Riesgo de Ocurrencia de Delitos Sexuales en el Departamento de Nariño a partir de un Modelo de Regresión para Datos de Recuento

Andrés F. Jaramillo Mejía & Euclides Díaz Arcos

Universidad Tecnológica de Pereira
Facultad de Ingeniería Industrial
Maestría en Investigación Operativa y Estadística
Pereira - Risaralda
Febrero de 2018

Riesgo de Ocurrencia de Delitos Sexuales en el Departamento de Nariño a partir de un Modelo de Regresión para Datos de Recuento

Andrés F. Jaramillo Mejía & Euclides Díaz Arcos

Trabajo presentado como requisito parcial para optar al título de
Magister en Investigación Operativa y Estadística

Director
Arsenio Hidalgo Troya
Magister en Estadística

Universidad Tecnológica de Pereira
Facultad de Ingeniería Industrial
Maestría en Investigación Operativa y Estadística
Pereira - Risaralda
Febrero de 2018

Título en español

Riesgo de Ocurrencia de Delitos Sexuales en el Departamento de Nariño a partir de un Modelo de Regresión para Datos de Recuento.

Resumen: En este trabajo se realiza un estudio poblacional sobre el riesgo de ocurrencia de delitos sexuales durante los periodos *2015-I*, *2015-II*, *2016-I* y *2016-II*, en Nariño. Para ello se hace uso de los modelos de regresión para datos de recuento, tales como: Modelo de Regresión Poisson, Modelo de Regresión Binomial Negativo y Modelo de Regresión Inflados con Ceros. Las variables explicativas consideradas para realizar el ajuste del modelo son: sexo, rango de edad, región del hecho y temporada. La variable observada Y es el número de delitos sexuales que ocurren en una determinada población de $N_{m\times 1}$, donde $N_{m\times 1}$ es la variable de control o variable de exposición. Así, lo que se modela en este caso no es el conteo, sino la tasa que se expresa como y_i/N_i. Se logró establecer que las variables que mejor explican estos eventos son: rango de edad, región del hecho y sexo de la víctima. Se calcularon las tasas de incidencia IRR y se mostró que los las mujeres son aproximadamente 10 veces más vulnerables que los hombres a estos sucesos. La región con mayor riesgo es la central, además con respecto a la edad, el grupo que se encuentra entre 10 y 19 años son los de mayor riesgo. En cuanto a la temporada, la tasa de delitos sexuales fue levemente decreciendo coforme pasa el tiempo. Haciendo uso de las pruebas de bondad de ajuste y tomando en cuenta los criterios de selección AIC y BIC, se logra seleccionar el modelo de regresión binomial negativo (MRBN) como el mejor modelo que se acerca a la representación de los datos.

Abstract: In this work, a population study is conducted on the risk of the occurrence of sexual crimes during the periods *2015-I*, *2015-II*, *2016-I* y *2016-II*, in Nariño. This is done using regression models for counting data, such as: Poisson Regression Model, Negative Binomial Regression Model and Regression Model Inflated with Zeros. The explanatory variables considered to perform the adjustment of the model are: sex, age range, region of the event and season. The observed variable Y is the number of sexual crimes that occur in a given population of $N_{m\times 1}$, where $N_{m\times 1}$ is the control variable or exposure variable. Thus, what is modeled in this case is not the count, but the rate that is expressed as y_i/N_i. It was established that the variables that best explain these events are: age range, region of the event and sex of the victim. The IRR incidence rates were calculated and it was shown that women are approximately 10 times more vulnerable than men to these events. The region with the greatest risk is the central one, in addition with respect to age, the group that is between 10 and 19 years old are the most at risk. As for the season, the rate of sexual offenses was slightly decreasing as time passes. Using the goodness-of-fit tests and taking into account the AIC and BIC selection criteria, the negative binomial regression model (NBRM) is selected as the best model that approaches the representation of the data.

Palabras clave: Modelo lineal generalizado, Modelo inflado con ceros, Función de enlace, Población a riesgo, Variable *offset*.

Keywords: Generalized linear model, Model inflated with zeros, Link function, Population at risk, Variable *offset*.

Nota de aceptación

Jurado
PhD: Herman José Serrano López

Jurado
MSc: Alvaro Antonio Trejos Carpintero

Director
MSc: Arsenio Hidalgo Troya

Pereira, Febrero de 2018

Índice general

Índice de tablas

Índice de figuras

Introducción

La Organización Mundial de la Salud (OMS), señala en su *Informe Mundial Sobre la Violencia y Salud (2002)*, donde se tocan temas de violencia doméstica, suicidio, delitos sexuales, entre otros, que:

"La violencia esta tan presente, que se la percibe a menudo como un componente ineludible de la condición humana ... en la que el papel de los profesionales de la salud se limita a tratar las consecuencias".

Así pues, podemos razonar cuán usual es la exposición de las personas a eventos violentos (tan sólo como ejemplo, los medios de información muestran todos los días casos de violencia), esto ha hecho que se perciba la violencia como un evento común y aceptable, particularmente para la violencia sexual.

