Sobre el uso adecuado de la regresión lineal: conceptualización básica mediante un ejemplo aplicado a las ciencias de la salud

On the proper use of linear regression: basic conceptualization through an example applied to health sciences

Juan Hernández-Lalinde[1], Mgtr. https://orcid.org/0000-0001-6768-1873; j.hernandezl@unisimonbolivar.edu.co, Jhon-Franklin Espinosa-Castro[1], Mgtr. https://orcid.org/0000-0003-2186-3000; j.espinosa@unisimonbolivar.edu.co Diego García Álvarez[2], Mgtr. https://orcid.org/0000-0002-9350-785X; diegogarcia_123@hotmail.com, Valmore Bermúdez-Pirela3, Dr. https://orcid.org/0000-0003-1880-8887; v.bermudez@unisimonbolivar.edu.co
[1]Universidad Simón Bolívar, Departamento de Cs. Básicas, Sociales y Humanas, Cúcuta, Colombia.
[2]Universidad Rafael Urdaneta, Facultad de Cs. Sociales y Administrativas, Maracaibo, Venezuela.
[3]Universidad Simón Bolívar, Facultad de Ciencias de la Salud, Cúcuta, Colombia.
Autor para correspondencia: Juan Hernández-Lalinde. Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas.
Calle 14 entre avenidas 4 y 5, Barrio La Playa. C. P.: 540006. Cúcuta, Colombia. Correo electrónico: j.hernandezl@unisimonbolivar.edu.co.

Resumen

La regresión lineal es una de las técnicas estadísticas más difundidas y con mayores posibilidades de aplicación en contextos diversos; desde escenarios técnicos, prácticos o de ingeniería; hasta situaciones vinculadas a las ciencias sociales, del comportamiento o de la salud. Las ecuaciones de regresión posibilitan la interpretación de fenómenos al modelar parsimoniosamente la realidad, constituyéndose en una importante herramienta de ayuda para académicos, investigadores o profesionales. El objetivo del presente artículo de revisión es el de suministrar una guía teórico-práctica sobre los aspectos básicos de la regresión lineal, haciendo hincapié en el desarrollo del método de mínimos cuadrados ordinarios para obtener los estimadores de regresión, presentando las fórmulas requeridas para ello y empleando un ejemplo relacionado con la medicina para ilustrar el procedimiento. El nivel matemático es relativamente básico, aunque se requerirá de cierto conocimiento de estadística y álgebra matricial. El caso analizado gira en torno al estudio sobre la prevalencia de síndrome metabólico de la ciudad de Maracaibo, Venezuela. Se utiliza como variable dependiente la resistencia a la insulina a través del HOMA-2, y como regresores la edad, el índice de masa corporal y la razón triglicéridos-colesterol. También se ofrece la solución al problema mediante la utilización del programa de código abierto R-Studio.

Palabras clave: regresión lineal, mínimos cuadrados ordinarios, síndrome metabólico, Maracaibo, HOMA-2, triglicéridos-colesterol, edad, índice de masa corporal.

Abstract

Linear regression is one of the most widespread statistical techniques and with greater possibilities of application in diverse contexts; from technical, practical or engineering scenarios; to situations related to social, behavioral and health sciences. Regression equations make possible to interpret phenomena by the parsimonious modeling of reality, becoming an important tool of help for academics, researchers or professionals. The aim of this review is to provide a theoretical-practical guide on the basic aspects of linear regression, emphasizing the development of the ordinary least squares method to obtain the regression estimators, presenting the formulas required for this and using an example related to medicine to illustrate the procedure. The mathematical level is relatively basic, although some knowledge of statistics and matrix algebra will be required. The analyzed case revolves around the study on the prevalence of metabolic syndrome from the city of Maracaibo, Venezuela. Insulin resistance measured by HOMA-2 is used as a dependent variable, and age, body mass index and triglyceride-cholesterol ratio as predictor variables. The solution to the problem is also offered by using the R-Studio program.

Keywords: linear regression, ordinary least squares, metabolic syndrome, Maracaibo, HOMA-2, triglyceride-cholesterol, age, body mass index.

Introducción

La regresión lineal es una de las técnicas estadísticas más utilizadas. Quizá una de las razones que expliquen su popularidad es la «elegancia» de los modelos que de ella se derivan y que permiten representar la compleja realidad de los fenómenos estudiados, de forma clara, resumida y sencilla. En casi cualquier contexto, la multiplicidad de factores hace de su modelización un arduo trabajo, en especial si se desea incorporar a este esquema la totalidad de variables involucradas o si se aspira a reproducir de manera exacta el estado de la naturaleza. Un modelo de regresión lineal, por el contrario, hace gala del principio de parsimonia atribuido a William de Ockham, alejándose ligeramente de la enrevesada realidad y mostrándola de la forma más simple posible.

Ahora bien, no es esta la única ventaja que ofrece la regresión lineal, ya que el análisis subyacente a esta técnica también posibilita medir la relación entre las variables modeladas, sirviendo como herramienta para investigaciones en múltiples campos del saber. Un ejemplo de lo anterior se encuentra en la publicación de He et al.[1], estudio en el que se evaluó la asociación entre la obesidad abdominal y el síndrome metabólico en adolescentes de la Universidad de Pensilvania; o en el artículo de Gloria y Steinhardt[2], investigación en la que se usaron modelos jerárquicos de regresión para examinar el efecto moderador que tenía la resiliencia sobre las combinaciones estrés-ansiedad y estrés-depresión. Otro caso se encuentra en el trabajo de Khamisa et al.[3], quienes plantearon ecuaciones de regresión lineal para encontrar predictores de salud a partir de variables como el estrés laboral, el síndrome de agotamiento y la satisfacción por el trabajo en enfermeras de Sudáfrica.

