancias y que el modelo esté libre de violaciones a los supuestos, son otros aspectos que debemos tener en cuenta.

Retomemos la medida $Count\ R^2$, como se dijo, esta medida está basada en la comparación entre los datos observados y los datos estimados. Es fundamental entender el uso del comando **estat classification** o su abreviatura **estat class,** debido a que muestran una serie de estadísticos de clasificación que nos permiten ampliar nuestro punto de vista sobre el contraste en el que se basan $Count\ R^2$ y $Adj\ Count\ R^2$. Se trata de un comando de postestimación que calcula una medida de bondad de ajuste basada en el porcentaje correcto de observaciones clasificadas. No olvidemos que, los modelos Logit predicen la probabilidad de ocurrencia de la variable dependiente, entonces en aquellas observaciones donde nuestro modelo predice más de 0.5 de probabilidad que la variable dependiente tenga éxito ($Y_i = 1$), la predicción es que ocurra "+Classified", por otro lado, en las observaciones donde el modelo predice una probabilidad inferior o igual a 0.5 entonces se predice que la variable dependiente no tendrá éxito ($Y_i = 0$) "-Classified", así lo explican (Escobar M., Fernández M., & Bernardi, 2012). En términos

matemáticos podemos utilizar la expresión que brindan (Colin C. & Trivedi, 2009) Refiriéndose a la clasificación.

$$\hat{Y}_i = 1 \leftarrow G(X'\beta) > 0.5 \;\&\; \hat{Y}_i = 0 \leftarrow G(X'\beta) \leq 0.5 \quad (4.5.11.)$$

En la siguiente figura se muestra los resultados del comando **estat clasification**.

```
. estat class

Logistic model for niv_pobreza

              ——————— True ———————
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      1551          1042  |       2593
     -     |      5198         29671  |      34869
-----------+--------------------------+-----------
   Total   |      6749         30713  |      37462

Classified + if predicted Pr(D) >= .5
True D defined as niv_pobreza != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   22.98%
Specificity                     Pr( -|~D)   96.61%
Positive predictive value       Pr( D| +)   59.81%
Negative predictive value       Pr(~D| -)   85.09%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    3.39%
False - rate for true D         Pr( -| D)   77.02%
False + rate for classified +   Pr(~D| +)   40.19%
False - rate for classified -   Pr( D| -)   14.91%
--------------------------------------------------
Correctly classified                        83.34%
--------------------------------------------------
```

Figura 4.139. Ejecución del comando **estat class**.

Primero veamos la tabla en la parte superior, en este modelo hay 1551 observaciones que están clasificadas correctamente como 1 y 29671 observaciones correctamente clasificadas como 0. Si sumamos 1551 con 29671 obtenemos 31222 observaciones correctamente clasificadas, dividamos ahora 31222 entre el total de observaciones que son 37462 y obtenemos el porcentaje de observaciones correctamente clasificadas “Correctly clasified”, el cual es 83.34%. Lo descrito anteriormente, es justamente la fórmula del $Count\ R^2$. Al mismo tiempo, la tabla muestra que hay 1042 observaciones incorrectamente clasificadas como 1 cuando su clasificación correcta debió ser 0 y hay 5198 observaciones clasificadas incorrectamente como 1 cuando debieron estar clasificadas como 0. No obstante, la interpretación de $Count\ R^2$ en ocasiones puede ser irrelevante si tomamos todas las observaciones de la categoría con más casos, lo que provoca una excesiva capacidad predictiva del modelo. Por ejemplo, se

sabe que los hogares en situación no pobre son el 81.99% de toda la muestra, entonces pronosticando para los hogares que no son pobres ya se tiene más de un 81.99% de aciertos. Para arreglar esa exageración en la capacidad predictiva podemos usar el $Adj\ Count\ R^2$, lo podemos calcular restando tanto al denominador como al numerador, la frecuencia marginal más alta entre la ocurrencia o no.

$Adj\ Count\ R^2 = \frac{31222-30713}{37462-30713} = 0.07$ (4.5.12.)

Esta medida tiene una interpretación más justa y relevante que la interpretación del $Count\ R^2$, podríamos interpretarlo como la medida de capacidad de acierto con respecto a lo que se tendría si solo predecimos las observaciones con la categoría más común siendo del 7% la capacidad de predicción en el modelo.

Por debajo de la tabla encontramos dos estadísticos absolutamente cruciales para entender la capacidad predictiva del modelo estimado, se tratan de la Sensibilidad "Sensitivity" y la Especificidad "Specificity", dos estadísticos cuyas interpretaciones se concentra respectivamente, en el cálculo de la probabilidad de clasificar correctamente a aquellas observaciones con la categoría positiva, es decir ($Y_i = 1$), y en la probabilidad de clasificar a las observaciones con la categoría negativa correctamente, es decir ($Y_i = 0$) . Ambos estadísticos se calculan mediante la división de las observaciones correctamente clasificadas entre el total de observaciones para cada categoría, por ejemplo, en el modelo, la tasa de la sensibilidad es $1551/6749 = 22.98\%$ y la tasa de especificidad es $29671/30713 = 96.60\%$. Por último, en las 4 últimas filas encontramos los ratios de los observaciones que han sido incorrectamente clasificadas, "False + rate for true ~D" se calcula mediante $1042/30713 = 3.39\%$ y "False - rate for true D" se calcula mediante $5198/6749 = 77.02\%$.

Si agregamos la opción **cutoff(),** podemos especificar el valor para determinar si una observación tiene un resultado positivo predicho. Los autores consideran el uso de dos tasas de pobreza en la opción **cutoff()** siendo las tasas de pobreza poblacional y muestral, 20.42% y 18.01% respectivamente. A continuación, veremos los resultados del comando **estat class** utilizando ambas tasas de pobreza para ordenar a STATA que considere a aquellas tasas como la probabilidad de ocurrencia. Es decir, siguiendo las siguientes expresiones:

$Pobreza\ poblacional$: $\hat{Y}_i = 1 \leftarrow G(X'\beta) > 0.2042$ & $\hat{Y}_i = 0 \leftarrow G(X'\beta) \leq 0.2042$ (4.5.13.)

$Pobreza\ muestral$: $\hat{Y}_i = 1 \leftarrow G(X'\beta) > 0.1801$ & $\hat{Y}_i = 0 \leftarrow G(X'\beta) \leq 0.1801$ (4.5.14.)

```
. estat class,cutoff(0.2042)

Logistic model for niv_pobreza

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
         + |      4959        7687 |      12646
         - |      1790       23026 |      24816
-----------+----------------------+-----------
     Total |      6749       30713 |      37462

Classified + if predicted Pr(D) >= .2042
True D defined as niv_pobreza != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   73.48%
Specificity                     Pr( -|~D)   74.97%
Positive predictive value       Pr( D| +)   39.21%
Negative predictive value       Pr(~D| -)   92.79%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   25.03%
False - rate for true D         Pr( -| D)   26.52%
False + rate for classified +   Pr(~D| +)   60.79%
False - rate for classified -   Pr( D| -)    7.21%
-------------------------------------------------
Correctly classified                        74.70%
-------------------------------------------------
```

Figura 4.140. Ejecución del comando **estat class** con la tasa de pobreza poblacional.

```
. estat class,cutoff(0.1801)

Logistic model for niv_pobreza

                 True
Classified |         D          ~D |      Total
-----------+----------------------+-----------
         + |      5244        8845 |      14089
         - |      1505       21868 |      23373
-----------+----------------------+-----------
     Total |      6749       30713 |      37462

Classified + if predicted Pr(D) >= .1801
True D defined as niv_pobreza != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   77.70%
Specificity                     Pr( -|~D)   71.20%
Positive predictive value       Pr( D| +)   37.22%
Negative predictive value       Pr(~D| -)   93.56%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   28.80%
False - rate for true D         Pr( -| D)   22.30%
False + rate for classified +   Pr(~D| +)   62.78%
False - rate for classified -   Pr( D| -)    6.44%
-------------------------------------------------
Correctly classified                        72.37%
-------------------------------------------------
```

Figura 4.141. Ejecución del comando **estat class** con la tasa de pobreza muestral.

Usando las tasas de pobreza tanto, muestral como poblacional, podemos notar que sus respectivos ratios de Sensibilidad “Sensitivity” y Especificidad “Specificity” están más cercanos a sus respectivos ratios de observaciones clasificadas correctamente “Correctly classified”, que en los resultados sobre la tabla de clasificación utilizando el 0.5 de probabilidad. Debido a que, estamos estimando el modelo Logit desde una muestra es que, el ratio Sensibilidad “Sensitivity” es mayor usando la tasa de pobreza muestral que usando la tasa de pobreza poblacional.

En la teoría sobre modelos de elección binaria existen algunas formas gráficas que pueden ayudar a elegir entre un modelo u otro, es el caso de la **curva ROC**, cuyo nombre proviene de **Receiver Operating Characteristics** (Característica Operativa del Receptor), se trata de una curva que representa el ratio entre la razón de las observaciones clasificadas correctamente como positivas ($Y_i = 1$) contra la razón de las observaciones clasificadas incorrectamente como negativas ($Y_i = 0$) según un umbral de decisión.

En una gráfica ROC, tenemos en el eje Y a la Sensibilidad y en el eje X se encuentra la tasa de las observaciones clasificadas como negativas incorrectamente (1-Especificidad), desde el origen se encuentra una línea diagonal que divide el espacio de la gráfica y por encima se encuentra la curva de ROC. Según (Colin C. & Trivedi, 2005) Si el modelo tiene una pésima capacidad predictiva la curva ROC es la línea diagonal, mientras la curva se aleje más de la línea diagonal hacia arriba, entonces la capacidad predictiva es más óptima. Concretamente si el valor del **AUC**, **Area Under the Curve** (área bajo la curva) se encuentra entre 0.5 y 0.6 el test es malo, si se encuentra entre 0.6 y 0.75 es un test regular, 0.75 y 0.9 el test es bueno, 0.9 y 0.97 el test es muy bueno y si se encuentra entre 0.97 y 1 el test es excelente. Para realizar la curva de ROC en STATA digitamos el comando **lroc** y nos devolverá una gráfica con la curva ROC y en la consola el valor del AUC.

```
. lroc

Logistic model for niv_pobreza

number of observations =    37462
area under ROC curve   =   0.8246
```

Figura 4.142. Ejecución del comando **lroc** (1).

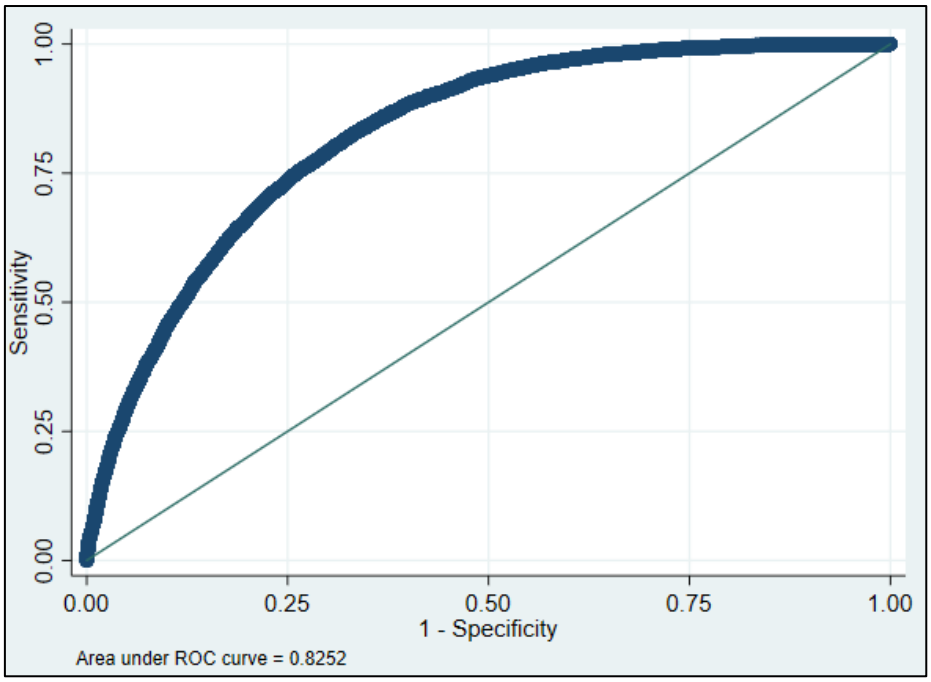


Figura 4.143. Ejecución del comando **lroc** (2).

El valor del AUC es 0.8252, lo que se traduce en un modelo que tiene una buena capacidad predictiva. Otra grafica parecido a la curva de ROC es la que proporciona el comando **lsens**, la cual genera una gráfica de la sensibilidad y especificidad versus al corte de probabilidad.

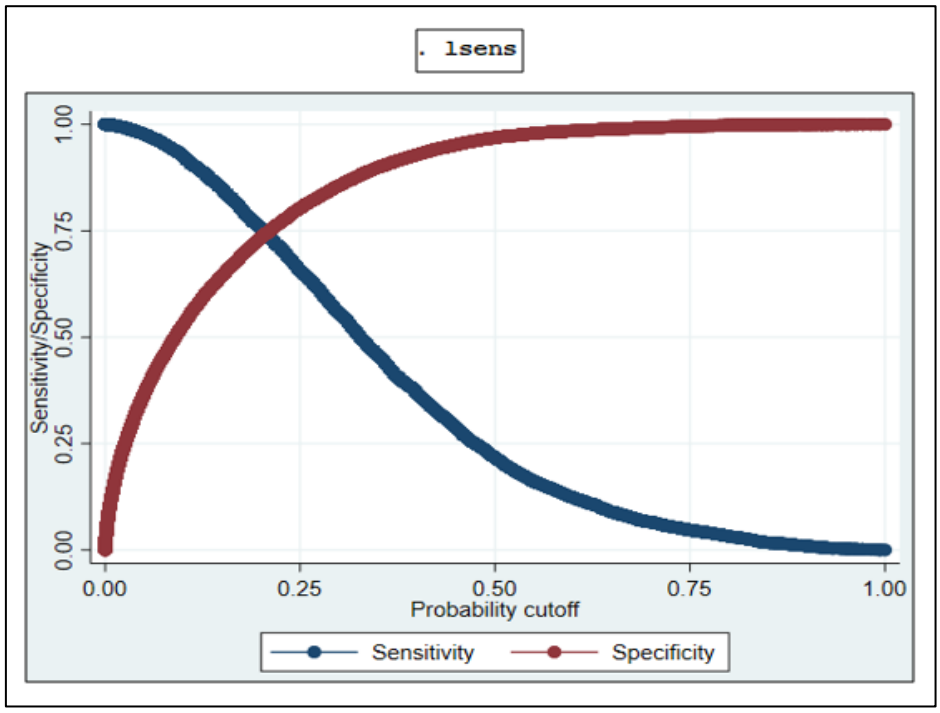


Figura 4.144. Ejecución del comando **lsens**.