Al respecto, Williams (1984) citado en [1], donde enfoca la violencia sexual contra la pareja VSCP, resalta que la relación existente entre la víctima y el agresor figura como una de las principales dificultades para el registro de estos casos donde es muy frecuente que la víctima no acepte que las agresiones ejercidas por su pareja sean un delito. Por su parte, Petrzelová en [21], expone que la violencia sexual en menores de edad es un fenómeno aún mas complejo, donde se guarda silencio, por que a menudo esta involucrado un miembro de la familia como agresor, luego no se realizan las correspondientes denuncias por prejuicios sociales o la vergüenza de la víctima.

Cabe anotar que, según Petrzelová, el ser humano se puede reponer o adaptar fácilmente a muchas condiciones adversas, pero no sucede así cuando se trata de violencia sexual, ya que por las características de este tipo de delito se convierte en un fenómeno que afecta a la víctima más allá del momento en el que fue agredida.

En efecto, según varios autores[1], la violencia sexual repercute en la salud física y mental de las víctimas, ésta se asocia a un mayor riesgo de diversos problemas sexuales y reproductivos con consecuencias que se manifiestan tanto de inmediato como muchos años después de la agresión.

Según el mismo informe se afirma que la violencia sexual se presenta con mayor frecuencia en mujeres y niñas donde el agresor es un hombre, pero también existen casos documentados de violencia sexual donde las víctimas son hombres y niños con agresor masculino, efectivamente las estadísticas presentadas en el capítulo 1 de este trabajo reafirman tal hecho para ambos géneros.

[1]Informe Mundial Sobre la Violencia y Salud (2002).

Señalamos aquí la creciente preocupación de la OMS frente al aumento de casos de violencia en todas sus formas, con el ánimo de pedir a los países que tomen iniciativas para dar mejor respuesta a estos eventos en la prevención, mitigación, pronta respuesta, y penalización de los mismos.

En Colombia, los esfuerzos para atender los casos de violencia sexual y de género se visibilizan en cuanto a la creación de nuevas leyes y el fortalecimiento y/o modificación de las que ya existen, como también en aumentar la severidad de las penas para éstos delitos, para eso el Ministerio de Salud y Protección Social cuenta con la ley 1146 de 10 de julio de 2007, la *Guía de atención al menor y a la mujer maltratada* resolución 412 de 2000, y el *Protocolo de atención integral en salud para las víctimas de violencia sexual* Resolución 459 de 2012.

Por lo anterior, es de nuestro interés realizar un estudio demográfico en el departamento de Nariño, acerca de éstos sucesos, utilizando modelos de regresión para datos de recuento. Para ello se utiliza la información que recopila el *Instituto Nacional de Medicina Legal y Ciencias Forenses INMLCF*[2], con el fin de realizar un estudio poblacional sobre las variables que mas influyen en el evento de ser una víctima potencial de violencia sexual.

En este trabajo se habla de riesgo de ocurrencia de un delito sexual con base en las denuncias que las víctimas interpusieron ante Medicina Legal, esto significa que cuando se haga alusión a un delito sexual es porque se esta hablando del delito sexual denunciado ante *INMLCF*.

Con la intención de delimitar esta investigación, dadas las diferentes formas de violencia presentes, nos centraremos en el estudio de las denuncias de violencia sexual, ésta comprende una variedad de actos bien descritos por el código penal, Ley 599 de 2000. El propósito de este trabajo es evaluar los factores asociados con el riesgo de delitos sexuales en el departamento de Nariño utilizando un modelo de regresión para datos de recuento. Para ello, inicialmente se realiza un análisis descriptivo que permite identificar características de la población a riesgo. Posteriormente se evalúan los modelos de regresión de Poisson, Binomial Negativo, Poisson inflado con ceros y Binomial Negativo Inflado con Ceros, seleccionando entre éstos el modelo que mejor representa la población de estudio.

Los modelos mencionados anteriormente se utilizan para datos de recuento. Los recuentos se definen como el número de sucesos o eventos que ocurren en una misma unidad de observación durante un intervalo temporal o espacial definido (ver [14]), para nuestro estudio, la unidad de observación es el tamaño de una población específica de N, siendo N un vector con tamaños de individuos. Las variables de recuento se caracterizan por su naturaleza discreta y no negativa. Es decir, si Y es una variable aleatoria de recuento entonces los valores que toma son $0, 1, 2, \ldots$.

Generalmente las variables de recuento acumulan una gran cantidad de ceros en las observaciones, por tal motivo es importante recurrir en estos casos a modelos de regresión que consideren este fenómeno. De igual manera, en sucesos reales muchas veces se presenta sobredispersión en los datos, es decir, cuando la varianza supera el valor esperado en la variable respuesta. Este acontecimiento mas general que el primero y que muchas veces conlleva a exceso de ceros, se puede modelar con distribuciones de probabilidad que consideran esta coyuntura. Uno de los modelos de regresión que consideran este asunto es el Modelo de Regresión Binomial Negativo (MRBN), en el cual la variable respuesta Y sigue una distribución binomial negativa (BN).