Tomando en cuenta estas referencias, se plantea la presente revisión. El objetivo principal es el de servir como orientación teórico-práctica a investigadores de las ciencias sociales, muy particularmente a aquellos que se desempeñan en el área de la medicina. Se propone esta iniciativa como respuesta a la escaza formación estadística que reciben muchos de estos académicos durante su trayectoria profesional, hecho que acarrea inconvenientes en algunas situaciones en las que se desconoce la fundamentación matemática del procedimiento empleado, pudiendo conducir a interpretaciones erradas o a predicciones que no concuerdan de la realidad. La revisión hace uso de un ejemplo auténtico en el que se determinó la prevalencia de síndrome metabólico en la ciudad de Maracaibo, Venezuela. El aspecto matemático se aborda en cierto detalle, haciendo énfasis en el desarrollo del método de estimación por mínimos cuadrados ordinarios y presentando las fórmulas necesarias para lograrlo. Con el propósito de facilitar la comprensión de la técnica y de validar los resultados obtenidos manualmente, se añade al artículo la resolución del ejemplo por medio de R-Studio.

Breve reseña histórica

Cualquiera podría suponer que fue Karl Pearson quien desarrolló el coeficiente producto-momento de Pearson, aunque en realidad las nociones de correlación y regresión fueron originalmente concebidas por Sir Francis Galton debido a su fascinación por la genética y la herencia[4]. Este conocido polímata británico afirmó en su famosa publicación de 1885, titulada «Regression Towards Mediocrity in Heriditary Stature», lo siguiente: «it appeared from these experiments that the offspring did not tend to resemble their parent seeds in size, but to be always more mediocre than they —to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small»[5]. Sin ánimos de realizar una traducción exacta sino una interpretación básica de esta aseveración, lo que Galton planteó fue que el tamaño de las semillas descendientes tendía a distanciarse del de sus progenitoras «regresando» al promedio. Estos hallazgos fueron obtenidos mientras experimentaba con plantas de guisantes dulces, aunque posteriormente se encontraron resultados similares al evaluar la relación entre la estatura de padres e hijos adultos[5,6]. Así pues, la palabra «regresión» fue acuñada inicialmente por Galton para describir este fenómeno y rápidamente fue adoptada para caracterizar otro tipo de situaciones, incluso aquellas en las que la «regresión a la media» no estaba presente. A pesar de esta inconsistencia, el término siguió usándose entonces y continua empleándose hoy en día[6–12].

Definición del análisis de regresión

El análisis de regresión permite contestar interrogantes que tienen que ver con la dependencia de una variable respuesta a partir de uno o varios regresores, incluyendo aspectos como la predicción de futuros valores, la identificación de predictores significativos o el impacto originado por modificaciones en las variables explicativas[13]. Si la correlación se limita a medir la fuerza de asociación entre dos características tratándolas simétricamente, la regresión propone un modelo lineal en el que los cambios observados en una variable se explicarían debido al efecto de otras. Por tanto, esta herramienta estadística supera las restricciones del coeficiente de correlación al plantear asimétricamente el vínculo entre variables, considerando a una como dependiente y a otras como independientes[14–19]. Tal y como señalan Montgomery et al.[20], el análisis de regresión es una de las técnicas más utilizadas para trabajar con datos multifactoriales. Los autores proponen dos causas básicas que explicarían la vasta difusión de esta metodología: por un lado, el proceso lógico que implica plantear la relación entre variables a partir de una ecuación; y por el otro, la elegancia asociada a la matemática subyacente a los modelos que permiten representar la realidad de forma sencilla y comprensible.

Modelos lineales generalizados

La regresión lineal pertenece a la familia de modelos conocida como modelos lineales generalizados (MLG). En términos simples, un MLG es una ecuación que posee la siguiente forma:

$$Y_i = \mu_i + \varepsilon_i, \qquad (1)$$

donde $Y_i Y_i$ es la variable respuesta para la *i*-ésima observación, $\mu_i \mu_i$ es alguna función monótona diferenciable y ε_i es el error aleatorio del modelo, aquel que reúne toda la variación de Y_i que no es explicada por los regresores. Como se verá a continuación, la regresión lineal es un caso particular de los MLG cuando μ_i es una función lineal de los predicto-

res y ε_i se distribuye normalmente con media cero y varianza constante[1] **[21]**.

Modelo de regresión lineal simple mediante un ejemplo

Para explicar en términos sencillos qué es un modelo de regresión lineal simple, considérese el ejemplo que habrá de utilizarse en el artículo. Bermúdez et al.[22,23] investigaron la prevalencia de síndrome metabólico (SM) en adultos de la ciudad de Maracaibo, Venezuela, basándose en los criterios de la «International Diabetes Federation» y de la «Adult Treatment Panel», mediante un estudio descriptivo que contó con 2230 sujetos escogidos por muestreo aleatorio polietápico[24]. Hernández-Lalinde et al. seleccionaron mediante muestreo aleatorio simple a 30 individuos de esta base de datos, submuestra que luego fue utilizada para presentar una revisión acerca de la verificación de los supuestos del coeficiente de correlación de Pearson[25]. Este conjunto de datos, que puede ser descargado gratuitamente desde el enlace http://bit.ly/2Zm2hIA, será utilizado en esta publicación como guía para la exposición de los aspectos teóricos del análisis de regresión.