Podemos ver que, la probabilidad del punto de corte se acerca mucho a las tasas de pobreza poblacional y muestral, en lugar del punto de corte predeterminado por

STATA siendo de 0.5. Ya que, el punto de corte que maximiza las medidas de sensibilidad y especificidad es el punto de corte visto en la gráfica, generado por el comando **lsens**, concluimos que sería un mejor el punto de corte que se acerca a las tasas de pobreza poblacional y muestral.

4.5.7. Interpretación de los resultados.

Recordemos, en los modelos Logit y Probit no podemos interpretar sus estimadores sino el signo de sus estimadores, debido a que son modelos no lineales. El sentido interpretativo que le damos al modelo radica principalmente en la interpretación de sus respectivos **odds ratio**, sus **efectos marginales** y en la **predicción de las probabilidades**. Empezamos con los **odss ratio**, primero veamos las estimaciones del modelo Logit usando toda la muestra.

niv_pobreza	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.1571427	.0414704	3.79	0.000	.0758622	.2384232
desague	-.4223579	.0419784	-10.06	0.000	-.5046341	-.3400816
electricidad	-.174614	.0501915	-3.48	0.001	-.2729875	-.0762405
telefono	-.5533436	.0471211	-11.74	0.000	-.6456993	-.4609879
primaria	-.3898558	.0385292	-10.12	0.000	-.4653717	-.3143398
secundaria	-.8104147	.0472125	-17.17	0.000	-.9029495	-.7178798
superior	-1.93279	.0931229	-20.76	0.000	-2.115308	-1.750273
propiedad	-.3123753	.0410453	-7.61	0.000	-.3928226	-.231928
cocina	-.4478482	.1335809	-3.35	0.001	-.7096619	-.1860345
auto	-1.363958	.1073245	-12.71	0.000	-1.57431	-1.153606
camion	-.9852788	.2933507	-3.36	0.001	-1.560236	-.4103219
habitaciones	-.2309376	.0118527	-19.48	0.000	-.2541685	-.2077066
asociacion	.0311951	.0384817	0.81	0.418	-.0442276	.1066178
personas	.3910248	.0089836	43.53	0.000	.3734172	.4086324
edad	-.0559453	.0060901	-9.19	0.000	-.0678817	-.0440088
edad2	.0004853	.0000555	8.74	0.000	.0003764	.0005941
lengua_nativa	.4281439	.0324876	13.18	0.000	.3644694	.4918183
rural	.5507213	.0409653	13.44	0.000	.4704308	.6310118
transferencias_jub	-1.033465	.1300681	-7.95	0.000	-1.288394	-.7785359
_cons	.1220407	.1692762	0.72	0.471	-.2097346	.4538159

Figura 4.108. Resultados de la estimación del modelo Logit (2).

Para ver los respectivos **odds ratio** de un modelo Logit, existen tres formas que se complementan, la primera forma es introduciendo la opción **or** en el comando **logit**. Podemos agregar la opción **nolog** para pedir a STATA que no nos muestra las iteraciones.

```
. logit $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias, nolog or
Logistic regression                                   Number of obs     =     37,462
                                                      LR chi2(19)       =    7971.30
                                                      Prob > chi2       =     0.0000
Log likelihood = -13682.543                           Pseudo R2         =     0.2256

------------------------------------------------------------------------------
  niv_pobreza | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        agua |   1.170163   .0485271     3.79   0.000     1.078814    1.269246
     desague |   .6554994   .0275168   -10.06   0.000     .6037264    .7117122
 electricidad |   .8397811   .0421499    -3.48   0.001     .7611023    .9265933
    telefono |    .575024   .0270958   -11.74   0.000     .5242958    .6306603
     primaria |   .6771545   .0260903   -10.12   0.000     .6279017    .7302708
   secundaria |   .4446736   .0209942   -17.17   0.000     .4053722    .4877854
     superior |   .1447438    .013479   -20.76   0.000     .1205962    .1737265
   propiedad |   .7317069   .0300331    -7.61   0.000     .6751485    .7930032
      cocina |   .6390017   .0853584    -3.35   0.001     .4918105    .8302449
        auto |   .2556469   .0274372   -12.71   0.000     .2071504     .315497
      camion |   .3733351   .1095181    -3.36   0.001     .2100866    .6634367
 habitaciones |    .793789   .0094086   -19.48   0.000     .7755611    .8124453
   asociacion |   1.031687    .039701     0.81   0.418     .9567362    1.112509
     personas |   1.478495   .0132823    43.53   0.000      1.45269    1.504758
        edad |   .9455909   .0057588    -9.19   0.000      .934371    .9569455
       edad2 |   1.000485   .0000556     8.74   0.000     1.000376    1.000594
lengua_nativa |   1.534407   .0498491    13.18   0.000      1.43975    1.635287
       rural |   1.734504   .0710544    13.44   0.000     1.600684    1.879511
transferencias_jub |   .3557721   .0462746    -7.95   0.000     .2757133    .4590777
       _cons |     1.1298   .1912482     0.72   0.471     .8107994    1.574308
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Figura 4.145. Odds Ratios (1).

Comparemos la primera columna de ambas tablas, que se ven en las figuras 4.108., y 4.145. La primera figura muestra el valor de los estimadores mientras que en la segunda figura se ve el valor de los odds ratio de cada variable. Por el contrario, el resto de columnas se mantienen iguales.

De forma similar ocurre cuando utilizamos el comando **logistic**.

```
. logistic $y $x_servicios $x_cap_humano $x_cap_fisico $x_cap_social $x_caracteristicas $x_transferencias
Logistic regression                                   Number of obs     =     37,462
                                                      LR chi2(19)       =    7971.30
                                                      Prob > chi2       =     0.0000
Log likelihood = -13682.543                           Pseudo R2         =     0.2256

------------------------------------------------------------------------------
  niv_pobreza | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        agua |   1.170163   .0485271     3.79   0.000     1.078814    1.269246
     desague |   .6554994   .0275168   -10.06   0.000     .6037264    .7117122
 electricidad |   .8397811   .0421499    -3.48   0.001     .7611023    .9265933
    telefono |    .575024   .0270958   -11.74   0.000     .5242958    .6306603
     primaria |   .6771545   .0260903   -10.12   0.000     .6279017    .7302708
   secundaria |   .4446736   .0209942   -17.17   0.000     .4053722    .4877854
     superior |   .1447438    .013479   -20.76   0.000     .1205962    .1737265
   propiedad |   .7317069   .0300331    -7.61   0.000     .6751485    .7930032
      cocina |   .6390017   .0853584    -3.35   0.001     .4918105    .8302449
        auto |   .2556469   .0274372   -12.71   0.000     .2071504     .315497
      camion |   .3733351   .1095181    -3.36   0.001     .2100866    .6634367
 habitaciones |    .793789   .0094086   -19.48   0.000     .7755611    .8124453
   asociacion |   1.031687    .039701     0.81   0.418     .9567362    1.112509
     personas |   1.478495   .0132823    43.53   0.000      1.45269    1.504758
        edad |   .9455909   .0057588    -9.19   0.000      .934371    .9569455
       edad2 |   1.000485   .0000556     8.74   0.000     1.000376    1.000594
lengua_nativa |   1.534407   .0498491    13.18   0.000      1.43975    1.635287
       rural |   1.734504   .0710544    13.44   0.000     1.600684    1.879511
transferencias_jub |   .3557721   .0462746    -7.95   0.000     .2757133    .4590777
       _cons |     1.1298   .1912482     0.72   0.471     .8107994    1.574308
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Figura 4.146. Odds Ratios (2).

Es posible que, ante tantos números que se ven en las tablas de resultados provistos por los comandos, la persona que está llevando a cabo la investigación se sienta agobiado por tantos resultados engorrosos. Para superar este problema, se suele utilizar el comando **listcoef** con su opción **help**. La finalidad de este comando es crear una tabla donde estén los estimadores, sus respectivos odds ratio y otros estadísticos de los estimadores con una breve descripción en la parte inferior de la tabla.

```
. listcoef, help

logistic (N=37462): Factor Change in Odds

  Odds of: Pobre vs No_pobre
```

niv_pobreza	b	z	P>|z|	e^b	e^bStdX	SDofX
agua	0.15714	3.789	0.000	1.1702	1.0606	0.3744
desague	-0.42236	-10.061	0.000	0.6555	0.8135	0.4888
electricidad	-0.17461	-3.479	0.001	0.8398	0.9530	0.2755
telefono	-0.55334	-11.743	0.000	0.5750	0.8438	0.3069
primaria	-0.38986	-10.118	0.000	0.6772	0.8381	0.4531
secundaria	-0.81041	-17.165	0.000	0.4447	0.6932	0.4522
superior	-1.93279	-20.755	0.000	0.1447	0.4845	0.3750
propiedad	-0.31238	-7.611	0.000	0.7317	0.8609	0.4793
cocina	-0.44785	-3.353	0.001	0.6390	0.9335	0.1535
auto	-1.36396	-12.709	0.000	0.2556	0.6677	0.2962
camion	-0.98528	-3.359	0.001	0.3733	0.9272	0.0768
habitaciones	-0.23094	-19.484	0.000	0.7938	0.6794	1.6738
asociacion	0.03120	0.811	0.418	1.0317	1.0131	0.4170
personas	0.39102	43.526	0.000	1.4785	2.1079	1.9070
edad	-0.05595	-9.186	0.000	0.9456	0.4171	15.6303
edad2	0.00049	8.737	0.000	1.0005	2.3225	1736.4149
lengua_nat~a	0.42814	13.179	0.000	1.5344	1.2189	0.4624
rural	0.55072	13.444	0.000	1.7345	1.3094	0.4895
transferen~b	-1.03346	-7.946	0.000	0.3558	0.7708	0.2520

```
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

Figura 4.147. Odds Ratios (3).

De izquierda a derecha tenemos las siguientes columnas: estimadores del modelo Logit "b", estadístico *Z* calculado "z", valor-p "P>|z|", odds ratios "e^b", cambio en las odds ratio por un incremento de la variable independiente en su desviación típica "e^bStdX" y desviación estándar de la variable independiente "SDofX". La pregunta que surge a continuación es: ¿Cómo se interpretan los odds ratio? Recordemos que, se le define a los odds ratio como una razón entre la probabilidad de éxito que tiene la variable dependiente sobre la probabilidad de fracaso de la variable dependiente, también recuerde

que se determina que la variación del odds ratio es negativa si el valor que acompaña a la regresora se encuentra entre 0 y 1, y la variación del odds ratio es positiva si es mayor a 1 y dependiendo si la regresora es cuantitativa o cualitativa la interpretación es distinta. Por ejemplo, tomemos a la variable ***personas***, su odds ratio se interpreta de la siguiente manera "Si incrementa en una persona el número de miembros en un hogar, entonces la razón de probabilidad que el hogar sea pobre aumenta 1,47 veces", ahora tomemos a la variable ***desague***, "Si el hogar cuenta con servicio de red de desagüe, entonces la razón de probabilidad que el hogar sea pobre disminuye 0,65 veces".

Cuando una variable dicotómica tiene un odds ratio menor a 1 conviene calcular su inversa, con el fin de comparar el efecto relativo entre sus categorías, por ejemplo, tomemos una vez más a la variable ***desague*** para calcular el inverso de su odds ratio siendo $1/0.6555 = 1.5255$, y podemos interpretarlo como "los hogares que no tienen acceso al servicio de desagüe tiene la razón que el hogar sea pobre 1,52 veces más que los hogares que tienen servicio de desagüe", el cálculo de su inversa no solo se limita a comparar las categorías de una dicotómica, sino también entre variables dicotómicas, por ejemplo, ¿Qué variable tiene más efectos sobre la probabilidad que el hogar sea pobre $(Y = 1)$, ***desague*** o ***agua***? Como la variable ***agua*** tiene un odds ratio de 1,17 no es necesario calcular su inversa, mientras que para la variable ***desague*** si ha sido necesario calcular su inversa siendo de 1,52; entonces, podemos ver que la variable ***desague*** tiene un efecto superior a la variable ***agua***.

La comparación anteriormente explicada usando la inversa de algunas variables dicotómicas, no permite comparar los odds ratio entre variables cuantitativas y cualitativas ya que no tienen un rango de valores iguales. Para superar esta dificultad podemos recurrir a la quinta columna que se ve en tabla de la figura 4.147. Se trata de la columna "e^bStdX", la cual no utiliza un incremento unitario en la variable independiente, sino utiliza a la desviación estándar como el incremento de dicha variable independiente.

Si comparamos las variables ***edad*** y ***desague***, entonces podemos interpretar que la primera tiene un efecto mayor en la razón que la segunda, ya que "e^bStdX" de la variable ***edad*** se acerca más a 0 que el de la variable ***desague***. Puede parecer contradictorio, entonces calculemos sus inversos y observemos cual variable tiene el mayor efecto, siendo $1/0.8135 = 1.2292$ el inverso de la variable ***desague*** y

$1/0.4171 = 2.3976$ el inverso de la variable ***edad***, entonces es más visible que la variable ***edad*** tiene un efecto superior que la variable ***desague***.

Como se dijo cuando se explicó la interpretación de los odds ratios, estos no calculan la probabilidad de ocurrencia y de fracaso de la variable dependiente, sino la razón entre la probabilidad de ocurrencia sobre la probabilidad de fracaso. Sin embargo, STATA permite la predicción, tanto de la probabilidad de ocurrencia como la probabilidad de no ocurrencia de la variable dependiente, a partir del uso de condicionales predeterminadas de valores de las variables regresoras con el uso del comando **prvalue**. Por ejemplo, si queremos saber la probabilidad que un hogar sea pobre y no pobre cuando el hogar tiene acceso a los servicios básicos de agua, desagüe, electricidad y teléfono entonces utilicemos el comando **prvalue** con su opción **x()** para determinar las condiciones a partir de las regresoras que representan a los servicios básicos determinados. Por otro lado, la opción **rest(mean)** indica a STATA que tome el promedio de las variables regresoras que no están predeterminadas, esta opción por defecto indica el promedio (mean) pero podemos cambiar el estadístico descriptivo según cada investigador. Veamos un ejemplo donde solicitamos que utilice el promedio de las regresoras que no han sido utilizadas como condicionantes.