[2]Datos obtenidos por medio del Observatorio de Género de Nariño.

Para el caso más general, el modelado estadístico via regresión para datos de recuento permite analizar una variable respuesta Y en función de variables explicativas $X_1, X_2, \ldots, X_p$ mediante la expresión

$$E(Y/X = x),$$

es decir, la variable Y dado $X_1, X_2, \ldots, X_p$ sigue una distribución discreta, ya sea Poisson o Binomial Negativa, en cualquiera de sus formulaciones (ver[6, 7, 10, 11, 12, 14, 19]).

La principal característica de los modelos para datos de recuento mencionados, es la *función de enlace*, la cual se utiliza para relacionar la variable repuesta con los predictores o variables explicativas. En los modelos lineales generales (ML), la función de enlace que emplean para relacionar la variable respuesta con las variables independientes es la función identidad, es decir, $g(\mu) = \mu$. Sin embargo, esto no ocurre en los modelos discretos cuando la variable dependiente es un conteo y sus valores no pueden ser negativos. Para estos sucesos, la función de enlace mas utilizada es la función logaritmo, la cual se relaciona con $g(\lambda) = \ln(\lambda)$, donde λ representa el valor esperado.

Los modelos de regresión de Poisson y Binomial Negativo, a diferencia del Modelos Inflados con Ceros, pertenecen a la familia de los modelos lineales generalizados (MLG), por ende, todas las propiedades que caracterizan a los MLG la cumplen los modelos estándar de regresión Poisson y Binomial negativo (ver [16, 20]).

En este informe, la variable respuesta muestral tipo recuento se define como el número de delitos sexuales presentados en una determinada población específica de $N_{m\times 1}$, bajo las variables explicativas *sexo*, *rango de edad*, *región del hecho* y *semestre* de ocurrencia. La variable N se denomina variable de exposición o de control, es una variable discreta de tipo numérico que permite establecer la población a riesgo en cada observación. La variable de exposición también es conocida como variable *Offset*, se expresa como:

$$Offset = \ln N\,.$$

Mientras que en los modelos lineales (ML) se produce una relación de identidad entre los valores ajustados y el predictor lineal, $\mu = \eta$, en los MLG la linealidad se establece en la escala del predictor lineal pero no en la escala de los valores ajustados. No se da, por tanto, la identidad entre valores ajustados y valores predichos, sino que entre ellos media una funcion que los relaciona, la función de enlace:

$$g(\lambda) = \ln(\lambda) = \eta = X\beta\,,$$

donde η es el predictor lineal, X es una matriz con p variables explicativas y β es un vector columna con p coeficientes de regresión, los cuales se asocian con su respectiva variable independiente. El valor λ se encuentra en la escala de la variable respuesta.

En este sentido, el presente trabajo se ha estructurado en tres capítulos. En el primer capítulo se presenta un análisis descriptivo, en su mayor parte enfocado a las víctimas, con los casos de denuncias sobre violencia general y delitos sexuales en el departamento de Nariño, durante las temporadas *2015* y *2016*. Aquí se determinan tasas y razón de tasas que permiten describir el riesgo que presenta una población específica en sucederle el evento, teniendo en cuenta variables como: región del hecho, rango de edad, sexo, ocupación de la víctima, tipo de agresor, etc. En el segundo capítulo se detallan los modelos teóricos de regresión para datos de recuento, entre ellos: Modelo de regresión de Poisson

(MRP), Modelo de Regresión Binomial Negativo (MRBN), Modelo de Regresión de Poisson Inflado con Ceros, conocido mas popularmente como Modelo de Regresión ZIP (*Zero Inflate Poisson*) y Modelo de Regresión Binomial Negativo Inflado con Ceros, denominado Modelo de Regresión ZINB (*Zero Inflate Negative Binomial*). En la primera sección del tercer capítulo se da a conocer como se obtuvo la población a riesgo por medio de la página oficial del DANE `http://www.dane.gov.co`, posteriormente se presentan algunos test estadísticos que permiten probar sobredispersión de los datos. Luego se muestra la estimación de parámetros en los modelos y las pruebas de bondad de ajuste. Mas adelante, en el mismo capítulo se presenta una pequeña sección donde aparece la interpretación del modelo de regresión binomial negativo (MRBN). Luego se presenta un apéndice donde se detallan algunos elementos que no fueron incluidos en el trabajo central. Se finaliza el informe con algunas conclusiones y recomendaciones.