Supóngase que un grupo de endocrinólogos desea explorar la posible relación entre la resistencia a la insulina medida a través del «Homeostatic Model Assessment 2» (HOMA-2) y la relación triacilglicéridos-colesterol (TAG/HDL). Asúmase además que la intención de los investigadores no es únicamente la de evaluar la fuerza y la dirección de esta asociación, sino también la de construir un modelo que pueda predecir la resistencia a la insulina a partir de la razón TAG/HDL. Por último, presuma que los médicos utilizan la matriz de datos que se presenta en la **tabla 1** y obtienen la siguiente fórmula:

$$\text{HOMA2} = 0.594\left(\frac{\text{TAG}}{\text{HDL}}\right) + 0.599. \qquad (2)$$

Sin ahondar por ahora en el aspecto matemático, la ecuación (2) plantea que, por cada cambio unitario de la relación triglicéridos-colesterol, el HOMA-2 aumentaría 0.594 veces. Por otro lado, el término independiente indicaría que, si la relación TAG/HDL fuese nula, el valor mínimo de la resistencia a la insulina sería de 0.599. Como se discutirá en las próximas secciones, la interpretación práctica de esta cantidad no siempre es pertinente.

En la expresión (2), el HOMA-2 es la variable dependiente o respuesta, aquella sobre la que recae el interés y sobre la que se pretenden realizar las predicciones. En otros ámbitos, como el econométrico, suele llamársele variable explicada, endógena o regresando[26]. Por su parte, el cociente TAG/HDL es la variable independiente, la que se supone está relacionada con el HOMA-2 de tal forma que, las variaciones observadas en ella, ocasionarían cambios lineales en aquella a una tasa fija. Esta variable también es conocida como variable explicativa, exógena, regresor o predictor[6]. La cantidad 0.594 es la pendiente de la ecuación, cifra que indica cuánto cambiaría la resistencia a la insulina cuando se modifica unitariamente el TAG/HDL. Es costumbre que la pendiente también sea identificada como razón de cambio del modelo.

1 La suposición de normalidad no es necesaria para la estimación por MCO, pero sí para la inferencia estadística sobre la significación de la regresión (ANOVA) y de los predictores (prueba t-Student).

Por último, la constante 0.599 es el término independiente, intercepto o corte con la ordenada. Como se ha dicho, su interpretación es inviable en muchas ocasiones.

Estructura general del modelo de regresión lineal simple

Considérese una variable respuesta *Y* y un regresor *X*. La relación lineal entre *Y* y *X* queda definida a partir de la siguiente expresión:

$$Y = \beta_1 X + \beta_0 + \varepsilon, \qquad (3)$$

donde β_1 y β_0 son los coeficientes de regresión o parámetros del modelo y $\varepsilon\varepsilon$ es el componente del error o perturbación aleatoria. Es importante aclarar que, aunque se afirme taxativamente que la asociación entre *Y* y *X* es lineal, en realidad se asume que la ecuación (3) proveerá una aproximación bastante aceptable de la verdadera relación y que las desviaciones del modelo serán recogidas por $\varepsilon\varepsilon$[21]. Por otro lado, β_1 y β_0 son la pendiente y el intercepto de la recta de regresión, respectivamente.

Ahora supóngase que se dispone de *n* pares ordenados $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ como los que se suministran en la **tabla 1**. La ecuación (3) pasaría a constituir un conjunto de *n* ecuaciones de la forma:

$$\begin{array}{lll} y_1 & = \beta_1 x_1 + \beta_0 + & \varepsilon_1 \\ y_2 & = \beta_1 x_2 + \beta_0 + & \varepsilon_2 \\ \vdots & \vdots & \vdots \\ y_n & = \beta_1 x_n + \beta_0 + & \varepsilon_n \end{array}, \qquad (4)$$

donde y_i, x_i y ε_i son los valores respectivos de la respuesta, del predictor y del error para el *i*-ésimo sujeto. La ecuación (4) puede reescribirse de forma abreviada como:

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i, \qquad (5)$$

para $i = 1, 2, 3, \cdots, n$. La importancia de las fórmulas (4) y (5) se verá cuando se contemple el método de estimación de parámetros por mínimos cuadrados ordinarios (MCO).

Estructura general del modelo de regresión lineal múltiple

Extiéndase la situación anterior al caso en el que *Y* está relacionada con *k* regresores. El modelo de regresión lineal múltiple quedaría definido por:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \beta_0 + \varepsilon. \qquad (6)$$

Nótese que ahora hay $k + 1$ coeficientes de regresión puesto que se incluye al intercepto; además, los $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ ya no representan meramente a la pendiente de la recta, sino a las pendientes del hiperplano formado por el espacio *k*-dimensional de los predictores[20,27]. Si ahora se reescribe la ecuación (6) para todas las observaciones de la muestra, se obtiene:

$$\begin{array}{lllllll} y_1 = & \beta_1 x_{11} + & \beta_1 x_{21} + & \cdots & \beta_1 x_{k1} + & \beta_0 + & \varepsilon_1 \\ y_2 = & \beta_1 x_{12} + & \beta_1 x_{22} + & \cdots & \beta_1 x_{k2} + & \beta_0 + & \varepsilon_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n = & \beta_1 x_{1n} + & \beta_1 x_{2n} + & \cdots & \beta_1 x_{kn} + & \beta_0 + & \varepsilon_n \end{array}. \qquad (7)$$