```
. prvalue, x(agua=1 desague=1 electricidad=1 telefono=1) rest(mean)

logit: Predictions for niv_pobreza

Confidence intervals by delta method

                              95% Conf. Interval
  Pr(y=Pobre|x):      0.0856   [ 0.0811,    0.0900]
  Pr(y=No_pobre|x):   0.9144   [ 0.9100,    0.9189]

            agua       desague  electricidad      telefono      primaria    secundaria      superior
x=             1             1             1             1     .28863915     .28671721     .16923816

       propiedad        cocina          auto        camion  habitaciones    asociacion      personas
x=     .35764241     .02415781     .09716513     .00592601     3.2491858     .22414714     3.5330468

            edad         edad2  lengua_nat~a         rural  transferen~b
x=      53.26651     3081.6204      .3097005      .3980567     .06812236
```

Figura 4.148. Cálculo de las probabilidades de ocurrencia de la variable dependiente cuando el hogar tiene acceso a los servicios básicos de agua, desagüe, electricidad y teléfono.

En la parte superior, el término "logit: Predictions for **niv_pobreza**" indica que modelo hemos utilizado y la variable dependiente del modelo para calcular sus respectivas probabilidades. El término "Pr(y=Pobre|x)" es la probabilidad que el hogar sea pobre dada las condiciones de las regresoras seleccionadas en la opción **x()**, la cual es 0.0856, es decir si un hogar tiene acceso a los servicios básicos de agua, desagüe,

electricidad y teléfono tiene una probabilidad de 8.56% de ser pobre. Por otro lado el término "Pr(y=No_pobre|x)" es la probabilidad que el hogar tiene de no ser pobre si recibe acceso a los servicios básicos, siendo esta probabilidad del 91.44%. Al lado de las probabilidades se encuentran sus intervalos de confianza al 95%. Estas probabilidades se consiguen si ordenamos a STATA que utilice el promedio de las demás variables regresoras que no se han tomado en cuenta como condicionantes. Si digitamos solamente el comando **prvalue**, entonces estaríamos ordenando a STATA que calcule las probabilidades de éxito y fracaso utilizando el promedio de todas las variables.

Recordemos que los estimadores del modelo Logit no pueden ser interpretados de forma literal debido a que estamos ante un modelo no Lineal, por lo que solo podríamos tomar los signos que acompañan a los estimadores. Para lograr cuantificar los efectos de las variables independientes sobre la probabilidad de ocurrencia de la variable dependiente, necesitamos calcular los respectivos **efectos marginales** de las variables independientes. En STATA, es posible el cálculo de dos tipos de efectos marginales, siendo estos MER "Marginal effects at a Representative value" (Efectos Marginales a un valor Representativo) y MEM "Marginal Effects at the Mean" (Efecto Marginal en la Media). Para los dos tipos de efectos marginales se pueden utilizar los comandos **mfx** y **prchange** como comandos que se complementan. Empecemos explicando el MEM, para ello ejecutamos el comando **mfx**.

```
. mfx

Marginal effects after logit
      y  = Pr(niv_pobreza) (predict)
         =  .10375213
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
agua*	.0140247	.00355	3.95	0.000	.007059	.020991	.831429
desague*	-.0408051	.00425	-9.61	0.000	-.049128	-.032482	.605494
electr~d*	-.0172024	.00525	-3.28	0.001	-.027487	-.006918	.917276
telefono*	-.061153	.00616	-9.93	0.000	-.073229	-.049077	.894747
primaria*	-.0340415	.0032	-10.65	0.000	-.040307	-.027776	.288639
secund~a*	-.0664512	.00354	-18.77	0.000	-.073389	-.059513	.286717
superior*	-.1156326	.00331	-34.95	0.000	-.122117	-.109148	.169238
propie~d*	-.0280886	.00357	-7.88	0.000	-.035079	-.021098	.357642
cocina*	-.0351878	.00874	-4.03	0.000	-.052313	-.018063	.024158
auto*	-.0840547	.00397	-21.17	0.000	-.091835	-.076275	.097165
camion*	-.062636	.01182	-5.30	0.000	-.085811	-.039461	.005926
habita~s	-.0214743	.00111	-19.35	0.000	-.02365	-.019299	3.24919
asocia~n*	.0029206	.00363	0.81	0.421	-.004189	.01003	.224147
personas	.0363605	.00093	39.15	0.000	.03454	.038181	3.53305
edad	-.0052022	.00057	-9.17	0.000	-.006314	-.00409	53.2665
edad2	.0000451	.00001	8.73	0.000	.000035	.000055	3081.62
lengua~a*	.0425715	.00347	12.28	0.000	.035774	.049369	.3097
rural*	.0538042	.00426	12.64	0.000	.045462	.062146	.398057
transf~b*	-.0681645	.00563	-12.11	0.000	-.079201	-.057128	.068122

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Figura 4.149. Cálculo de los efectos marginales MEM (1).

En la parte superior de la tabla en la figura 4.149. Se observa la probabilidad predicha tomando en cuenta el valor medio de todas las regresoras. Es equivalente a ejecutar el comando **prvalue** sin agregar opciones. En la tabla de la figura 4.149., se pueden apreciar los respectivos **cambios discretos** o **efectos marginales** para cada regresora en la primera columna, en las columnas siguientes se pueden ver los errores estándares de los efectos marginales, los respectivos estadísticos *Z* calculados, sus respectivos valores-p, los intervalos de confianza al 95% de los efectos marginales y la desviación estándar de la regresora. La utilización del comando **mfx** sin añadir ninguna opción, genera los efectos marginales que tienen los promedios de los regresores sobre la media condicional de la variable dependiente dadas las variables independientes. En otras palabras, los efectos marginales MEM que son equivalentes a la interpretación que le damos a los estimadores de los MRLC y dependiendo de la naturaleza de la regresora tiene una interpretación diferente. El comando **mfx** se puede complementar con el comando **prchange**.

```
. prchange

logit: Changes in Probabilities for niv_pobreza

              min->max      0->1      -+1/2     -+sd/2   MargEfct
        agua     0.0140    0.0140    0.0146    0.0055    0.0146
     desague    -0.0408   -0.0408   -0.0394   -0.0192   -0.0393
electricidad    -0.0172   -0.0172   -0.0162   -0.0045   -0.0162
    telefono    -0.0612   -0.0612   -0.0517   -0.0158   -0.0515
    primaria    -0.0340   -0.0340   -0.0364   -0.0164   -0.0363
  secundaria    -0.0665   -0.0665   -0.0763   -0.0342   -0.0754
    superior    -0.1156   -0.1156   -0.1911   -0.0680   -0.1797
   propiedad    -0.0281   -0.0281   -0.0291   -0.0139   -0.0290
      cocina    -0.0352   -0.0352   -0.0418   -0.0064   -0.0416
        auto    -0.0841   -0.0841   -0.1310   -0.0377   -0.1268
      camion    -0.0626   -0.0626   -0.0932   -0.0070   -0.0916
habitaciones    -0.1893   -0.0340   -0.0215   -0.0360   -0.0215
  asociacion     0.0029    0.0029    0.0029    0.0012    0.0029
    personas     0.9495    0.0130    0.0365    0.0700    0.0364
        edad    -0.4728   -0.0120   -0.0052   -0.0824   -0.0052
       edad2     0.7043    0.0000    0.0000    0.0794    0.0000
lengua_nat~a     0.0426    0.0426    0.0399    0.0184    0.0398
       rural     0.0538    0.0538    0.0515    0.0251    0.0512
transferen~b    -0.0682   -0.0682   -0.0979   -0.0242   -0.0961

         No_pobre      Pobre
Pr(y|x)     0.8962    0.1038

             agua       desague  electricidad      telefono      primaria    secundaria      superior
   x=      .831429       .605494       .917276       .894747       .288639       .286717       .169238
sd_x=      .374377       .488751       .275468       .306884       .453136       .452234       .374967

         propiedad        cocina          auto         camion  habitaciones    asociacion      personas
   x=      .357642       .024158       .097165       .005926       3.24919       .224147       3.53305
sd_x=      .479312       .153541       .296186       .076753       1.67385       .417025       1.90704

              edad         edad2  lengua_nat~a         rural  transferen~b
   x=      53.2665       3081.62         .3097       .398057       .068122
sd_x=      15.6303       1736.41       .462376       .489504       .251959
```

Figura 4.150. Cálculo de los efectos marginales MEM (2).

En la parte superior de la figura 4.150. Encontramos cinco columnas, de las cuales las cuatro primeras corresponden al **cambio discreto**. **STATA define al cambio discreto como una diferencia en el valor predicho a medida que cambia una variable independiente, mientras las demás regresoras permanecen constantes**. De derecha a izquierda tenemos, el cambio discreto de una variable desde su valor mínimo al máximo "min→max", el cambio discreto de una variable desde 0 a 1 "0→1", el cambio discreto de una variable independiente en torno a los valores medios de dicha variable independiente "-+1/2" y el cambio discreto de una variable independiente en un incremento de una desviación estándar "-+sd/2". Por último, la columna "MargEfct" corresponde al efecto marginal de la variable regresora.

Utilicemos los resultados del comando **prchange** para la interpretación. Dependiendo si la regresora es cualitativa o cuantitativa la interpretación es distinta. Por ejemplo, tomemos el caso de la regresora ***superior*** cuyos valores son "0" cuando el jefe de hogar no tiene educación superior máxima y "1" cuando el jefe de hogar tiene educación superior máxima, debido a que es una variable dicotómica sus cambios discretos en las columnas "min→max" y "0→1" son iguales y sus interpretaciones son similares siendo, "si el jefe de hogar pasa de no tener educación superior máxima a tener educación superior máxima, la probabilidad que el hogar sea pobre se reduce en 0.116", por otro lado las columnas "-+1/2" y "-+sd/2" no son relevantes para las variables ficticias en palabras de (Escobar M., Fernández M., & Bernardi, 2012). En cuanto a la interpretación de su efecto marginal en la columna "MargEfct", podemos decir, "si el jefe de hogar tiene educación superior máxima, entonces la probabilidad que el hogar sea pobre se reduce en 0.1797".

Veamos ahora un ejemplo con una regresora cuantitativa tomando a la variable ***personas*** que indica el número de miembros en el hogar, el valor de la columna "min→max" se interpreta como "si el hogar pasa desde su valor mínimo hasta su valor máximo, la probabilidad que el hogar sea pobre aumenta en 0.9495", mientras el valor de la columna "0→1" no tiene sentido de interpretación, puesto que no existe hogar alguno que tenga un número de miembros 0. De igual forma sucede con cualquier variable cuantitativa, el valor de la columna "-+1/2" se puede interpretar como la estimación del efecto marginal, ya que es la tasa de cambio estimada en torno a los valores medios de la variable independiente, por ello es que es igual al valor "MargEftc" y podemos interpretarlo como "si el número de miembros aumenta en una persona entonces la

probabilidad que el hogar sea pobre aumenta en 0.0364. En cuanto al valor de la columna "-+sd/2" se interpreta de forma similar a la anterior columna, la única diferencia es que se utiliza la desviación estándar, lo que ocasiona que se estandarice la estimación del efecto marginal y se pueda comparar distintas tasas de cambio marginal de distintas variables regresoras con distintos rangos.

Otro tipo de efecto marginal es el MER y a diferencia del MEM utiliza valores predeterminados, previamente de las regresoras. Podríamos utilizar tanto los comandos **mfx** o **prchange** para el cálculo de los efectos marginales MER, apoyándonos de las opciones **at()** y **x()** respectivamente, pero introducir el comando de tal forma que los comandos nos otorguen los mismos resultados es tedioso. Por ejemplo, veamos el efecto marginal para un jefe de hogar con 20 años, teniendo acceso solo a los servicios básicos, con 2 miembros en el hogar en el área urbana.

```
. mfx, at(1 1 1 1 0 0 0 0 0 0 0 0 0 2 20 400 0 0 0)

Marginal effects after logit
      y  = Pr(niv_pobreza) (predict)
         =  .26622208
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
agua*	.0295511	.00763	3.87	0.000	.014596	.044507	1
desague*	-.0900646	.00953	-9.46	0.000	-.108734	-.071396	1
electr~d*	-.0354682	.01064	-3.33	0.001	-.056323	-.014613	1
telefono*	-.1206375	.01122	-10.76	0.000	-.142621	-.098654	1
primaria*	-.0689974	.00772	-8.94	0.000	-.084132	-.053863	0
secund~a*	-.1273022	.01018	-12.50	0.000	-.14726	-.107345	0
superior*	-.2163278	.01377	-15.71	0.000	-.243314	-.189341	0
propie~d*	-.0564419	.00734	-7.69	0.000	-.070825	-.042058	0
cocina*	-.0780183	.02103	-3.71	0.000	-.119245	-.036791	0
auto*	-.1813434	.01258	-14.41	0.000	-.206009	-.156677	0
camion*	-.1469303	.03154	-4.66	0.000	-.208757	-.085104	0
habita~s	-.0451132	.00313	-14.41	0.000	-.051248	-.038979	0
asocia~n*	.0061382	.00762	0.81	0.421	-.008798	.021074	0
personas	.0763859	.00316	24.16	0.000	.070189	.082583	2
edad	-.0109288	.00148	-7.38	0.000	-.013831	-.008027	20
edad2	.0000948	.00001	7.24	0.000	.000069	.00012	400
lengua~a*	.0913927	.00762	12.00	0.000	.07646	.106326	0
rural*	.1200157	.00977	12.28	0.000	.100859	.139173	0
transf~b*	-.1519007	.01522	-9.98	0.000	-.18174	-.122062	0

(*) dy/dx is for discrete change of dummy variable from 0 to 1

Figura 4.151. Cálculo de los efectos marginales MER para un jefe de hogar con 20 años, teniendo acceso solo a los servicios básicos, con 2 miembros en el hogar en el área urbana (1).

El comando **mfx** utiliza la opción **at()** para el cálculo del efecto marginal MER, colocando dentro del paréntesis el valor que le otorgamos a cada variable según el

requerimiento de su investigación, ubicando cada valor según el lugar que le corresponde a las variables implicadas. Veamos ahora la ejecución del comando **prchange**.