El MRBN y el modelo ZINB resultaron los más apropiados para explicar el número de delitos sexuales. Estos modelos son los que mejor representan los datos, teniendo en cuenta las pruebas de bondad de ajuste y los criterios AIC y BIC. La prueba de Vuong **no** permitió decidir entre el MRBN y el modelo de regresión ZINB, sin embargo, la prueba arrojo una mejora a favor del modelo ZIP versus el MRP. Es importante anotar que el estadístico de Voung prueba la hipótesis nula de que los dos modelos son igualmente cercanos al verdadero proceso de generación de datos, frente a la alternativa de que un modelo está más cerca, pero no toma ninguna decisión sobre si el modelo "más cercano"es el verdadero modelo.

Como en todo proceso de modelado, el objetivo es la obtención de un modelo que sea capaz de representar los datos y, al mismo tiempo reducir la complejidad, es decir, atender a los criterios de **bondad de ajuste**. Podemos afirmar que el MRBN cumple tales condiciones para estos datos.

Gran parte de los resultados presentados en este informe fueron desarrollados en el Software estadístico STATA, algunos otros fueron complementados con R y Gretl.

CAPÍTULO 1

Estadísticas sobre Delitos Sexuales en Nariño

En la primera sección de este capítulo se presenta las estadísticas de denuncias por violencia en el departamento de Nariño en los años *2015* y *2016*. Posteriormente, en la sección 1.2 se muestra un análisis descriptivo de las denuncias sobre delitos sexuales en las temporadas anteriormente mencionadas. Las gráficas y tablas reveladas en este capítulo son producción propia que los autores desarrollaron.

1.1. Denuncias por violencia en el departamento de Nariño

De acuerdo a los reportes de Medicina Legal, en los años *2015* y *2016* se presentaron 9242 denuncias por diferentes tipos de violencia, entre los que se encuentra: violencia interpersonal, violencia intrafamiliar, delitos sexuales, homicidios, entre otros. En los 9242 casos presentados en los dos años, 3703 corresponden al año *2015* y 5539 al *2016*. De los 3703 denuncias presentadas sobre violencia en *2015*, 283 corresponden a delitos sexuales. De igual forma, de las 5539 denuncias presentadas ante Medicina Legal de violencia en *2016*, 192 son de delitos sexuales. La variable DELITOS SEXUALES enmarca las categorías de abuso sexual, asalto sexual y acceso carnal violento, entre otros, conceptos bien definidos en la Legislación Colombiana, Ley 599 de 2000. La tabla 1.1 presenta el número de denuncias por violencia que se presentaron en los dos años mencionados. El término Violencia General en la tabla hace referencia a las denuncias sobre violencia intrafamiliar, violencia interpersonal, homicidios, entre otros.

Tabla 1.1. Número de denuncias por violencia.

TIPO VIOLENCIA	**AÑO 2015**	**AÑO 2016**	**TOTAL**
Delitos sexuales	283	192	475
Violencia General	3443	5348	8791
TOTAL	3726	5540	9266

Esto demuestra que aunque las denuncias sobre violencia general aumentaron aproximadamente en un $55,33\,\%$ en el año *2016*, con respecto al año *2015*, las denuncias sobre delitos sexuales decayeron en un $32,15\,\%$ en el año *2016* con respecto al año anterior. Por otro lado, con respecto a cada año, el porcentaje de delitos sexuales en *2015* representa

aproximadamente el $7,6\,\%$ del total de casos denunciados como violencia, sin embargo en *2016*, el porcentaje de violencia sexual corresponde al $3,47\,\%$ del total de denuncias interpuestas ante Medicina Legal en el departamento.

Importante anotar que a pesar de que el número de denuncias sobre delitos sexuales disminuyeron en un $32,15\,\%$ en el año *2016*, con respecto al año *2015*, esto no necesariamente implica que en la realidad los eventos hayan ido en tendencia a la baja, sino que pueden existir factores de ruido, por ejemplo, las denuncias no se presentaron, ya que por ser eventos de carácter reservado, la víctima, o acudiente de la víctima cuando el agredido es menor de edad, algunas veces no dan a conocer la situación ante las entidades competentes. Sin embargo, el hecho de conocer que las denuncias sobre violencia general aumentaron en el año *2016* en $55,35\,\%$ con respecto al año anterior, debe tomarse como un estimativo de crecimiento de violencia.

En este sentido, es importante analizar las características tanto del agresor como de la víctima, en cualquier tipo de violencia de las mencionadas anteriormente (ver [5]), esto con el objeto de proteger a las víctimas e implementar planes de seguridad y políticas sociales de convivencia que permitan intervenir este tipo de hechos.

La tabla 1.2 muestra el número de denuncias por violencia, de todo tipo, interpuestas ante Medicina Legal en el Departamento de Nariño a través del tiempo, en ella se observa una pequeña tendencia positiva. La información está ordenada por trimestres, los cuatro primeros corresponden al año *2015* y los otros cuatro al *2016*.

Tabla 1.2. Tendencia de denuncias por violencia.