El sistema de ecuaciones anterior puede presentarse en forma matricial como:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad (8)$$

o en forma compacta de la siguiente manera:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (9)$$

donde **Y** es el vector respuesta de $n \times 1$, **X** es la matriz de regresores de $n \times k$, $\boldsymbol{\beta}$ es el vector de coeficientes de $k \times 1$ y $\boldsymbol{\varepsilon}$ es el vector de los errores de $n \times 1$. Obsérvese en **X** que los términos de la primera columna son todos iguales a 1. Estos corresponden al «regresor ficticio», aquellos que se incluyen en el sistema de ecuaciones para estimar el valor de β_0[13,15,28–30].

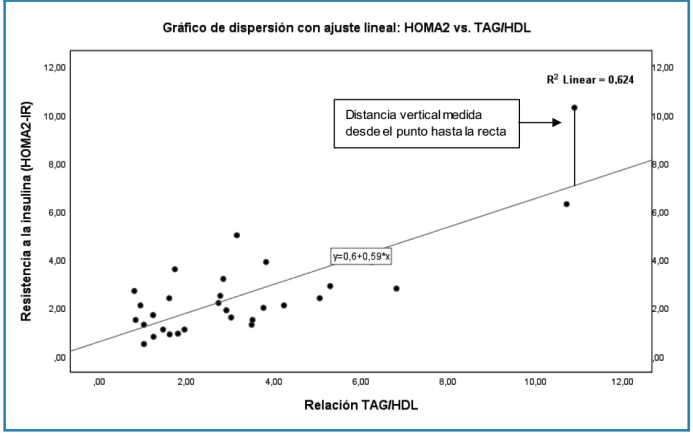
Mínimos cuadrados ordinarios para regresión lineal simple

En cierta forma, esta etapa del análisis de regresión es similar a la inferencia estadística, en la que se estiman los parámetros poblacionales a partir de los estadísticos obtenidos de una muestra. Uno de los métodos que permite lograr esto y que se presenta en esta revisión es el de mínimos cuadrados ordinarios, o como se le conoce por su forma en inglés, «ordinary least squares». Dicho procedimiento consiste en minimizar la suma de cuadrados del error; es decir, la suma de cuadrados de las distancias verticales que separan los puntos de la línea recta. Para visualizar esto con facilidad, remítase a la **figura 1**: nótese que en el gráfico de dispersión se ha señalado por simplicidad solo la observación ubicada en la abscisa más extrema; sin embargo, el método de estimación toma en cuenta todas las distancias verticales. Lo anterior puede sintetizarse mediante la siguiente expresión:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2. \quad (10)$$

Figura 1. Diagrama de dispersión con recta de regresión ajustada. Se muestra como referencia una sola distancia vertical, pero los MCO toman en cuenta todas las distancias verticales con respecto a la recta.

Figura 1



El objetivo que se persigue con los MCO es encontrar las estimaciones de los coeficientes de regresión que minimicen *S*. Aunque no se profundizará en el procedimiento que permite lograrlo, sí se mencionará que para ello debe derivarse parcialmente la fórmula (10) con respecto a β_1 y β_0 e igualar a cero. El detalle del desarrollo matemático puede encontrarse en Drapper y Smith[6], Montgomery et al.[20] y Uriel[31]. En definitiva, los valores de b_1 y b_0 que minimizan *S* se hallan a través de:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} = \frac{S_{XY}}{S_{XX}}, \quad (11)$$

$$b_0 = \bar{y} - b_1 \bar{x}, \quad (12)$$

donde $\bar{y}$ y $\bar{x}$ son las medias de la variable independiente y dependiente, respectivamente. Al reunir los términos calculados con las ecuaciones (11) y (12) se obtiene el modelo estimado de regresión simple:

$$\hat{Y} = b_1 X + b_0. \quad (13)$$

Volviendo al ejemplo, asígnese *Y* al HOMA-2 y *X* a la relación triacilglicéridos-colesterol. Como puede apreciarse en la última fila de la **tabla 1**, las sumatorias que aparecen en la ecuación (11) ya están calculadas, de manera que solo bastaría reemplazar para hallar b_1. Así:

$$b_1 = \frac{367.01 - (76.44)(98.45)/30}{518.58 - (98.45)^2/30} = \frac{116.16}{195.50} = 0.594.$$

Al sustituir b_1 en la fórmula (12) se obtiene la estimación del intercepto. Entonces:

$$b_0 = \left(\frac{76.44}{30}\right) - (0.59)\left(\frac{98.45}{30}\right) = 0.599.$$