```
. prchange, x( agua==1 desague==1 electricidad==1 telefono==1 primaria==0 secundaria==0 superior==0 propiedad==0 cocina==0 auto==0 camion==0 habitaciones=
> =0 asociacion==0 personas==2 edad==20 edad2==400 lengua_nativa==0 rural==0 transferencias_jub==0 )

    logit: Changes in Probabilities for niv_pobreza

                  min->max       0->1      -+1/2     -+sd/2   MargEfct
           agua     0.0296     0.0296     0.0307     0.0115     0.0307
        desague    -0.0901    -0.0901    -0.0824    -0.0403    -0.0825
    electricidad   -0.0355    -0.0355    -0.0341    -0.0094    -0.0341
        telefono   -0.1206    -0.1206    -0.1079    -0.0332    -0.1081
        primaria   -0.0690    -0.0690    -0.0761    -0.0345    -0.0762
      secundaria   -0.1273    -0.1273    -0.1576    -0.0715    -0.1583
        superior   -0.2163    -0.2163    -0.3668    -0.1410    -0.3776
       propiedad   -0.0564    -0.0564    -0.0610    -0.0292    -0.0610
          cocina   -0.0780    -0.0780    -0.0874    -0.0134    -0.0875
            auto   -0.1813    -0.1813    -0.2628    -0.0788    -0.2664
          camion   -0.1469    -0.1469    -0.1911    -0.0148    -0.1925
    habitaciones   -0.2550    -0.0426    -0.0451    -0.0754    -0.0451
      asociacion    0.0061     0.0061     0.0061     0.0025     0.0061
        personas    0.8013     0.0547     0.0763     0.1451     0.0764
            edad   -0.3076    -0.0140    -0.0109    -0.1699    -0.0109
           edad2    0.7165     0.0001     0.0001     0.1638     0.0001
    lengua_nat~a    0.0914     0.0914     0.0835     0.0387     0.0836
           rural    0.1200     0.1200     0.1073     0.0526     0.1076
    transferen~b   -0.1519    -0.1519    -0.2003    -0.0508    -0.2019

              No_pobre      Pobre
    Pr(y|x)     0.7338     0.2662

         agua     desague  electricidad     telefono     primaria   secundaria     superior    propiedad       cocina         auto
x=          1           1             1            1            0            0            0            0            0            0
sd_x=  .374377    .488751      .275468      .306884     .453136      .452234      .374967      .479312      .153541      .296186

        camion  habitaciones   asociacion     personas         edad        edad2  lengua_nat~a        rural  transferen~b
x=           0             0            0            2           20          400             0            0             0
sd_x=  .076753       1.67385      .417025      1.90704      15.6303      1736.41      .462376      .489504      .251959
```

Figura 4.152. Cálculo de los efectos marginales MER para un jefe de hogar con 20 años, teniendo acceso solo a los servicios básicos, con 2 miembros en el hogar en el área urbana (2).

Como se observa, ambos comandos nos brindan los mismos resultados sobre los cambios discretos y los efectos marginales de las variables regresoras según la especificación determinada en la opción del comando, en el caso del comando **prchange** se ha utilizado su opción **x()** para indicar el requerimiento. La interpretación de las variables regresoras cuantitativas y cualitativas son las mismas a los efectos marginales MEM. Por ejemplo la interpretación del efecto marginal de la variable ***superior*** es, "si un jefe de hogar tiene 20 años, acceso solo a los servicios básicos, tiene 2 miembros en el hogar, se ubica en el área urbana y tiene educación superior máxima, entonces la

probabilidad que su hogar sea pobre se reduce 0.3736", mientras tanto la interpretación del cambio discreto "0→1" es, "si un jefe de hogar tiene 20 años, acceso solo a los servicios básicos, además tiene 2 miembros en el hogar, se ubica en el área urbana y pasa de no tener una educación superior máxima a tener educación superior máxima entonces la probabilidad que el hogar sea pobre se reduce en 0.2163".

Tomemos la variable ***edad*** e interpretemos su efecto marginal, "si un jefe de hogar tiene 20 años, acceso solo a los servicios básicos, tiene 2 miembros en el hogar, se ubica en el área urbana aumenta en un año su edad, entonces la probabilidad que el hogar sea pobre se reduce en 0.0764" y su cambio discreto "min→max" se interpreta como "si un jefe de hogar tiene 20 años, acceso solo a los servicios básicos, tiene 2 miembros en el hogar, se ubica en el área urbana y llega a su máxima edad, entonces la probabilidad que su hogar sea pobre se reduce en 0.3076". En ocasiones la especificación no es tan detallada, por ejemplo, (Aparicio, Jaramillo, & San Román , 2011) Recomiendan el cálculo de los efectos marginales, cuando el jefe de hogar tiene sexo masculino y femenino y cuando el hogar se encuentra en una zona urbana y rural. Utilicemos la condicional **if** y la opción **x()** en el comando **prchange**.

```
. prchange if sexo==1

logit: Changes in Probabilities for niv_pobreza

                min->max       0->1      -+1/2     -+sd/2   MargEfct
      agua        0.0145     0.0145     0.0151     0.0058     0.0151
   desague       -0.0419    -0.0419    -0.0408    -0.0201    -0.0406
electricidad    -0.0178    -0.0178    -0.0168    -0.0048    -0.0168
  telefono      -0.0634    -0.0634    -0.0535    -0.0155    -0.0533
  primaria      -0.0355    -0.0355    -0.0376    -0.0173    -0.0375
secundaria      -0.0699    -0.0699    -0.0789    -0.0361    -0.0780
  superior      -0.1205    -0.1205    -0.1972    -0.0709    -0.1860
 propiedad      -0.0290    -0.0290    -0.0301    -0.0143    -0.0301
    cocina      -0.0365    -0.0365    -0.0433    -0.0069    -0.0431
      auto      -0.0890    -0.0890    -0.1354    -0.0420    -0.1313
    camion      -0.0651    -0.0651    -0.0964    -0.0082    -0.0948
habitaciones    -0.1984    -0.0353    -0.0222    -0.0378    -0.0222
asociacion       0.0030     0.0030     0.0030     0.0013     0.0030
  personas       0.9511     0.0123     0.0377     0.0722     0.0376
      edad      -0.4568    -0.0120    -0.0054    -0.0833    -0.0054
     edad2       0.7194     0.0000     0.0000     0.0793     0.0000
lengua_nat~a     0.0440     0.0440     0.0413     0.0191     0.0412
     rural       0.0549     0.0549     0.0533     0.0263     0.0530
transferen~b    -0.0710    -0.0710    -0.1013    -0.0258    -0.0995

        No_pobre    Pobre
Pr(y|x)   0.8921   0.1079

         agua    desague  electricidad    telefono    primaria   secundaria    superior   propiedad      cocina        auto
x=   .820839   .581311      .909393    .906444     .30837     .309453    .172105    .342119    .026618    .114761
sd_x=   .383494   .493353      .287055    .291216     .461829     .462277    .377479    .474428    .160968     .31874

         camion  habitaciones   asociacion     personas         edad        edad2  lengua_nat~a       rural  transferen~b
x=   .007616      3.31856      .236168      3.77869      52.3383      2972.79      .313522      .429329       .07265
sd_x=   .086938       1.6956      .424735       1.90089      15.2807       1679.1      .463933       .49499      .259566
```

Figura 4.153. Cálculo de los efectos marginales para un jefe de hogar masculino.

```
. prchange if sexo==2

logit: Changes in Probabilities for niv_pobreza

                min->max       0->1      -+1/2     -+sd/2   MargEfct
      agua        0.0128     0.0128     0.0134     0.0047     0.0134
   desague       -0.0382    -0.0382    -0.0361    -0.0170    -0.0360
electricidad     -0.0158    -0.0158    -0.0149    -0.0036    -0.0149
  telefono       -0.0556    -0.0556    -0.0474    -0.0161    -0.0471
  primaria       -0.0307    -0.0307    -0.0333    -0.0142    -0.0332
secundaria       -0.0585    -0.0585    -0.0699    -0.0291    -0.0690
  superior       -0.1042    -0.1042    -0.1763    -0.0613    -0.1646
 propiedad       -0.0260    -0.0260    -0.0267    -0.0130    -0.0266
    cocina       -0.0320    -0.0320    -0.0383    -0.0051    -0.0381
      auto       -0.0726    -0.0726    -0.1204    -0.0261    -0.1162
    camion       -0.0568    -0.0568    -0.0856    -0.0034    -0.0839
habitaciones     -0.1677    -0.0308    -0.0197    -0.0317    -0.0197
asociacion        0.0027     0.0027     0.0027     0.0011     0.0027
  personas        0.9275     0.0146     0.0334     0.0598     0.0333
      edad       -0.4778    -0.0119    -0.0048    -0.0787    -0.0048
     edad2        0.6604     0.0000     0.0000     0.0775     0.0000
lengua_nat~a      0.0392     0.0392     0.0366     0.0167     0.0365
     rural        0.0511     0.0511     0.0472     0.0219     0.0469
transferen~b     -0.0615    -0.0615    -0.0899    -0.0204    -0.0880
        No_pobre    Pobre
Pr(y|x)   0.9060   0.0940

            agua     desague  electricidad     telefono     primaria   secundaria     superior   propiedad       cocina         auto
x=       .857999      .666167      .937055      .865399      .239135      .229674      .162046      .39659      .017984      .053016
sd_x=    .349067      .471603      .242875      .341313      .426575      .420643       .36851      .489213       .1329      .224076

            camion  habitaciones   asociacion     personas         edad        edad2  lengua_nat~a        rural  transferen~b
x=       .001686       3.07512      .193987       2.91673      55.5953      3354.67      .300112      .319595      .056763
sd_x=    .041029        1.6049      .395437       1.77896      16.2437      1844.55      .458328      .466342        .2314
```

Figura 4.154. Cálculo de los efectos marginales para un jefe de hogar femenino.

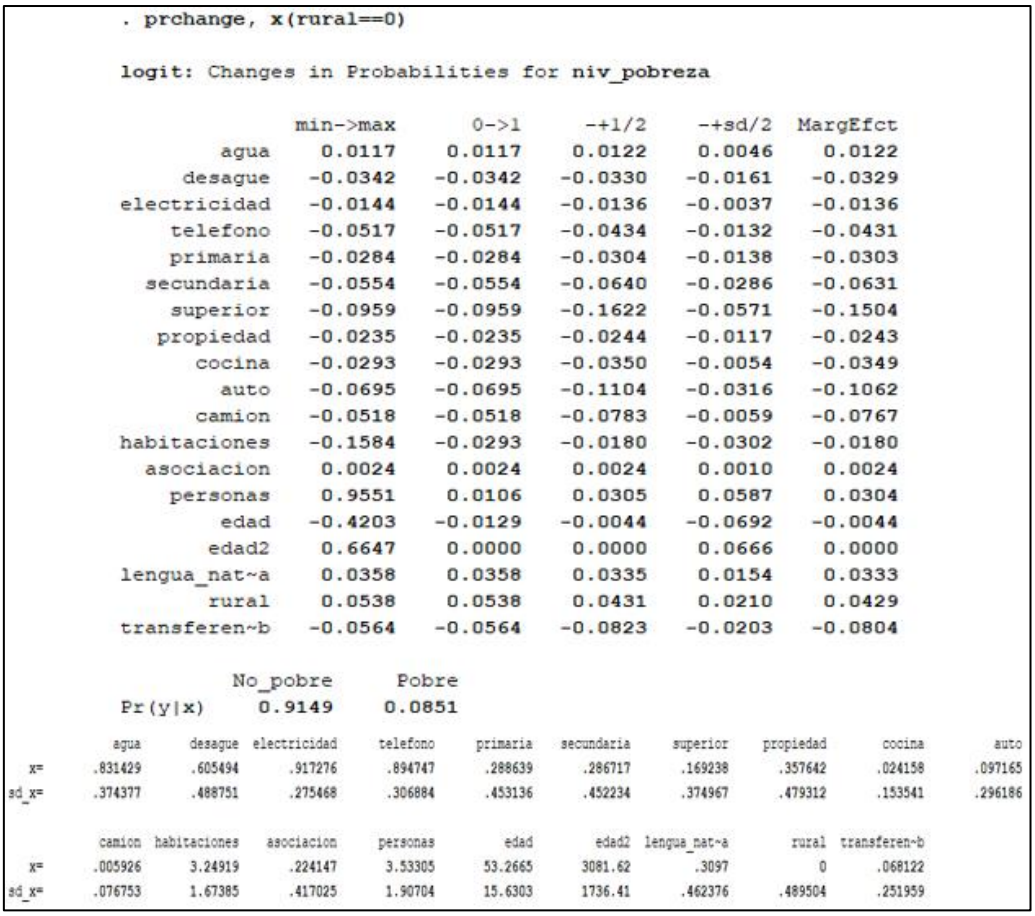
Los efectos marginales del modelo estimado según el sexo del jefe de hogar no podrían ser considerados como MER, debido a que no se está usando la opción **x()** para indicar el uso de un valor predeterminado, sino la media de las regresora. Podemos comprobarlo comparando los resultados en la parte inferior donde se aprecian las medidas "x" y desviaciones típicas "sd_x" con el comando **sum**.

```
. sum agua if sexo==2

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        agua |     10,676    .8579993    .3490673         0          1
```

Figura 4.155. Cuadro de estadísticos descriptivos de la variable ***agua***.

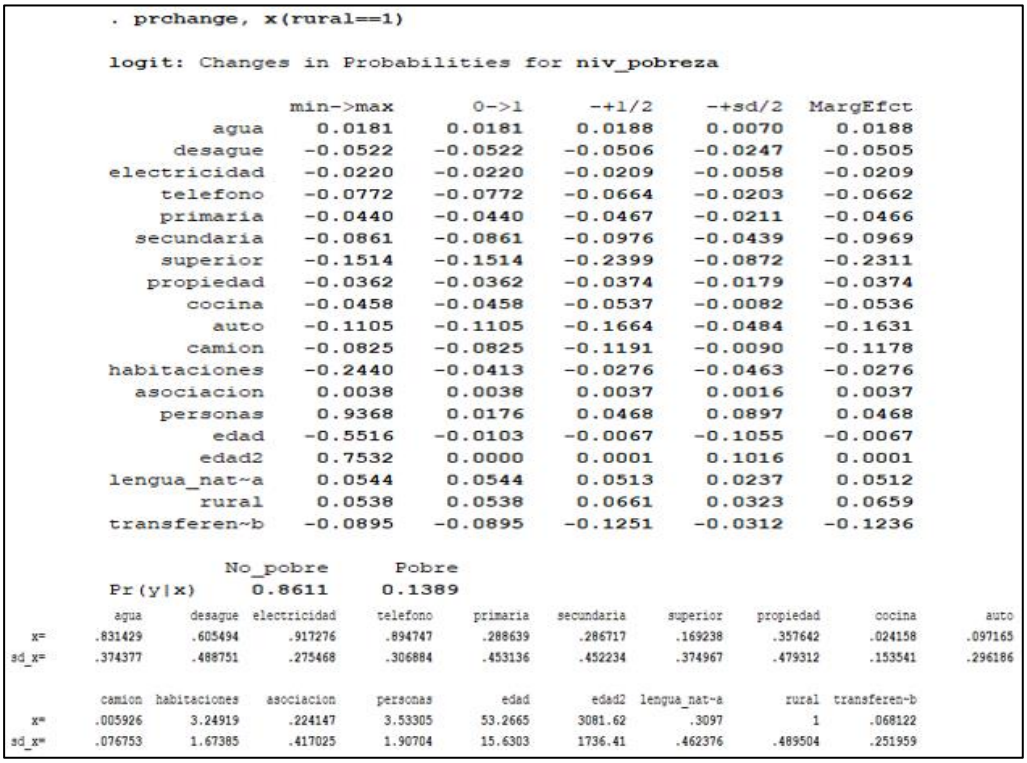
En la columna "Mean" se puede ver el promedio y en la columna "Std. Dev" la desviación estándar de la variable ***agua*** los cuales son los mismos en la figura 4.154.

Veamos los efectos marginales para los hogares según la ubicación de su hogar (área urbana y rural), la cual utilizaremos en la opción **x()** para detallar el requerimiento usando la variable ***rural***. Estos efectos marginales si pueden ser considerados como efectos marginales MER.