Trimestre	Número de casos
2015-1	997
2015-2	960
2015-3	939
2015-4	830
2016-1	1389
2016-2	1447
2016-3	1391
2016-4	1313
TOTAL	9266

Cabe resaltar, que del total de casos denunciados por violencia en el año *2015*, el $32,02\,\%$ corresponden a violencia interpersonal, en este ámbito el tipo de agresor es de carácter múltiple, desde grupos al margen de la ley hasta la misma fuerza pública. En el año *2016*, el $64,92\,\%$ de los hechos denunciados por violencia corresponden a violencia interpersonal, al igual que el año anterior, el presunto agresor es de categoría múltiple. De acuerdo a lo anterior y teniendo en cuenta la matriz de datos de trabajo, el tipo de denuncias que se incremento en el año *2016*, con respecto al año inmediatamente anterior, fue las denuncias sobre violencia interpersonal. Con respecto a los otros tipos, las denuncias por violencia algunas veces disminuyeron y otras el número de ocurrencias se mantuvo constante (con la misma tasa).

En la figura 1.1 se presenta, a través de una serie de tiempo, el número de denuncias por violencia de los dos años mencionados. En este gráfico, al igual que la tabla 1.1, se aprecia una pequeña tendencia positiva del comportamiento de las denuncias que las

víctimas interpusieron ante Medicina Legal. También se observa un crecimiento de los casos iniciando el primer mes del año *2016* hasta el sexto mes del mismo.

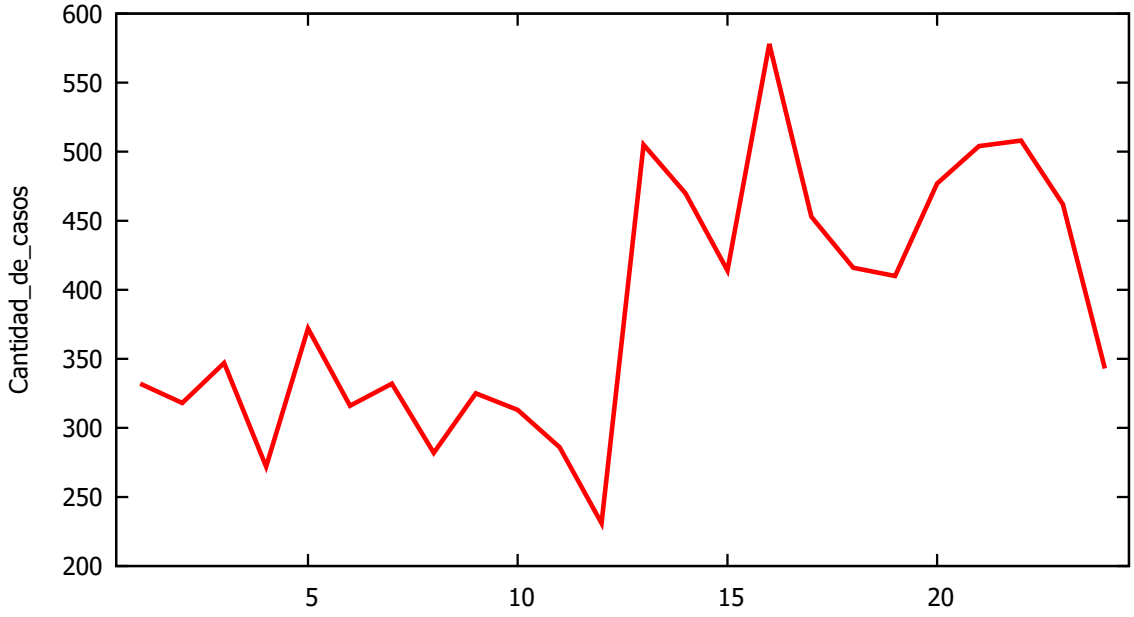

FIGURA 1.1. Tendencia de las denuncias por violencia en el tiempo.

Por otro lado, para una mejor interpretación y lectura de los resultados, se define la **Tasa** de denuncias por violencia, en cada 100.000 habitantes como:

$$Tasa = \left(\frac{\text{Número de casos presentados}}{\text{Población a riesgo por año}} \right) 100.000 \text{ Habitantes.} \tag{1.1}$$

Una de las variables a tener en cuenta en el desarrollo de este capítulo, es la región del departamento donde sucedió el posible delito sexual. Los municipios que conforman cada región se muestran en el apéndice A, tabla A.1.

En este sentido, la figura 1.2 presenta el número de denuncias por violencia para sucesos presentados en cada región en la temporada mencionada. El diagrama de barras muestra que la región con más número de denuncias es la región central. Sin embargo, para establecer diferencias se debe tener en cuenta la población a riesgo de cada región. Cabe señalar que la población a riesgo para la región Central entre los años *2015* y *2016*, de acuerdo a las proyecciones del DANE, es de 1.167.806 habitantes, asi que la tasa de denuncias por violencia para esta región es de 517,81 por cada 100.000 habitantes. Para la región de la Costa Pacífica, la población a riesgo en el transcurso de los dos años es de 828.233 habitantes, luego la tasa de denuncias por violencia en esta región es de 139,09 por cada 100.000 habitantes. Siguiendo con este procedimiento, la población a riesgo para la región Sur es de 567.785 habitantes, luego la tasa de denuncias por violencia para esta zona es 228,96 por cada 100.000 habitantes. En la región Norte, la población a riesgo es de 492.273, así que la tasa de denuncias por violencia para esta región es de 43,27 por 100.000 habitantes. Finalmente la tasa de denuncias por violencia para la región Suroccidente,

teniendo en cuenta la población a riesgo de 454.037 habitantes, es de 122,02 por cada 100.000 habitantes.