Finalmente, al suplantar en la expresión (13) los términos generales por las estimaciones, se tiene:

$$\hat{Y} = 0.594X + 0.599.$$

Compruébese que este modelo coincide con el mostrado en la ecuación (2). Ahora podría utilizarse esta fórmula para predecir la resistencia a la insulina sin necesidad de haberla medido; siempre y cuando, por su puesto, se cuente con los registros del paciente asociados a los triglicéridos y al colesterol de alta densidad. Supóngase que la razones TAG/HDL de tres sujetos cualesquiera son de 2.52, 4.15 y 12.08. Los valores predichos para el HOMA-2 serían:

$$\hat{y}_1 = 0.59(2.52) + 0.61 = 2.09.$$

$$\hat{y}_2 = 0.59(4.15) + 0.61 = 3.06.$$

$$\hat{y}_3 = 0.59(12.08) + 0.61 = 7.34.$$

Conviene aclarar algo en este punto. La cantidad 12.08 ha sido usada para advertir sobre la incorrección en la que se incurre muchas veces al usar ecuaciones de regresión para predecir resultados. Recuérdese que b_1 y b_0 se obtuvieron de una base de datos en la que el mínimo y el máximo para el TAG/HDL fueron de 0.80 y 10.89, respectivamente (véase **tabla 1**). Por tanto, usar valores que excedan este rango po-

dría generar resultados que se alejen sistemáticamente del valor real, ocasionando un incremento significativo del error. Así pues, la sugerencia pasa por sustituir cantidades consideradas normales según el conjunto de datos utilizado para hallar los coeficientes estimados de regresión.

Tabla 1. Base de datos extraída del estudio del síndrome metabólico de la ciudad de Maracaibo. El HOMA-2 y la relación TAG/HDL se representan mediante *Y* y *X*, respectivamente. Al final de la tabla se muestran las sumatorias empleadas para la obtención de los estimadores.

Cód.	Y	X	Y²	X²	XY	Cód.	Y	X	Y²	X²	XY
S01	2.00	3.76	4.00	14.14	7.52	S16	2.80	2.77	6.25	7.67	6.93
S02	1.70	1.23	2.89	1.51	2.09	S17	0.80	6.81	7.84	46.38	19.07
S03	1.50	0.83	2.25	0.69	1.25	S18	10.30	1.24	0.64	1.54	0.99
S04	1.10	1.95	1.21	3.80	2.15	S19	1.50	10.89	106.09	118.59	112.17
S05	2.20	2.73	4.84	7.45	6.01	S20	2.90	3.51	2.25	12.32	5.27
S06	3.20	2.84	10.24	8.07	9.09	S21	2.10	5.29	8.41	27.98	15.34
S07	0.50	1.02	0.25	1.04	0.51	S22	2.10	4.23	4.41	17.89	8.88
S08	1.30	1.02	1.69	1.04	1.33	S23	3.90	0.94	4.41	0.88	1.97
S09	1.30	3.49	1.69	12.18	4.54	S24	1.90	3.82	15.21	14.59	14.90
S10	2.40	1.60	5.76	2.56	3.84	S25	3.60	2.91	3.61	8.47	5.53
S11	0.90	1.61	0.81	2.59	1.45	S26	2.70	1.73	12.96	2.99	6.23
S12	6.30	10.71	39.69	114.70	67.47	S27	1.60	0.80	7.29	0.64	2.16
S13	5.01	3.15	25.10	9.92	15.78	S28	1.10	3.02	2.56	9.12	4.83
S14	0.93	1.80	0.86	3.24	1.67	S29	3.90	1.46	1.21	2.13	1.61
S15	2.50	3.76	4.00	14.14	7.52	S30	2.40	6.24	15.21	38.94	24.34
Sumatorias: $\sum x = 98.45, \sum y = 76.44, \sum x^2 = 518.58, \sum y^2 = 305.40, \sum xy = 367.01.$											

Mínimos cuadrados ordinarios para regresión lineal múltiple

En este caso, el método de MCO parte de la ecuación (10) con la modificación respectiva debido a que se plantea un modelo de regresión lineal múltiple. En consecuencia, para obtener el vector de estimadores de los β_i debe minimizarse:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (14)$$

Lo anterior se logra cuando[6,20,27,31,32]:

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0. \quad (15)$$

Omitiendo el aspecto matemático que permite resolver la expresión (15), se presenta la fórmula con la que pueden encontrarse los coeficientes de regresión estimados; a saber:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (16)$$

La ecuación de regresión múltiple ajustada resulta de:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}, \quad (17)$$

donde $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es conocida como matriz de proyección o «Hat Matrix», la cual es de suma importancia para conocer el efecto que tiene cada observación de la variable respuesta sobre los valores ajustados o predichos[32,33].

Para ilustrar la utilidad de los elementos matemáticos suministrados arriba, supóngase que el equipo de investigadores desea incluir dos variables adicionales como posibles predictores del HOMA-2: el índice de masa corporal (IMC) y la edad. El detalle de estos datos se provee en la **tabla 2**. Para hallar los b_i, debe encontrarse primero $\mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X}$, lo que se alcanza al premultiplicar a la matriz de regresores su matriz

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 3.76 & 1.23 & \cdots & 5.05 \\ 32.95 & 24.60 & \cdots & 33.19 \\ 35 & 39 & \cdots & 31 \end{bmatrix} \cdot \begin{bmatrix} 1 & 3.76 & 32.95 & 35 \\ 1 & 1.23 & 24.60 & 39 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 5.05 & 33.19 & 31 \end{bmatrix} = \begin{bmatrix} 30 & 98.45 & 921.44 & 1057.00 \\ 98.45 & 518.58 & 3195.97 & 3440.04 \\ 921.44 & 3195.97 & 29\,417.11 & 32\,259.87 \\ 1057.00 & 3440.04 & 32\,259.87 & 39\,411.00 \end{bmatrix}.$$