```
. prchange, x(rural==0)

logit: Changes in Probabilities for niv_pobreza

                min->max       0->1      -+1/2     -+sd/2   MargEfct
agua              0.0117     0.0117     0.0122     0.0046     0.0122
desague          -0.0342    -0.0342    -0.0330    -0.0161    -0.0329
electricidad     -0.0144    -0.0144    -0.0136    -0.0037    -0.0136
telefono         -0.0517    -0.0517    -0.0434    -0.0132    -0.0431
primaria         -0.0284    -0.0284    -0.0304    -0.0138    -0.0303
secundaria       -0.0554    -0.0554    -0.0640    -0.0286    -0.0631
superior         -0.0959    -0.0959    -0.1622    -0.0571    -0.1504
propiedad        -0.0235    -0.0235    -0.0244    -0.0117    -0.0243
cocina           -0.0293    -0.0293    -0.0350    -0.0054    -0.0349
auto             -0.0695    -0.0695    -0.1104    -0.0316    -0.1062
camion           -0.0518    -0.0518    -0.0783    -0.0059    -0.0767
habitaciones     -0.1584    -0.0293    -0.0180    -0.0302    -0.0180
asociacion        0.0024     0.0024     0.0024     0.0010     0.0024
personas          0.9551     0.0106     0.0305     0.0587     0.0304
edad             -0.4203    -0.0129    -0.0044    -0.0692    -0.0044
edad2             0.6647     0.0000     0.0000     0.0666     0.0000
lengua_nat~a      0.0358     0.0358     0.0335     0.0154     0.0333
rural             0.0538     0.0538     0.0431     0.0210     0.0429
transferen~b     -0.0564    -0.0564    -0.0823    -0.0203    -0.0804

            No_pobre      Pobre
Pr(y|x)       0.9149     0.0851

         agua      desague  electricidad  telefono  primaria  secundaria  superior  propiedad  cocina    auto
x=    .831429     .605494      .917276  .894747   .288639    .286717   .169238   .357642  .024158  .097165
sd_x= .374377     .488751      .275468  .306884   .453136    .452234   .374967   .479312  .153541  .296186

         camion  habitaciones  asociacion  personas     edad     edad2  lengua_nat~a     rural  transferen~b
x=    .005926       3.24919    .224147   3.53305  53.2665   3081.62        .3097         0      .068122
sd_x= .076753       1.67385    .417025   1.90704  15.6303   1736.41        .462376    .489504   .251959
```

Figura 4.156. Cálculo de los efectos marginales para hogar que se encuentra que se encuentra en un área urbana.



```
. prchange, x(rural==1)

logit: Changes in Probabilities for niv_pobreza

                min->max       0->1      -+1/2     -+sd/2   MargEfct
agua              0.0181     0.0181     0.0188     0.0070     0.0188
desague          -0.0522    -0.0522    -0.0506    -0.0247    -0.0505
electricidad     -0.0220    -0.0220    -0.0209    -0.0058    -0.0209
telefono         -0.0772    -0.0772    -0.0664    -0.0203    -0.0662
primaria         -0.0440    -0.0440    -0.0467    -0.0211    -0.0466
secundaria       -0.0861    -0.0861    -0.0976    -0.0439    -0.0969
superior         -0.1514    -0.1514    -0.2399    -0.0872    -0.2311
propiedad        -0.0362    -0.0362    -0.0374    -0.0179    -0.0374
cocina           -0.0458    -0.0458    -0.0537    -0.0082    -0.0536
auto             -0.1105    -0.1105    -0.1664    -0.0484    -0.1631
camion           -0.0825    -0.0825    -0.1191    -0.0090    -0.1178
habitaciones     -0.2440    -0.0413    -0.0276    -0.0463    -0.0276
asociacion        0.0038     0.0038     0.0037     0.0016     0.0037
personas          0.9368     0.0176     0.0468     0.0897     0.0468
edad             -0.5516    -0.0103    -0.0067    -0.1055    -0.0067
edad2             0.7532     0.0000     0.0001     0.1016     0.0001
lengua_nat~a      0.0544     0.0544     0.0513     0.0237     0.0512
rural             0.0538     0.0538     0.0661     0.0323     0.0659
transferen~b     -0.0895    -0.0895    -0.1251    -0.0312    -0.1236

            No_pobre      Pobre
Pr(y|x)       0.8611     0.1389

         agua      desague  electricidad  telefono  primaria  secundaria  superior  propiedad  cocina    auto
x=    .831429     .605494      .917276  .894747   .288639    .286717   .169238   .357642  .024158  .097165
sd_x= .374377     .488751      .275468  .306884   .453136    .452234   .374967   .479312  .153541  .296186

         camion  habitaciones  asociacion  personas     edad     edad2  lengua_nat~a     rural  transferen~b
x=    .005926       3.24919    .224147   3.53305  53.2665   3081.62        .3097         1      .068122
sd_x= .076753       1.67385    .417025   1.90704  15.6303   1736.41        .462376    .489504   .251959
```

Figura 4.157. Cálculo de los efectos marginales para hogar que se encuentra que se encuentra en un área rural.

Los resultados en las figuras podemos resumirlas en la siguiente tabla en forma de porcentaje.

Variable	Toda la muestra		Urbano		Rural		Hombre		Mujer	
Infraestructura	E.M.	%	E.M.	%	E.M.	%	E.M.	%	E.M.	%
Agua Potable	0.0146	1.46	0.0122	1.22	0.0188	1.88	0.0151	1.51	0.0134	1.34
Desagüe	-0.0393	-3.93	-0.0329	-3.29	-0.0505	-5.05	-0.0406	-4.06	-0.036	-3.6
Electricidad	-0.0162	-1.62	-0.0136	-1.36	-0.0209	-2.09	-0.0168	-1.68	-0.0149	-1.49
Teléfono	-0.0515	-5.15	-0.0431	-4.31	-0.0662	-6.62	-0.0533	-5.33	-0.0471	-4.71
Capital Humano										
Primaria completa	-0.0363	-3.63	-0.0303	-3.03	-0.0466	-4.66	-0.0375	-3.75	-0.0332	-3.32
Secundaria completa	-0.0754	-7.54	-0.0631	-6.31	-0.0969	-9.69	-0.078	-7.8	-0.069	-6.9
Superior completa	-0.1797	-17.97	-0.1504	-15.04	-0.2311	-23.11	-0.186	-18.6	-0.1646	-16.46
Capital Física										
Título de propiedad	-0.029	-2.9	-0.0243	-2.43	-0.0374	-3.74	-0.0301	-3.01	-0.0266	-2.66
Cocina	-0.0416	-4.16	-0.0349	-3.49	-0.0536	-5.36	-0.0431	-4.31	-0.0381	-3.81
Auto	-0.1268	-12.68	-0.1062	-10.62	-0.1631	-16.31	-0.1313	-13.13	-0.1162	-11.62
Camión	-0.0916	-9.16	-0.0767	-7.67	-0.1178	-11.78	-0.0948	-9.48	-0.0839	-8.39
Habitaciones	-0.0215	-2.15	-0.018	-1.8	-0.0276	-2.76	-0.0222	-2.22	-0.0197	-1.97
Capital Social										
Asociaciones	0.0029	0.29	0.0024	0.24	0.0037	0.37	0.003	0.3	0.0027	0.27
Características del hogar o del jefe de hogar										
Miembros	0.0364	3.64	0.0304	3.04	0.0468	4.68	0.0376	3.76	0.0333	3.33
Edad	-0.0052	-0.52	-0.0044	-0.44	-0.0067	-0.67	-0.0054	-0.54	-0.0048	-0.48
$Edad^2$	0	0	0	0	0.0001	0.01	0	0	0	0
Lengua indígena	0.0398	3.98	0.0333	3.33	0.0512	5.12	0.0412	4.12	0.0365	3.65
Rural	0.0512	5.12	0.0429	4.29	0.0659	6.59	0.053	5.3	0.0469	4.69
Transferencias										
Transf. Jubilación	-0.0961	-9.61	-0.0804	-8.04	-0.1236	-12.36	-0.0995	-9.95	-0.088	-8.8
Probabilidad de la variable dependiente										
Probabilidad de ocurrencia Pr $(Y = 1)$	0.1038		0.0851		0.1389		0.1079		0.0940	
Probabilidad de no ocurrencia Pr $(Y = 0)$	0.8962		0.9149		0.8611		0.8921		0.9060	

Tabla 4.3. Efectos Marginales sobre la probabilidad que el hogar sea pobre para un modelo Logit estimado usando la muestra completa.

Cuando estamos trabajando con variables cuantitativas, podemos analizar su comportamiento con respecto a las probabilidades de ocurrencia o no ocurrencia de la variable dependiente mediante gráficas. El comando **prgen** computa los valores predichos y los intervalos de confianza para un modelo. La instrucción es la siguiente.

```
. prgen personas, from(1) to(21) gen(personas1) ci
```

Figura 4.158. Ejecucion del comando **prgen**.

Hemos elegido a la variable ***personas*** que indica el número de miembros que hay en un hogar, como la variable cuantitativa para el siguiente ejemplo, en la opción **from()** indicamos el mínimo valor de la variable, en la opción **to()** señalamos el máximo valor de la variable, la opción **gen()** creará 3 variables nuevas con el nombre "pesonas1" seguido de sufijos que analizaremos posteriormente y la opción **ci** generará los intervalos de confianza. Sus resultados son los siguientes.

```
logit: Predicted values as personas varies from 1 to 21.

           agua       desague  electricidad      telefono      primaria    secundaria      superior     propiedad        cocina
x=     .83142918     .60549357     .91727617     .89474668     .28863915     .28671721     .16923816     .35764241     .02415781

           auto        camion  habitaciones     asociacion       personas          edad         edad2  lengua_nat~a         rural
x=     .09716513     .00592601     3.2491858     .22414714     3.5330468     53.26651     3081.6204     .3097005     .3980567

    transferen~b
x=     .06812236
```

Figura 4.159. Resultados del comando **prgen**.

Los resultados corresponden a las medias de las variables regresoras. En realidad, lo importante se encuentra en la creación de estas variables.

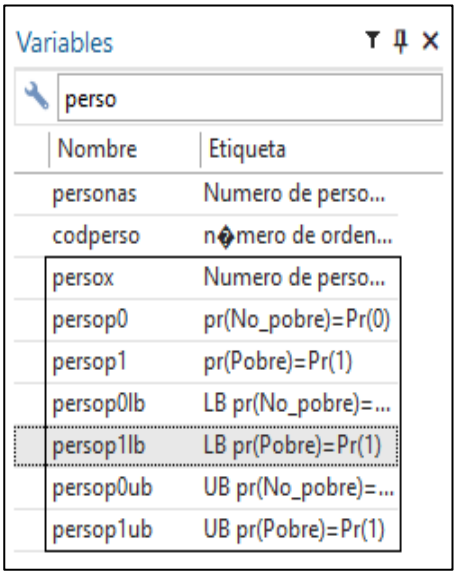


Figura 4.160. Variables creadas por el comando **prgen**.

La variable con terminación "***x***" representa a los valores de la variable ***personas*** en intervalos de cantidades iguales, la variable con terminación "***p0***" significa las probabilidades que el hogar no sea pobre, la variable con terminación "***p1***" significa las probabilidad que el hogar sea pobre, las variables ***perso0lb*** y ***persop0ub*** son los intervalos de confianza (inferior y superior, respectivamente) para la probabilidad que el hogar no sea pobre, y ***perso1lb*** y ***persop1ub*** son los intervalos de confianza (inferior y superior, respectivamente) para la probabilidad que el hogar es pobre.

El comando **graph** con sus componentes **twoway** y **rarea** nos ayudarán a graficar.

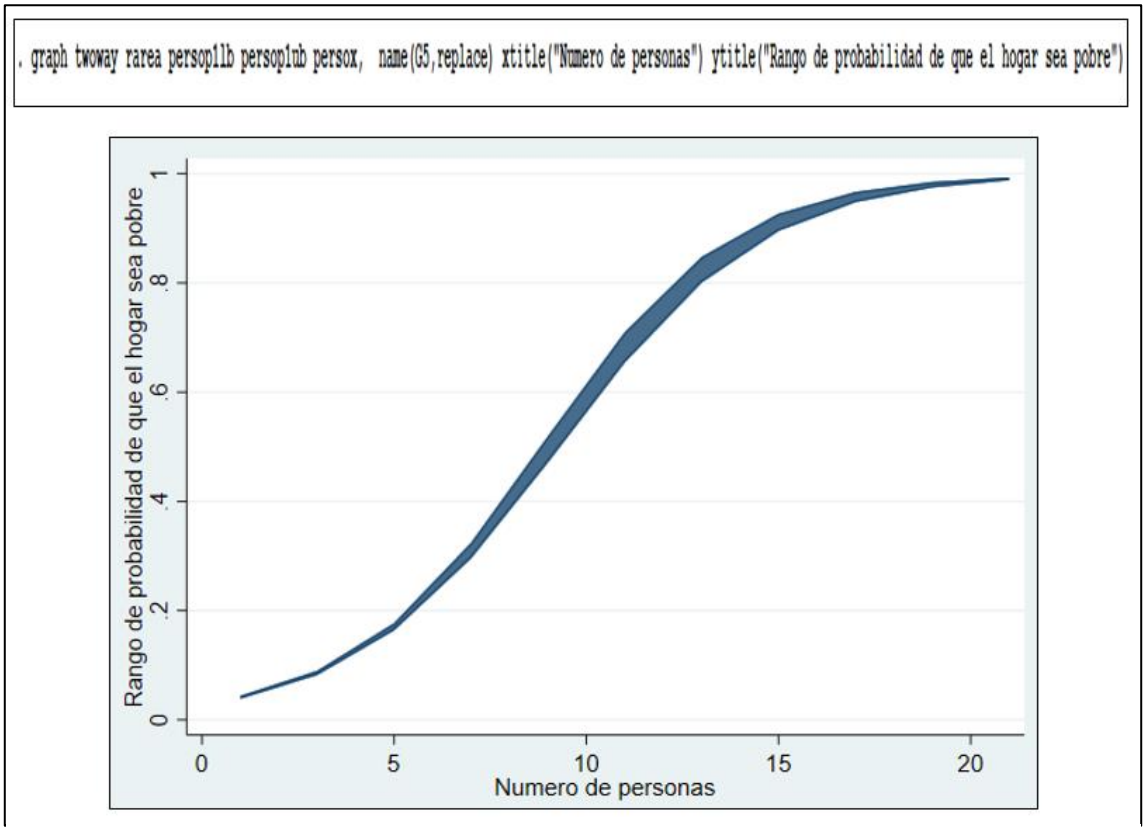


Figura 4.161. Gráfico de probabilidades predichas para distintos valores de la variable ***personas***.

Se puede apreciar como varía la probabilidad que el hogar sea pobre a medida que aumentan los miembros en un hogar, manteniendo constantes las demás variables en su media. El mismo efecto se puede ver cuantificado en los comandos **mfx** y **prchange**. La gráfica indica que a medida que el número de miembros pasa de ser aproximadamente 5 a 15, la probabilidad que el hogar sea pobre aumenta más rápido que en los extremos, la línea de arriba del área que muestra la gráfica es el intervalo de confianza y la línea inferior corresponde al intervalo de confianza.

Finalmente, en los modelos Logit podemos estimar el valor de las elasticidades, los cuales miden el cambio porcentual de la probabilidad de ocurrencia de la variable dependiente en proporción al aumento del 1% de las variables regresoras. El comando para realizar tal estimación de las elasticidades es **margins** y su opción **eyex(_all)**.

El comando **margins** es un comando frecuente en la postestimación, no solo de modelos Logit y Probit, sino también de otros modelos como Poisson, truncados, logistic, etc. Merecería toda una guía sobre su uso, ya que permite calcular elasticidades, semielasticidades y en algunos modelos efectos marginales. Al igual que la estimación de los efectos marginales, STATA computa las elasticidades de las variables regresoras tomando sus respectivos valores medios.