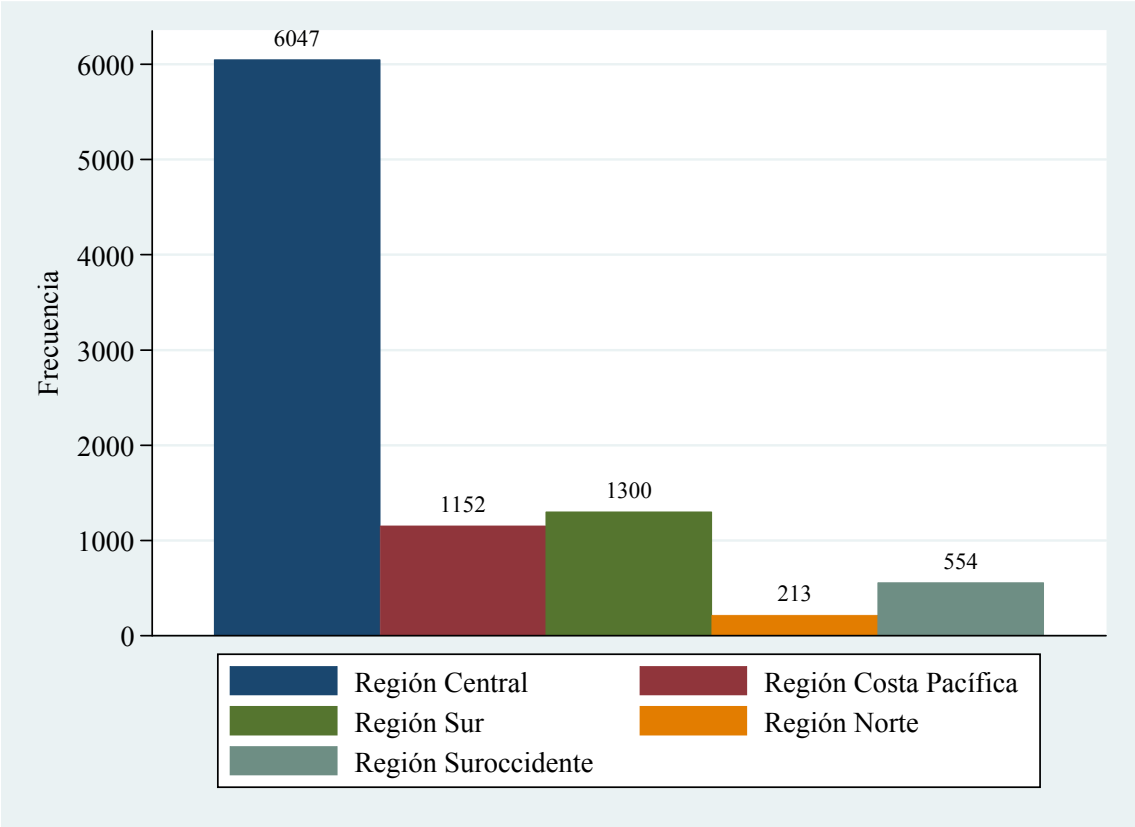

FIGURA 1.2. Denuncias de violencia por región.

Tomando en cuenta la información de la figura 1.2 y la ecuación (1.1), la tasa más alta de ocurrencias por violencia se encuentra en la región Centro, seguido por la región de la Costa Pacífica. La región con menor tasa de denuncias por violencia es la región Norte. Las tasas de denuncias más pequeñas no necesariamente implican disminución de violencia, sino mas bien, disminución en las denuncias pero no en los eventos reales. Sin embargo, un aumento en las denuncias es un buen estimativo para indicar que la violencia esta creciendo.

Por otro lado, como se observa en el diagrama de barras (ver figura 1.3), en el año *2016* el número de denuncias se incrementaron en cada una de las cinco regiones con respecto al año anterior. Hablando a nivel del departamento, la tasa de denuncias por violencia para el año *2015* fue de 213,62 denuncias por cada 100.000 habitantes. Para el año *2016*, la tasa de denuncias corresponde a 313,72 por cada 100.000 habitantes. Esto indica que la tasa de denuncias por violencia se incremento en 100 casos por cada 100.000 habitantes en el año *2016*, con respecto al año inmediatamente anterior.