El procedimiento anterior bien podría realizarse manualmente, aunque en este caso se han empleado funciones matriciales de hojas de cálculo. El lector podrá comprobar esto con facilidad al recurrir a cualquier paquete estadístico o al utilizar calculadoras científicas. Ahora se halla el vector $\mathbf{X}'\mathbf{y}$:

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 3.76 & 1.23 & \cdots & 5.05 \\ 32.95 & 24.60 & \cdots & 33.19 \\ 35 & 39 & \cdots & 31 \end{bmatrix} \cdot \begin{bmatrix} 2.00 \\ 1.70 \\ \vdots \\ 2.40 \end{bmatrix} = \begin{bmatrix} 76.44 \\ 367.01 \\ 2453.23 \\ 2655.23 \end{bmatrix}.$$

Los estimadores resultantes son:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 30 & 98.45 & 921.44 & 1057.00 \\ 98.45 & 518.58 & 3195.97 & 3440.04 \\ 921.44 & 3195.97 & 29\,417.11 & 32\,259.87 \\ 1057.00 & 3440.04 & 32\,259.87 & 39\,411.00 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 76.44 \\ 367.01 \\ 2453.23 \\ 2655.23 \end{bmatrix} = \begin{bmatrix} 0.899 \\ 0.591 \\ 0.001 \\ -0.010 \end{bmatrix}.$$

Finalmente, la ecuación ajustada de regresión múltiple es:

$$\hat{Y} = 0.591X_1 + 0.001X_2 - 0.010X_3 + 0.899,$$

o lo que es igual:

$$\text{HOMA2} = 0.591\left(\frac{\text{TAG}}{\text{HDL}}\right) + 0.001(\text{IMC}) - 0.010(\text{Edad}) + 0.899.$$

Tal y como se hizo en la sección anterior, podrían asignarse valores a X_1, X_2 y X_3 con la intención de predecir la resistencia a la insulina. Recuérdese que estas cantidades deben estar contenidas dentro del rango de los datos recolectados en la investigación.

Tabla 2. Base de datos extraída del estudio del síndrome metabólico de la ciudad de Maracaibo. El HOMA-2, TAG/HDL, IMC y la edad se representan mediante *Y*, X_1, X_2 y X_3, respectivamente.

Cód.	Y	X1	X2	X3	Cód.	Y	X1	X2	X3
S01	2,00	3,76	32,95	35	S16	2,80	6,81	43,53	40
S02	1,70	1,23	24,60	39	S17	0,80	1,24	28,32	45
S03	1,50	0,83	30,34	20	S18	10,30	10,89	30,71	44
S04	1,10	1,95	35,72	25	S19	1,50	3,51	24,08	36
S05	2,20	2,73	34,68	34	S20	2,90	5,29	41,01	38
S06	3,20	2,84	35,49	38	S21	2,10	4,23	25,46	41
S07	0,50	1,02	19,00	42	S22	2,10	0,94	22,52	45
S08	1,30	1,02	17,29	22	S23	3,90	3,82	36,19	47
S09	1,30	3,49	24,83	39	S24	1,90	2,91	32,34	35
S10	2,40	1,60	32,70	22	S25	3,60	1,73	40,00	24
S11	0,90	1,61	31,37	45	S26	2,70	0,80	28,08	49
S12	6,30	10,71	30,78	23	S27	1,60	3,02	33,30	37
S13	5,01	3,15	31,98	24	S28	1,10	1,46	26,19	40
S14	0,93	1,80	29,30	43	S29	3,90	6,24	38,54	28
S15	2,50	2,77	26,95	26	S30	2,40	5,05	33,19	31

Obtención de los modelos de regresión mediante R-Studio

A fin de ofrecer una alternativa menos rigurosa en términos matemáticos, se presenta en esta sección la rutina necesaria para obtener los modelos de regresión del ejemplo contemplado en el artículo. Para ello, deben instalarse versiones iguales o posteriores a la 3.1.6 y 1.2.1335 de R y R-Studio, respectivamente. A continuación, se describe esquemáticamente el procedimiento:

- Importar la base de datos: la importación desde fuentes externas en R-Studio puede realizarse de diversas formas. En esta publicación se recurrirá al comando «read.csv» descargando el fichero directamente desde la web. Si el lector encontrase algún inconveniente que le impidiese hacerlo de esta manera, siempre podrá importar la base de datos desde su computador, una vez haya descargado el archivo por medio del enlace que se mostrará a continuación. Así pues, cópiese y ejecútese en la consola del programa la instrucción siguiente: Muestra_SM =read.csv("http://bit.ly/2Zm2hIA").

- Crear el modelo de regresión lineal simple: para obtener la regresión lineal simple, cópiese y ejecútese la siguiente rutina: MRLS =lm(HOMA2IR~TAG_HDL, data =Muestra-SM). Con esta instrucción se construye el modelo que propone a la resistencia a la insulina como variable dependiente y a la razón triglicéridos-colesterol como único regresor.

- Crear el modelo de regresión lineal múltiple: procediendo de forma similar, cópiese y ejecútese la siguiente sucesión de comandos: MRLM=lm(HOMA2IR~TAG_HDL + IMC + Edad, data =Muestra-SM). Como puede apreciarse, en esta oportunidad los predictores son la relación TAG/HDL, el IMC y la edad.