```
. margins, eyex(_all)

Average marginal effects                        Number of obs     =     37,462
Model VCE    : OIM

Expression   : Pr(niv_pobreza), predict()
ey/ex w.r.t. : agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion habitaciones asociacion
               personas edad edad2 lengua_nativa rural transferencias_jub
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.1105142	.0290647	3.80	0.000	.0535485	.1674799
desague	-.2313937	.0232701	-9.94	0.000	-.2770022	-.1857853
electricidad	-.1346777	.0388059	-3.47	0.001	-.2107359	-.0586196
telefono	-.416758	.0358128	-11.64	0.000	-.4869499	-.3465661
primaria	-.0872706	.0089002	-9.81	0.000	-.1047146	-.0698265
secundaria	-.2050157	.0124116	-16.52	0.000	-.2293421	-.1806894
superior	-.3194661	.0159479	-20.03	0.000	-.3507235	-.2882087
propiedad	-.1017208	.0135699	-7.50	0.000	-.1283173	-.0751244
cocina	-.0099702	.0030634	-3.25	0.001	-.0159744	-.0039661
auto	-.1288154	.0104754	-12.30	0.000	-.1493469	-.108284
camion	-.0054443	.0017106	-3.18	0.001	-.008797	-.0020915
habitaciones	-.6373315	.0332516	-19.17	0.000	-.7025034	-.5721597
asociacion	.0058948	.0072501	0.81	0.416	-.0083151	.0201046
personas	1.076284	.0237277	45.36	0.000	1.029778	1.122789
edad	-2.45903	.2679857	-9.18	0.000	-2.984273	-1.933788
edad2	1.238991	.1416103	8.75	0.000	.96144	1.516542
lengua_nativa	.0953386	.0068913	13.83	0.000	.0818319	.1088454
rural	.1511242	.0109027	13.86	0.000	.1297552	.1724932
transferencias_jub	-.0683882	.0088194	-7.75	0.000	-.085674	-.0511025

Figura 4.162. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre del modelo Logit para una muestra completa.

De derecha a izquierda vemos las siguientes columnas, “ey/ex” son las elasticidades calculadas de las variables tomando sus valores medios, “Std. Err.” Son los errores estándares, “Z” el estadístico calculado Z, “P>|Z|” es el valor-p que determina la significancia al 5% y “[95% Conf. Interval]” los intervalos de confianza. Tomaremos a la variable **personas** para su interpretación, “si el número de miembros en un hogar

aumenta en un 1%, entonces la probabilidad de que el hogar sea pobre aumenta en 1.07%". (Aparicio, Jaramillo, & San Román , 2011) Sugieren el cálculo de las elasticidades para el modelo usando muestras según el sexo del jefe de hogar y el área de residencia del hogar. Usamos el condicional **if** para el requerimiento de los autores.

```
. margins if sexo==1, eyex(_all)
Average marginal effects                        Number of obs     =     26,786
Model VCE    : OIM

Expression   : Pr(niv_pobreza), predict()
ey/ex w.r.t. : agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion habitaciones asociacion
               personas edad edad2 lengua_nativa rural transferencias_jub
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.108317	.0284796	3.80	0.000	.052498	.164136
desague	-.2220021	.0223313	-9.94	0.000	-.2657706	-.1782336
electricidad	-.1325481	.0381948	-3.47	0.001	-.2074085	-.0576877
telefono	-.4176861	.0358828	-11.64	0.000	-.4880152	-.3473571
primaria	-.0908699	.0092947	-9.78	0.000	-.1090872	-.0726526
secundaria	-.2191972	.01331	-16.47	0.000	-.2452843	-.1931101
superior	-.324375	.0162281	-19.99	0.000	-.3561814	-.2925686
propiedad	-.0972643	.0129765	-7.50	0.000	-.1226979	-.0718307
cocina	-.0110632	.003392	-3.26	0.001	-.0177115	-.004415
auto	-.1520781	.0123735	-12.29	0.000	-.1763297	-.1278265
camion	-.0069863	.0021965	-3.18	0.001	-.0112912	-.0026813
habitaciones	-.6451053	.0337023	-19.14	0.000	-.7111606	-.57905
asociacion	.0061391	.0075495	0.81	0.416	-.0086577	.0209359
personas	1.135224	.0248177	45.74	0.000	1.086582	1.183866
edad	-2.397513	.261444	-9.17	0.000	-2.909934	-1.885092
edad2	1.189472	.1360902	8.74	0.000	.9227397	1.456203
lengua_nativa	.0957392	.0069107	13.85	0.000	.0821944	.109284
rural	.161626	.0116426	13.88	0.000	.1388068	.1844451

Figura 4.163. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre del modelo Logit para una muestra de los hogares con jefe de hogar con sexo masculino.

```
. margins if sexo==2, eyex(_all)
Average marginal effects                        Number of obs     =     10,676
Model VCE    : OIM

Expression   : Pr(niv_pobreza), predict()
ey/ex w.r.t. : agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion habitaciones asociacion
               personas edad edad2 lengua_nativa rural transferencias_jub
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.1160269	.0305329	3.80	0.000	.0561835	.1758703
desague	-.2549572	.0256261	-9.95	0.000	-.3051835	-.204731
electricidad	-.1400209	.0403396	-3.47	0.001	-.219085	-.0609568
telefono	-.4144294	.0356405	-11.63	0.000	-.4842835	-.3445752
primaria	-.0782399	.0079115	-9.89	0.000	-.0937462	-.0627337
secundaria	-.1694346	.0101588	-16.68	0.000	-.1893455	-.1495236
superior	-.3071496	.0152453	-20.15	0.000	-.3370298	-.2772694
propiedad	-.1129022	.0150588	-7.50	0.000	-.1424169	-.0833876
cocina	-.0072279	.002239	-3.23	0.001	-.0116162	-.0028396
auto	-.0704495	.0057133	-12.33	0.000	-.0816473	-.0592517
camion	-.0015754	.0004916	-3.20	0.001	-.002539	-.0006118
habitaciones	-.6178271	.0321284	-19.23	0.000	-.6807977	-.5548565
asociacion	.0052817	.0064987	0.81	0.416	-.0074555	.0180189
personas	.9284037	.021035	44.14	0.000	.8871758	.9696316
edad	-2.613377	.2844217	-9.19	0.000	-3.170834	-2.055921
edad2	1.363235	.1554721	8.77	0.000	1.058516	1.667955
lengua_nativa	.0943337	.006848	13.78	0.000	.080912	.1077554
rural	.1247754	.0090548	13.78	0.000	.1070283	.1425225
transferencias_jub	-.057413	.0073613	-7.80	0.000	-.071841	-.0429851

Figura 4.164. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre del modelo Logit para una muestra de los hogares con jefe de hogar con sexo femenino.

```
. margins if rural==0, eyex(_all)

Average marginal effects                        Number of obs     =     22,550
Model VCE    : OIM

Expression   : Pr(niv_pobreza), predict()
ey/ex w.r.t. : agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion habitaciones asociacion
               personas edad edad2 lengua_nativa rural transferencias_jub
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.1336538	.0352384	3.79	0.000	.0645878	.2027198
desague	-.3354192	.0335488	-10.00	0.000	-.4011737	-.2696646
electricidad	-.156774	.0450761	-3.48	0.001	-.2451216	-.0684264
telefono	-.4814369	.0411058	-11.71	0.000	-.5620027	-.4008711
primaria	-.0803259	.0080917	-9.93	0.000	-.0961853	-.0644665
secundaria	-.2653662	.0158605	-16.73	0.000	-.2964521	-.2342802
superior	-.4807874	.0238632	-20.15	0.000	-.5275584	-.4340165
propiedad	-.1437145	.0190832	-7.53	0.000	-.1811169	-.106312
cocina	-.0119641	.0036303	-3.30	0.001	-.0190793	-.0048489
auto	-.1810582	.0145703	-12.43	0.000	-.2096154	-.152501
camion	-.0053349	.0016446	-3.24	0.001	-.0085584	-.0021115
habitaciones	-.7418569	.038459	-19.29	0.000	-.8172351	-.6664786
asociacion	.0067006	.0082513	0.81	0.417	-.0094717	.0228729
personas	1.237479	.0277404	44.61	0.000	1.183108	1.291849
edad	-2.692322	.2933408	-9.18	0.000	-3.267259	-2.117384
edad2	1.343949	.1537644	8.74	0.000	1.042576	1.645322
lengua_nativa	.068926	.0050397	13.68	0.000	.0590484	.0788036
rural	0	(omitted)				
transferencias_jub	-.1002385	.0128635	-7.79	0.000	-.1254506	-.0750265

Figura 4.165. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre del modelo Logit para una muestra de los hogares ubicados en el área de residencia urbana.

```
. margins if rural==1, eyex(_all)

Average marginal effects                        Number of obs     =     14,912
Model VCE    : OIM