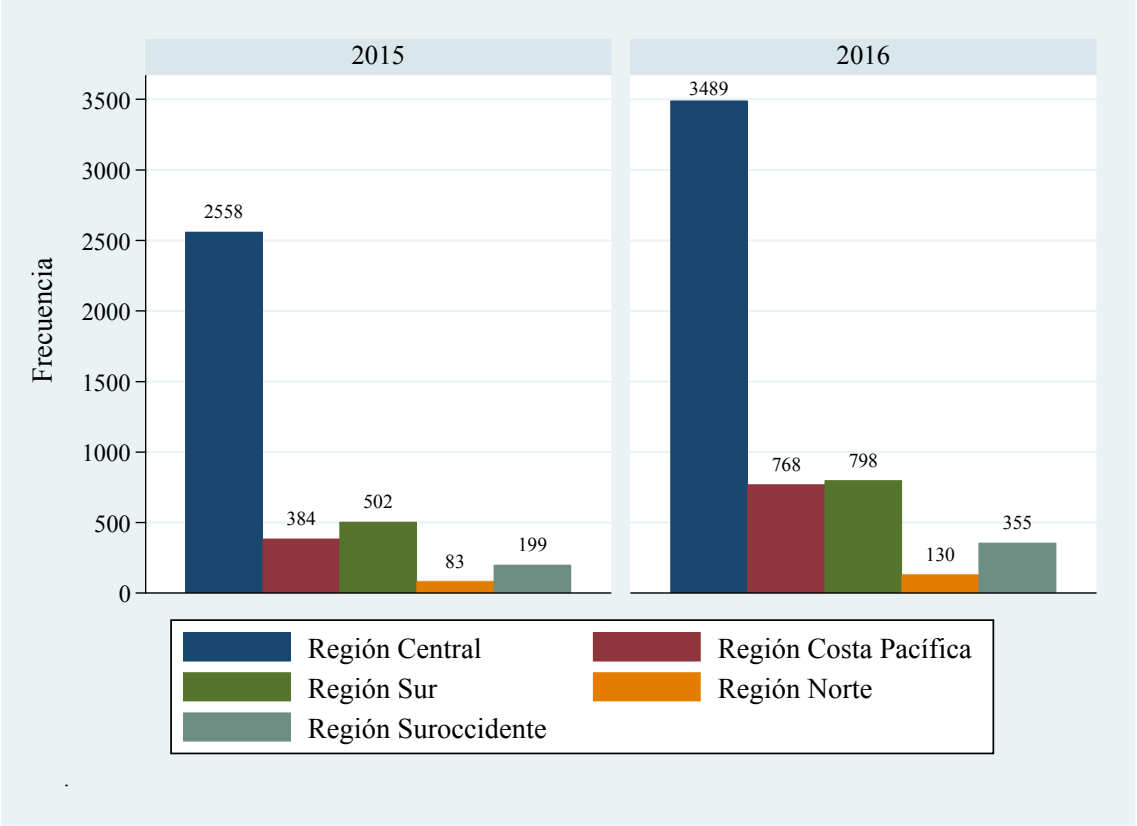

FIGURA 1.3. Denuncias de violencia por región y año.

La tabla 1.3 muestra las tasas de denuncia por violencia en cada una de las categorías de edad que proyecta el DANE en su página oficial de Internet. Las tasas, como se mostró en la ecuación (1.1), se expresan por cada 100.000 habitantes. Observe que de acuerdo a la población a riesgo, las tasas de denuncias aumentan o disminuyen dependiendo de la edad de las víctimas. Las mayores tasas de denuncias, con una población a riesgo mas o menos equidistante, se presentan en víctimas con edades entre 10 y 49 años.

TABLA 1.3. Tasas de denuncias por edad.

Edad	Casos	Población a Riesgo	Tasa
0 a 9 años	224	654.654	34,22
10 a 19 años	1.551	662.328	234,17
20 a 29 años	3.438	590.246	582,47
30 a 39 años	2.076	518.420	400,45
40 a 49 años	1.135	405.214	280,10
50 a 59 años	501	304.253	164,67
60 a 69 años	240	201.200	119,28
70 y más	101	173.819	58,11

Finalmente, el histograma 1.4 presenta el número de denuncias de violencia en Nariño para la variable edad en las temporadas *2015* y *2016*. En la figura se aprecia que la población entre los 10 y 49 años es la que mayor riesgo tiene de ser víctima de violencia, algo similar a lo que muestra la tabla 1.3. Sin embargo, en este caso como se esta hablando

de denuncias de cualquier tipo de violencia, la población entre los 0 y 9 años se observa en mínima proporción pero comparada con la totalidad de los casos presentados.