- Obtener las estimaciones: el procedimiento en R-Studio que genera las estimaciones de los coeficientes de regresión, aparte de otros resultados que no se toman en cuenta en esta revisión, se obtiene al copiar y ejecutar los siguientes comandos: summary (MRLS) y summary (MRLM) (tenga la precaución de copiarlos en líneas diferentes).

La **figura 2** exhibe cómo deben transcribirse las instrucciones recientemente descritas al cuadro de «scripts» de R-Studio, mientras que la **figura 3** muestra la salida del programa. Compruébese que, salvo por las diferencias debidas al redondeo, los resultados coinciden con los ya presentados.

Figura 2. Secuencia de instrucciones para el análisis de regresión en R-Studio

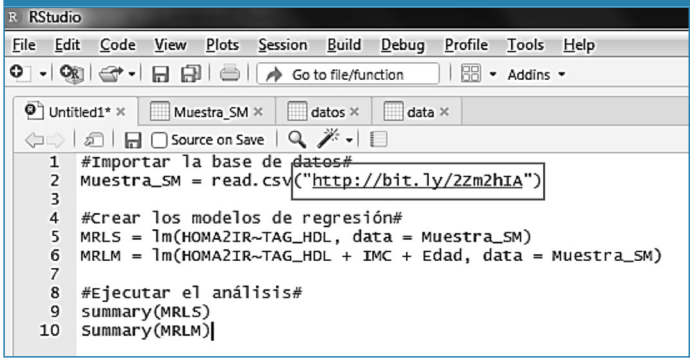
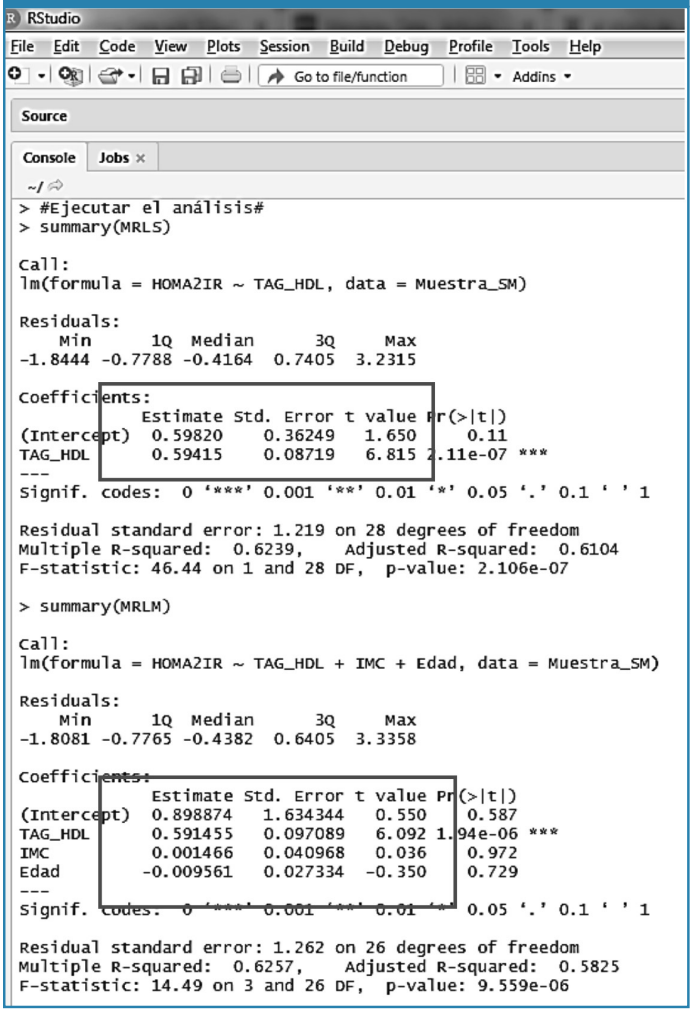


Figura 3. Salida de R-Studio para el ejemplo usado en el artículo



Referencias

1. He F, Rodriguez-Colon S, Fernandez-Mendoza J, Vgontzas AN, Bixler EO, Berg A, et al. Abdominal Obesity and Metabolic Syndrome Burden in Adolescents—Penn State Children Cohort Study. J Clin Densitom. enero de 2015;18(1):30-6.
2. Gloria CT, Steinhardt MA. Relationships among positive emotions, coping, resilience and mental health: positive emotions, resilience and health. Stress Health. abril de 2016;32(2):145-56.
3. Khamisa N, Oldenburg B, Peltzer K, Ilic D. Work related stress, burnout, job satisfaction and general health of nurses. Int J Environ Res Public Health. enero de 2015;12(1):652-66.
4. Stanton JM. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. J Stat Educ. 2001;9(3):1-13.
5. Galton F. Regression Towards Mediocrity in Heriditary Stature. J Anthropol Inst. 1885;15(2):246-63.
6. Drapper NR, Smith H. Applied Regression Analysis. 3.ª ed. New York, NY: John Wiley & Sons, Inc.; 1998. 705 p. (Wiley series in probability and statistics).
7. Stigler SM. Regression towards the mean, historically considered. Stat Methods Med Res. 1997;6(2):103-14.
8. Nesselroade JR, Stigler SM, Baltes PB. Regression toward the mean and the study of change. Psychol Bull. 1980;88(3):622-37.

9. Bland JM, Altman DG. Statistics Notes: Some examples of regression towards the mean. BMJ. 24 de septiembre de 1994;309(6957):780-780.