Expression   : Pr(niv_pobreza), predict()
ey/ex w.r.t. : agua desague electricidad telefono primaria secundaria superior propiedad cocina auto camion
               habitaciones asociacion personas edad edad2 lengua_nativa rural transferencias_jub
```

	ey/ex	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
agua	.0755224	.0197316	3.83	0.000	.0368491	.1141956
desague	-.074086	.0077322	-9.58	0.000	-.0892409	-.0589311
electricidad	-.1012637	.0293272	-3.45	0.001	-.158744	-.0437833
telefono	-.3189503	.0278413	-11.46	0.000	-.3735182	-.2643823
primaria	-.0977723	.0101323	-9.65	0.000	-.1176312	-.0779134
secundaria	-.1137536	.0072131	-15.77	0.000	-.127891	-.0996161
superior	-.0755151	.0039799	-18.97	0.000	-.0833156	-.0677146
propiedad	-.0382179	.0052344	-7.30	0.000	-.0484771	-.0279586
cocina	-.0069551	.0022063	-3.15	0.002	-.0112793	-.0026308
auto	-.0498136	.0042839	-11.63	0.000	-.0582098	-.0414174
camion	-.0056096	.0018104	-3.10	0.002	-.0091579	-.0020613
habitaciones	-.4792678	.0254513	-18.83	0.000	-.5291514	-.4293842
asociacion	.0046762	.005736	0.82	0.415	-.0065661	.0159185
personas	.8325239	.0182179	45.70	0.000	.7968174	.8682303
edad	-2.106247	.2298248	-9.16	0.000	-2.556695	-1.655798
edad2	1.080273	.1233235	8.76	0.000	.8385635	1.321983
lengua_nativa	.13528	.00971	13.93	0.000	.1162487	.1543112
rural	.379655	.0273899	13.86	0.000	.3259717	.4333383
transferencias_jub	-.020224	.0027041	-7.48	0.000	-.025524	-.014924

Figura 4.166. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre del modelo Logit para una muestra de los hogares ubicados en el área de residencia rural.

Los resultados de las figuras se pueden ver en la siguiente tabla.

Variable	Toda la muestra	Urbano	Rural	Hombre	Mujer
Infraestructura					
Agua Potable	0.1105	0.1337	0.0755	0.1083	0.1160
Desagüe	-0.2314	-0.3354	-0.0741	-0.2220	-0.2550
Electricidad	-0.1347	-0.1568	-0.1013	-0.1325	-0.1400
Teléfono	-0.4168	-0.4814	-0.3190	-0.4177	-0.4144
Capital Humano					
Primaria completa	-0.0873	-0.0803	-0.0978	-0.0909	-0.0782
Secundaria completa	-0.2050	-0.2654	-0.1138	-0.2192	-0.1694
Superior completa	-0.3195	-0.4808	-0.0755	-0.3244	-0.3071
Capital Física					
Título de propiedad	-0.1017	-0.1437	-0.0382	-0.0973	-0.1129
Cocina	-0.0100	-0.0120	-0.0070	-0.0111	-0.0072
Auto	-0.1288	-0.1811	-0.0498	-0.1521	-0.0704
Camión	-0.0054	-0.0053	-0.0056	-0.0070	-0.0016
Habitaciones	-0.6373	-0.7419	-0.4793	-0.6451	-0.6178
Capital Social					
Asociaciones	0.0059	0.0067	0.0047	0.0061	0.0053
Características del hogar o del jefe de hogar					
Miembros	1.0763	1.2375	0.8325	1.1352	0.9284
Edad	-2.4590	-2.6923	-2.1062	-2.3975	-2.6134
$Edad^2$	1.2390	1.3439	1.0803	1.1895	1.3632
Lengua indígena	0.0953	0.0689	0.1353	0.0957	0.0943
Rural	0.1511	0.0000	0.3797	0.1616	0.1248
Transferencias					
Transf. Jubilación	-0.0684	-0.1002	-0.0202	-0.0728	-0.0574

Tabla 4.4. Elasticidades de las variables regresoras sobre la probabilidad que el hogar sea pobre para un modelo Logit estimado usando la muestra completa, según el área de residencia del hogar y el sexo del jefe de hogar.

De la información de las tablas se puede inferir lo siguiente.

- A comparación de los resultados que calcularon (Aparicio, Jaramillo, & San Román , 2011), en los resultados sobre la probabilidad de que el hogar sea pobre, está más influenciada negativamente por el capital humano en la muestra y en todas las submuestras. Sobre todo, la variable ***superior***, en las zonas rurales, los jefes de hogar que tienen acceso a una educación superior reducen más la probabilidad que en las zonas urbanas. De igual forma sucede, cuando el jefe de hogar con sexo femenino tiene educación superior con respecto al jefe de hogar con sexo masculino. Otra variable con resultados interesantes es la variable ***lengua_nativa***, indica que en todas las submuestras si el jefe de hogar habla una lengua nativa aumentan sus probabilidades de ser pobre, señalando que las personas con lengua nativa tienen más dificultad de salir de la pobreza que las personas que no hablan una lengua nativa.
- Entre las variables que conforman los servicios básicos de infraestructura, si un hogar cuenta con **agua** tiene más probabilidad que sea pobre. Se puede entender que en el 2018, los hogares necesitan más que solo recibir agua potable desde una red pública para salir de una pobreza sobre todo en áreas rurales. Por el contrario, si el hogar también cuenta con teléfono reduce las probabilidades que el hogar sea pobre.
- En conclusión, se deberían crear programas que sean capaces de impulsar la integración económica en las zonas rurales, con el fin de ayudar a las familias que se encuentran en zonas rurales alejadas y que hablan una lengua nativa, además de promover una mayor tasa de estudiantes universitarios con acuerdos sobre becas, subsidios, entre otros y brindar información a las familias sobre la planificación familiar, ya que se ha visto que un mayor número de miembros en un hogar puede aumentar las probabilidad que el hogar sea pobre.

ANEXO 1. BASE DE DATOS PARA EL EJEMPLO DE ESTIMACIÓN DE MCO Y VERIFICACIÓN DEL CUMPLIMIENTO DE SUPUESTOS PARA STATA CON DATOS DE CORTE TRANSVERSAL.

ANEXO 1.1. BASE DE DATOS PARA EL MODELO ECONOMÉTRICO ESPECIFICADO PARA LOS TRABAJADORES INDEPENDIENTES DEDICADOS A ACTIVIDADES PRODUCTIVAS/EXTRACTIVAS.

Año	**Ganancia total neta**	**Ingresos**	**Gastos**	**Número de trabajadores**
	e25t3	**e14t**	**gastos**	**e8a**
2018	1300	3500	2200	3
2018	1288	12124	10888	5
2018	298	500	212	1
2018	594	950	365	2
2018	120	240	120	2
2018	187	300	113	1
2018	1380	3200	1820	2
2018	210	350	140	1
2018	1900	4000	2100	1
2018	1070	2500	1430	1
2018	1058	4500	3442	4
2018	572	600	28	1
2018	1750	4000	2250	3
2018	48	91	43	1
2018	171	281	123	1
2018	25	50	25	1
2018	232	520	288	1
2018	186	433	251	1
2018	262	520	278	2
2018	5507	10000	5190	7
2018	1014	2165	1151	2
2018	515	1000	485	1
2018	821	700	446	1
2018	525	600	75	1
2018	2079	4000	1950	3
2018	2383	300	1508	2
2018	1200	1800	600	1

ANEXO 1.2. BASE DE DATOS PARA EL MODELO ECONOMÉTRICO ESPECIFICADO PARA LOS TRABAJADORES INDEPENDIENTES DEDICADOS A ACTIVIDADES COMERCIALES.

Año	Ganancia total neta	Ingresos	Gastos	Número de trabajadores
	e25t3	e17t	gastosc	e8a
2018	790	3100	2310	1
2018	2320	32000	29680	2
2018	1000	3000	2000	1
2018	371	650	366	1
2018	2427	433	422	1
2018	35	150	115	1
2018	590	800	510	1
2018	983	2165	1182	3
2018	110	1299	1189	1
2018	138	400	292	1
2018	114	420	306	1
2018	70	250	180	2
2018	29	178	149	1
2018	300	600	300	1
2018	90	450	372	1
2018	50	250	200	1
2018	1222	5196	3974	1
2018	1380	3600	2220	1
2018	330	3000	2670	1
2018	135	500	365	1
2018	1132	7274	6152	1
2018	866	1732	866	1
2018	227	300	193	1
2018	2766	13380	10917	1
2018	170	300	130	1
2018	20	48	36	1
2018	476	3248	2772	1
2018	596	2165	1579	1
2018	34	240	206	1
2018	78	450	372	1
2018	125	600	525	2
2018	790	1500	760	2
2018	439	1083	670	1
2018	768	1000	275	2
2018	305	1000	695	2
2018	757	2273	1516	1

2018	566	1584	1188	2
2018	128	720	612	1
2018	2005	4500	2995	3
2018	234	530	426	1
2018	554	2500	2076	1
2018	128	935	807	2
2018	2232	5413	4480	1
2018	2604	4978	3240	2
2018	2760	12000	9240	2
2018	460	1500	1160	1
2018	270	6500	6230	1
2018	907	900	393	2
2018	5507	500	2990	7
2018	563	600	340	2
2018	104	350	250	1
2018	494	1000	586	1
2018	185	500	430	1
2018	121	200	84	1
2018	821	1000	533	1
2018	630	1200	650	1
2018	986	2598	1872	1
2018	1299	6495	5196	1
2018	216	700	524	1
2018	65	200	158	1
2018	350	700	350	1
2018	885	1516	643	1
2018	202	350	198	1
2018	25	140	125	1
2018	2750	14200	12250	1
2018	104	217	130	1
2018	500	1490	1005	1
2018	313	1000	730	1
2018	873	4000	3127	1
2018	140	350	210	1
2018	2680	10000	7370	1
2018	2000	2500	500	1
2018	1040	2900	2060	1
2018	200	550	350	1
2018	230	680	500	1
2018	110	450	375	1
2018	307	400	93	2
2018	107	185	78	1
2018	24	30	15	1
2018	395	760	375	1
2018	560	1200	640	1

2018	4370	10000	5730	3
2018	368	800	432	1
2018	415	800	416	1
2018	485	866	598	1
2018	265	600	360	1

ANEXO 1.3. BASE DE DATOS PARA EL MODELO ECONOMÉTRICO ESPECIFICADO PARA LOS TRABAJADORES INDEPENDIENTES DEDICADOS A ACTIVIDADES PRESTADORAS DE SERVICIOS.

Año	Ganancia total neta	Ingresos	Gastos	Número de trabajadores
	e25t3	e20t	gastoss	e8a
2018	215	433	218	1
2018	1614	1800	706	1
2018	1050	4700	3650	2
2018	1020	3000	1980	3
2018	3015	4438	1423	1
2018	2427	3637	2370	1
2018	67	130	85	2
2018	735	1732	1127	1
2018	460	850	390	1
2018	1650	6500	4850	5
2018	915	2165	1250	1
2018	561	2250	1689	1
2018	20	40	30	2
2018	945	1386	441	1
2018	564	1083	584	1
2018	123	217	146	1
2018	2165	8660	6495	1
2018	89	217	128	1
2018	953	1559	606	1
2018	1037	2165	1128	1
2018	500	750	250	1
2018	100	200	100	1
2018	450	7500	7050	3
2018	1900	4000	2100	2
2018	500	700	200	1
2018	1981	4114	2289	1
2018	2230	2500	270	1
2018	1377	2425	1048	1
2018	1598	3031	1433	1
2018	2529	7794	5265	2
2018	1280	2425	1145	1
2018	1136	2078	1124	1
2018	1236	1819	583	1
2018	520	1000	480	1
2018	1498	3486	1988	1
2018	66	85	19	1
2018	7110	8000	890	1
2018	318	390	72	1

2018	2007	4260	2253	1
2018	952	2122	1170	1
2018	1859	3031	1172	1
2018	574	909	335	1
2018	1324	2598	1274	1
2018	836	1212	376	1
2018	800	1212	412	1
2018	2074	6365	4715	4
2018	493	779	286	1
2018	1022	3486	2464	1
2018	2483	3637	1154	1
2018	305	600	395	2
2018	1556	3200	1650	2
2018	160	550	410	1
2018	939	1700	761	1
2018	1354	3000	1646	1
2018	1739	2200	461	1
2018	2190	2500	310	1
2018	635	650	15	1
2018	96	104	8	1
2018	1848	2598	750	1
2018	212	220	8	1
2018	1204	2598	1481	2
2018	1050	1732	682	1
2018	736	1169	433	1
2018	1988	2382	394	1
2018	1239	1819	606	1
2018	1764	2576	899	1
2018	240	260	20	1
2018	546	909	363	1
2018	593	1039	446	1
2018	2383	2802	2050	2
2018	4700	8500	3800	1
2018	4302	8300	3998	2
2018	30	195	187	1
2018	967	1603	636	1
2018	2245	3897	1704	2
2018	220	460	240	1
2018	1835	2000	165	1
2018	27	217	190	1
2018	1000	2550	1550	1
2018	1647	2382	735	1
2018	1299	2166	867	1
2018	866	1754	888	1
2018	800	1516	868	1

2018	928	2446	1518	1
2018	409	1775	1366	1
2018	1785	3616	1831	1
2018	1319	2100	781	1

ANEXO 2. BASE DE DATOS PARA EL EJEMPLO DE ESTIMACIÓN DE MCO Y VERIFICACIÓN DEL CUMPLIMIENTO DE SUPUESTOS PARA STATA CON DATOS DE SERIES TEMPORALES.

trimestre	imp	pbi	ibi	indp_v	impdolar	impi	impbc	tc_v
1999q1	9323.02	51214.6	7137.42	2.48898	1539.79	626.623	509.229	3.37413
1999q2	9428.56	55517.8	10875.7	2.55289	1598.87	683.174	541.324	3.3334
1999q3	9901.05	53196.1	6663.8	2.61606	1686.38	795.304	505.114	3.3925
1999q4	10728.8	56448.2	12119.8	2.66242	1885.44	874.745	561.739	3.49453
2000q1	9846.93	54674.8	9819.15	2.69514	1775.27	817.107	560.867	3.4722
2000q2	10211.3	58255.6	12641.6	2.74593	1839.58	901.656	546.225	3.4888
2000q3	9962.78	54621.8	6880.76	2.71631	1785.22	915.229	476.624	3.484
2000q4	10872.4	54654.6	6693.34	2.71071	1957.5	976.558	530.26	3.52147
2001q1	10470.4	51760.4	7483.44	2.73284	1833.97	865.903	555.715	3.52553
2001q2	10205.6	58431.1	10850	2.73068	1758.05	881.003	462.995	3.57193
2001q3	10775.2	56119.6	6866.19	2.66987	1856.09	958.717	464.201	3.48263
2001q4	10621.7	57268.5	8380.49	2.51411	1756.37	845.567	438.366	3.4392
2002q1	9920.47	55137.7	7866.86	2.62455	1630.79	799.277	436.386	3.4638
2002q2	10721.5	62307.2	12169.9	2.62849	1847.38	966.206	434.224	3.46737
2002q3	11204	58404.4	6359.74	2.75618	1954.99	1044.76	451.224	3.60183
2002q4	11203.2	59923.6	9117.06	2.61952	1959.64	930.115	520.438	3.53887
2003q1	11192.9	58249.3	9743.64	2.72958	2029.11	1087.49	505.603	3.47663
2003q2	10833.5	65202.5	12932.7	2.64611	1969.84	1034.02	465.925	3.47507
2003q3	11391.2	60551.7	7455.09	2.68432	2074.42	1081.14	519.772	3.47828
2003q4	11427.3	61589.2	8304.02	2.63864	2131.48	1137.24	482.927	3.47137
2004q1	11230.3	60913.8	8733.56	2.77194	2118.24	1161.38	527.377	3.4769
2004q2	12483.4	67639.7	14999.6	2.78084	2417.59	1347.2	585.091	3.47927
2004q3	12516.2	63145.8	5663.86	2.66444	2540.49	1394.04	614.257	3.3705
2004q4	12901.4	66070.5	8629.07	2.53366	2728.45	1461.01	634.255	3.29933
2005q1	12525.2	64340.9	8101.19	2.60296	2659.87	1467.43	657.663	3.258
2005q2	13761.6	71310.4	14458.6	2.61206	3006.45	1665.19	748.276	3.25383
2005q3	13907.2	67229.8	6915.81	2.64307	3161.51	1732.6	805.215	3.29263
2005q4	14277.4	71090.1	9401.92	2.70702	3253.77	1734.68	852.392	3.40463
2006q1	14643.1	69670.8	13112.5	2.72163	3380.49	1840.25	931.302	3.3196
2006q2	14908.3	75823.9	17111.1	2.6986	3629.6	2008.82	972.603	3.2853
2006q3	14877.5	72806.3	10484.4	2.65589	3670.58	2008.8	997.746	3.24247
2006q4	17158.5	76296.9	13729	2.5785	4163.42	2123.56	1221.73	3.2091
2007q1	17337.9	73353.8	14604	2.65098	4208.21	2207.16	1290.83	3.18923
2007q2	17775.2	80625.6	19618	2.62919	4490.36	2364.45	1359.45	3.17033
2007q3	19732.1	80689.1	16996	2.63251	5288.84	2868.3	1572.17	3.13547
2007q4	19891.1	85024.5	19969.8	2.47074	5603.11	2988.63	1631.87	2.99617
2008q1	21196.5	80813.1	18409.5	2.39501	6266.42	3437.77	1820.94	2.85397

2008q2	23027.7	89146.4	24719.4	2.4493	7552.95	4027.43	2394.29	2.88077
2008q3	24401.6	88439.8	22643.5	2.46767	7977.5	4156.26	2600.57	2.91347
2008q4	24081.8	90523.6	25801.8	2.49902	6652.31	2934.89	2416.78	3.10637
2009q1	18960.3	82894.9	15858.2	2.56383	4883.42	2071.97	1848.42	3.19363
2009q2	18007.4	88427.2	17509.8	2.41597	4826.66	2301.56	1590.44	2.99617
2009q3	19211	88283	14852.6	2.36507	5330.21	2655.65	1634.69	2.9383
2009q4	21025	92978.9	21058.7	2.33459	5970.4	3047.28	1776.1	2.88977
2010q1	21731.1	87418.2	18664.1	2.30989	6335.81	3170.57	1970.42	2.84673
2010q2	22667.6	96887.3	25307.2	2.28196	6610.15	3257.1	2056.83	2.8386
2010q3	26472.6	96918.8	23245.5	2.24655	7815.26	3743.31	2523.52	2.80163
2010q4	26334.1	101156	28623.5	2.22565	8054.1	3852.51	2522.94	2.81167
2011q1	25067.2	94996.3	23927.6	2.2657	8197.83	4024.66	2651.75	2.782
2011q2	27381.1	102176	26884.9	2.29981	9606.95	4864.23	3087.31	2.7786
2011q3	27740.8	102606	24701.1	2.19188	9692.03	4702.19	3037	2.74497
2011q4	28106.7	107274	30587.8	2.15215	9654.71	4741.41	2953.61	2.70023
2012q1	27758.5	100669	22766.9	2.14995	9524.66	4542.29	3108.01	2.6767
2012q2	29299.2	107961	29112.3	2.14688	9973.87	4604.86	3408.86	2.672
2012q3	32469.3	109625	29456.7	2.08827	10990.5	5258.47	3531.04	2.6112
2012q4	31152.7	113019	31730.2	1.99254	10528.9	4867.58	3299.48	2.56877
2013q1	30483.6	105428	29747.3	2.03527	10394.8	4846.57	3338.59	2.5847
2013q2	31129.2	114690	32947.1	2.1528	10514.6	4762.74	3562.61	2.7189
2013q3	33080	115431	31532.5	2.19441	11129.8	5221.44	3523.49	2.79287
2013q4	31024.7	120900	31112.9	2.14522	10317	4697.09	3238.95	2.78883
2014q1	30419.7	110643	28912.7	2.16389	10185.5	4673.69	3172.91	2.8089
2014q2	30538.6	116939	30815	2.18934	10363.6	4687.12	3449.84	2.78843
2014q3	31692.2	117592	30717.5	2.21316	10583.4	5028.04	3211.4	2.84277
2014q4	31490	122202	31007.8	2.22053	9909.68	4408.44	3076.75	2.93977
2015q1	30869.7	112788	28873.8	2.31345	9253.65	3998.27	2946.92	3.0824
2015q2	31134	120660	30372.8	2.40585	9344.94	4104.66	3025.27	3.15327
2015q3	32161.1	121315	29222.9	2.40329	9420.17	4039.71	3002.16	3.21497
2015q4	32674.3	127913	28251	2.44956	9312.04	3767.89	3027.94	3.35653
2016q1	30254.8	117963	26328.3	2.55211	8381.12	3446	2777.2	3.43707
2016q2	29580.3	125339	27574.2	2.52173	8399.47	3598.45	2798.33	3.31123
2016q3	31795.6	127091	27779.4	2.49938	9107.14	3987.88	2746.89	3.3798
2016q4	32302	131832	29303.8	2.52007	9240.67	3989.98	2909.17	3.37537
2017q1	30281.2	120628	25215	2.48949	8991.75	4308.72	2551.36	3.26307
2017q2	30886.2	128584	26097.8	2.44312	9213.44	4232.62	2684.8	3.25403

2017q3	33154	130569	27479.8	2.42307	10020.8	4473.56	3009.89	3.24463
2017q4	34432.4	134874	30621.7	2.45522	10496.1	4887.22	3070.41	3.23843
2018q1	32691.7	124393	27108.9	2.45991	10038.6	4819.68	2817.06	3.23357
2018q2	32871.8	135729	29501.1	2.51929	10503.6	5209.03	2896.78	3.26453
2018q3	33457.9	133824	29014.3	2.53958	10761.5	5401.95	2908.45	3.2886
2018q4	33823.6	141136	31650.5	2.58729	10566.3	5084.95	3018.64	3.37357
2019q1	32499.3	127435	26628.1	2.53277	9969.05	4783.42	2829.51	3.31663
2019q2	33080.9	137352	29437.9	2.57301	10216.2	4823.88	3086.59	3.323
2019q3	34439.9	138165	30378.6	2.55426	10536.6	4802.08	3202.92	3.36207
2019q4	34442.8	143699	31176.1	2.54715	10352.1	4691.34	3176.3	3.3515

BIBLIOGRAFÍA

Acosta G., E., Andrada F. Julián, & Fernández M., E. (2009). *Especificación de modelos econométricos utilizanco minería de datos.* Las Palmas.

Adkins C., L., & Carter H., R. (2011). *Using STATA for Principles of Econometrics.* Danvers: Clearence Center Inc.

Aguarto P., H. (2010). *La Metodología De La Investigación Econometrica.* Obtenido de WordPress: https://econometriai.files.wordpress.com/2010/01/la-metodologia-de-la-investigacion-econometrica.pdf

Aguilar-Barojas, S. (2005). *Fórmulas para el cálculo de la muestra en investigaciones de salud.* Obtenido de redalyc.org: https://www.redalyc.org/pdf/487/48711206.pdf

Ahumada, H. (2014). *Variables Endógenas en los Modelos Econométricos.* Obtenido de Asociacion Argentina de Economia Politica: https://aaep.org.ar/anales/download/2014/ahumada.pdf

Alonso, C. (2010). *Econometría Tema 6: Modelos con Variables Explicativas Endógenas.* Obtenido de Universidad Carlos III de Madrid: http://www.eco.uc3m.es/docencia/econometria/NotasdeClase/Tema6Slides.pdf

Alonso, C. (2012). *Tema 1: Datos Económicos y Modelización Econométrica.* Obtenido de Web de OCW-UC3M: http://ocw.uc3m.es/economia/econometria/material-de-clase-1/tema-1-datos-economicos-y-modelizacion-econometrica.

Aparicio, C., Jaramillo, M., & San Román , C. (2011). *Desarrollo de la Infraestructura y Reduccion de la Pobreza: el Caso Peruano.* Lima.

Attanasio, O., & Székely, M. (2001). *Portrair of the poor: an assets-based approach.* Washington: Inter-Americn Development Bank.

Baum, C. (2006). *An Introduction to Modern Econometrics Using Stata.* Brighton: STATA press.

Bravo, D., & Vásquez Javiera. (2008). *Microeconometría Aplicada.* Santiago de Chile.

Brooks, C. (2008). *Introductory Econometrics for Finance.* Cambridge: Cambridge University Press.

Casalí, P., & Pena, H. (2012). *Los trabajadores independientes y la seguridad social en el Perú.* Obtenido de Bvs.Minsa: http://bvs.minsa.gob.pe/local/minsa/1907.pdf

Chacaltana, J. (2006). *¿Se puede prevenir la pobreza? hacia la construccion de una red de proteccion de los activos en el Perú.* Lima: CIES.

Cid S., L., Mora C., A., & Valenzuela H., M. (1990). *Inferencia Estadistica.* Concepcion.

Colin C., A., & Trivedi, P. (2005). *Microeconometrics Methods and Applications.*

Colin C., A., & Trivedi, P. K. (2009). *Microeconometrics Using STATA.* Texas: STATA Press.

Costa A., F. (2018). *Perú: Indicadores de Empleo e Ingreso por departamento 2007-2017.* Obtenido de INEI: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1537/cap11.pdf

Court, E., & Rengifo, E. (2011). *Estadísticas y Econometría Financiera.* Buenos Aires: Cengage Learning Argentina.

De Grange C., L. (2005). *Apuntes de clases ICT-2950 Tópicos de Econometría.* Santiago de Chile.

De la Cruz-Ore, J. L. (2013). *¿Qué significan los grados de libertad?* Obtenido de redalyc.org: https://www.redalyc.org/pdf/2031/203129458002.pdf

elEconomistaAamerica. (2020). Macro Región Norte ejecutó 53.3% de presupuesto para inversión pública. *elEconomistaAmerica*.

Escobar M., M., Fernández M., E., & Bernardi, F. (2012). *Análisis de datos con STATA.* Madrid: Centro de Investigaciones Sociologicas .

Farrar, D., & Glauber R. (1967). *Multicollinearity in Regression Analysis: The Problemas Revisited.* Obtenido de The Review of Economics and Statistics: doi:10.2307/1937887

Flores C., C. (2020). 400 mil trabajadores de mypes se beneficiarían con el Seguro de Vida desde el primer día de trabajo. *infoMercado*.

Freund, J. E., & Walpole, R. E. (1990). *Estadística matemática con aplicaciones.* México D. F.: Prentice-Hall Hispanomericana S.A.

Galán F., J., Feregrino F., J., Ruíz G., L. A., Quintana R., L., Mendoza G., M. Á., & Andrés R., R. (2016). *Econometría Aplicada utilizando R.* México D.F.

Gallardo , Y., & Moreno, A. (1999). *Aprende a Investigar. Modulo 3 Recolección de la información.* Obtenido de Universidad Libre: http://www.unilibrebaq.edu.co/unilibrebaq/images/CEUL/mod3recoleccioninform.pdf

Gestión. (2020). Sunat elevó tope: independienes que ganan hasta S/ 3,135 al mes no pagarán Impuesto a la Rento este año. *Gestión*.

Gestión. (2020). WEF: Perú se ubica en el penúltimo lugar en movilidad social en Sudamérica. *Géstion*.

Gil F., J. (1994). *Análisis de Datos Cualitativos. Aplicaciones a la Investigación Educativa.* Barcelona: Edit. PPU.

Greene, W. H. (2012). *Econometric Analysis.* New York: Pearson.

Gujarati, D. N., & Porter, D. C. (2010). *Econometría.* Ciudad de México: McGraw-Hill.

Hanke , J. E., & Wichern, D. W. (2006). *Pronósticos en los Negocios.* México : PEARSON EDUACACION.

Hernández A., J., & Zúñiga R., J. (2013). *Modelos Econométricos para el análisis económico.* ESIC.

Hernández S., R., Fernández C., C., & Baptista L., P. (2010). *Metodología de la investigación.* Ciudad de México: McGraw-Hill .

Kendall, M. G., & Stuart , A. (1961). *The advanced theory of statistics.* Nueva York.

L. Webster, A. (2005). *Estadístia Aplicada a los Negocios y la Economía.* México D.F.: McGraw-Hill.

Lidia G., M., & H. Landro, A. (2015). Acerca de la evolución del concepto de aleatoriedad en los mdoelos econométricos. *Revista de investigación en modelos matemáticos aplicados a la gestión y la economía*.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). *Estadística aplicada a los negocios y la economía.* México D.F.: McGraw-Hill Education.

Mendoza B., W. (2014). *Cómo investigan los economistas Guía para elaborar y desarollar un proyecto de inversion.* Lima.

Moya C., R. (2007). *Estadística descriptiva Conceptos y Aplicaciones.* Lima: Editorial San Marcos.

Novales, A. (1998). *Estadística y Econometría.* Madrid: McGraw-Hill.

Núñez Z., R. (2007). *Introducción a la econometría.* Cuidad de México: Trillas.

Orellana, L. (2008). *Regresión Lineal Simple.* Obtenido de Departamento de Matemática: http://www.dm.uba.ar/materias/estadistica_Q/2011/1/clase%20regresion%20simple.pdf

Otzen, T., & Manterola, C. (2017). *Técnicas de Muestreo sobre una Población a Estudio.* Obtenido de Scielo: https://scielo.conicyt.cl/pdf/ijmorphol/v35n1/art37.pdf

Ouliaris, S. (2011). *¿Qué son los modelos económicos? Cómo tratan de simular la realidad los economistas.*

Pardo, A., Ruiz, M., & San Martín, R. (2009). *Análisis de datos en ciencias sociales y de la salud I.* Madrid: Editorial Síntesis .

Pérez L., C. (2005). *Muestreo estadístico. Conceptos y problemas resultos.* Madrid: Pearson Educacion .

Pérez L., C. (2005). *Técnicas Estadísticas con SPSS 12. Aplicaciones al análisis de datos.* Madrid: Pearson Educación.

Pérez L., C. (2012). *Econometría Básica. Aplicaciones con Eviews, STATA, SAS y SPSS.* Madrid: IBERGARCETA Publicaciones.

Pérez-Tejada, H. E. (2007). *Estadística para las ciencias sociales, del comportamiento y de la salud.* Mexico D.F.: Cengage Learning Editores.

Ponce A., M. E., & Nolberto S., V. A. (2008). *Estadística inferencial aplicada.* Obtenido de WordPress.com: https://edgarmartinlarosa.files.wordpress.com/2013/07/est_inf_aplicada.pdf

Portillo, F. (2006). *Introducción a la econometría.*

Pucutay V., F. G. (2002). *Los Modelos Logit y Probir en la Investigación Social.* Lima: INEI .

Reinikka, R., & Svensson, J. (1999). *How inadequate provision of public infrastructure and services affects private investment.* Washington: World Bank.

Rodríguez, J., & Higa, M. (2010). *Ministerio de la Mujer y Poblaciones Vulnerables.* Obtenido de Informalidad, empleo y productividad en el Perú: http://www.mimp.gob.pe/webs/mimp/sispod/pdf/353.pdf

RPP. (2017). Esto es lo que debes saber si eres un trabajador independiente. *RPP*.

Saavedra, J., & Suárez , P. (2002). *El Financiamiento de la Educación Pública en el Perú: el Rol de las Familias.* Obtenido de Grupo de Análisis para el Desarollo : http://www.grade.org.pe/wp-content/uploads/ddt38.pdf

Scheaffer, R. L., Mendenhall III, W., & Lyman O., R. (2007). *Elementos de Muestreo.* Madrid: Thomson Editores.

Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data.* Cambdrige.

Stock, J., & Watson , M. (2012). *Introducción a la Econometría.* Madrid: Pearson Educación.

Uriel, E. (2013). *Regresión lineal multiple: estimacion y propiedades.* Valencia.

Uriel, E., & Aldás, J. (2005). *Análisis Multivariante Aplicado. Aplicaciones al Marketing, Investigación de Mercados, Economía, Dirección de Empresas y Turismo.* Madrid: Thomson Editores.

Véliz C., C. (2011). *Estadística para la administración y los negocios.* México DF: Pearson Educación.

Verbeek, M. (2004). *A Guide to Modern Econometrics.* Chichester: John Wiley & Sons Ltd.

Verdera V., F. (1998). *International Labour Organization.* Obtenido de Trabajadores a domicilio en el Perú: https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_123596.pdf

Wooldrige, J. M. (2009). *Introducción a la econometría Un enfoque moderno.* México DF: Cengage Learning.

Yamada, G. (2009). *Universidad del Pacífico.* Obtenido de Determinantes del desempeño del trabajador independiente y la microempresa familiar en el Perú: http://repositorio.up.edu.pe/bitstream/handle/11354/347/DD0901.pdf?sequence=1&isAllowed=y