10. Regression to the mean [Internet]. Institute for Work & Health. 2014 [citado 10 de febrero de 2019]. Disponible en: https://www.iwh.on.ca/what-researchers-mean-by/regression-to-mean

11. Barnett AG. Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 27 de agosto de 2004;34(1):215-20.

12. Chen S, Chen H. Regression to the mean [Internet]. Encyclopedia Britannica. 2016 [citado 10 de febrero de 2019]. Disponible en: https://www.britannica.com/topic/regression-to-the-mean

13. Weisberg S. Applied Linear Regression. 3.ª ed. Hoboken, N.J: John Wiley & Sons, Ltd; 2005. 310 p. (Wiley series in probability and statistics).

14. Kissell R, Poserina J. Regression Models. En: Optimal Sports Math, Statistics, and Fantasy [Internet]. Elsevier; 2017 [citado 3 de febrero de 2019]. p. 39-67. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128051634000025

15. Wittek P. Regression Analysis. En: Quantum Machine Learning [Internet]. Elsevier; 2014 [citado 3 de febrero de 2019]. p. 85-8. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128009536000086

16. Liengme BV. Regression Analysis. En: A Guide to Microsoft Excel 2013 for Scientists and Engineers [Internet]. Elsevier; 2016 [citado 3 de febrero de 2019]. p. 157-79. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128028179000088

17. Mishra S, Datta-Gupta A. Regression Modeling and Analysis. En: Applied Statistical Modeling and Data Analytics [Internet]. Elsevier; 2018 [citado 3 de febrero de 2019]. p. 69-96. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128032794000043

18. Lalanne C, Mesbah M. Correlation and Linear Regression. En: Biostatistics and Computer-based Analysis of Health Data using R [Internet]. Elsevier; 2016 [citado 3 de febrero de 2019]. p. 89-110. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9781785480881500051

19. Hernández-Lalinde J, Espinosa-Castro J-F, Penaloza-Tarazona M-E, Rodríguez J, Chacón G, Toloza-Sierra C, et al. Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. Arch Venez Farmacol Ter. 2018;38(5):587-95.

20. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. Introduction to Linear Regression Analysis. 5th ed. New York, NY: John Wiley & Sons, Inc.; 2012. 872 p. (Wiley series in probability and statistics).

21. Samprit Chatterjee, Ali S. Hadi. Regression Analysis by Example. 4th ed. Hoboken, N.J: John Wiley & Sons, Ltd; 2006. 383 p. (Wiley series in probability and statistics).

22. Bermúdez V, Pacheco M, Rojas J, Córdova E, Velázquez R, Carrillo D, et al. Epidemiologic Behavior of Obesity in the Maracaibo City Metabolic Syndrome Prevalence Study. Maedler K, editor. PLoS ONE. 18 de abril de 2012;7(4):e35392.

23. Bermúdez V, Rojas J, Salazar J, Calvo MJ, Morillo J, Torres W, et al. The Maracaibo city metabolic syndrome prevalence study: primary results and agreement level of 3 diagnostic criteria. Rev Latinoam Hipertens. 2014;9(4):20-32.

24. Bermúdez V, Marcano RP, Cano C, Arráiz N, Amell A, Cabrera M, et al. The Maracaibo City Metabolic Syndrome Prevalence Study: Design and Scope: Am J Ther. mayo de 2010;17(3):288-94.

25. Hernández-Lalinde J, Espinosa-Castro J-F, Díaz-Camargo É, Bautista-Sandoval M, Riaño-Garzón ME, García Álvarez D, et al. Sobre el uso adecuado del coeficiente de correlación de Pearson: verificación de supuestos mediante un ejemplo aplicado a las ciencias de la salud. Arch Venez Farmacol Ter. 2018;37(5):452-61.

26. Ullah A, Wan A, Chaturvedi A. Handbook Of Applied Econometrics And Statistical Inference [Internet]. 1.ª ed. New York, NY: CRC Press; 2002 [citado 21 de agosto de 2019]. 749 p. (Statistics: textbooks and monographs; vol. 165). Disponible en: https://www.taylorfrancis.com/books/9780203911075

27. Montgomery DC, Runger GC. Applied statistics and probability for engineers. 3rd ed. New York: Wiley; 2003. 706 p.

28. Alkarkhi AFM, Alqaraghuli WAA. Regression Models. En: Easy Statistics for Food Science with R [Internet]. Elsevier; 2019 [citado 3 de febrero de 2019]. p.107-24. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128142622000078

29. Angelini C. Regression Analysis. En: Encyclopedia of Bioinformatics and Computational Biology [Internet]. Elsevier; 2019 [citado 3 de febrero de 2019]. p.722-30. Disponible en: https://linkinghub.elsevier.com/retrieve/pii/B9780128096338203609

30. Rawlings JO, Pantula SG, Dickey DA. Applied Regression Analysis: a Research Tool. 2nd ed. New York: Springer; 1998. 657 p. (Springer texts in statistics).

31. Uriel E. El modelo de regresión simple: estimación y propiedades. Univ Valencia. 2016;1(1):1-49.

32. Uriel E. Introducción a la econometría [Internet]. 1.ª ed. Universidad de Valencia; 2013 [citado 22 de agosto de 2019]. 234 p. Disponible en: https://www.uv.es/uriel/manual/Introduccion%20a%20la%20econometria.pdf

33. Hoaglin DC, Welsch RE. The Hat Matrix in Regression and ANOVA. Am Stat. febrero de 1978;32(1):17-22.