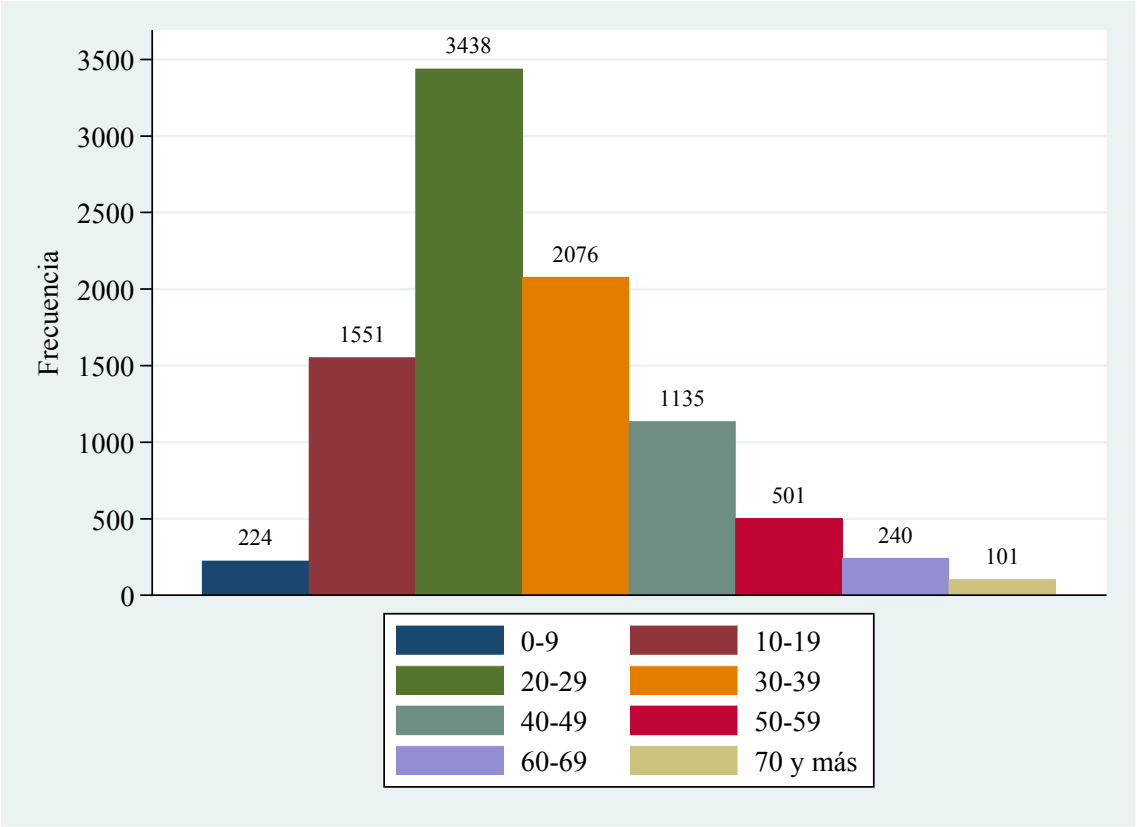

FIGURA 1.4. Denuncias de violencia por categoría de edad.

1.2. Delitos sexuales en el departamento Nariño

Las denuncias sobre delitos sexuales, de acuerdo a datos oficiales de Medicina Legal, en el año *2015* se presentaron 283 eventos y en *2016*, 192, para un total de 475 eventos en las dos temporadas. Sin embargo, como se dijo en la sección 1.1, es importante aclarar que el decrecimiento de las denuncias en *2016* no implican disminución de este tipo de violencia, porque se debe tener en cuenta que en algunas oportunidades las víctimas no denuncian y por ende se presentan falsos negativos.

Las variables a tener cuenta en esta sección son: Periodo del evento, Rango de edad de la víctima, Región donde sucedió el evento, Escolaridad de la víctima, Convivencia de la víctima con el presunto agresor, Ocupación de la víctima, Sexo de la víctima, Tipo de agresor de la víctima y Sexo del presunto agresor. Cada una de estas variables tiene asociadas unas respectivas categorías.

En la figura 1.5 se muestran representadas dos variables con sus respectivas categorías, las variables son: Región del evento y Escolaridad de la víctima. Observe que en las cinco regiones, las personas con grado de escolaridad Básica Primaria son las personas con

mayor riesgo de sufrir este tipo de eventos con 181 casos. Le sigue en orden decreciente, personas con escolaridad de Básica Secundaria, 126 denuncias. Luego, con 55 casos en todo el departamento, personas con escolaridad Inicial y Preescolar, con 51 casos, personas con Educación Media. Observe que de los 475 denuncias presentadas en los dos años, en 413 ocasiones la escolaridad de la víctima esta entre Preescolar y Bachillerato, es decir, de las 475 denuncias sobre delitos sexuales interpuestas ante Medicina Legal, en el 86, 95 % de éstas, las víctimas cuentan a lo sumo con Bachillerato como escolaridad. Importante comprobar que el grado de escolaridad de una persona influye en el factor de riesgo o en el factor de protección que tiene la misma con respecto a este tipo de violencia. En cuanto a las regiones, la región Centro y la región Sur son las que mayor número de denuncias presentaron por eventos sucedidos en las mismas regiones.

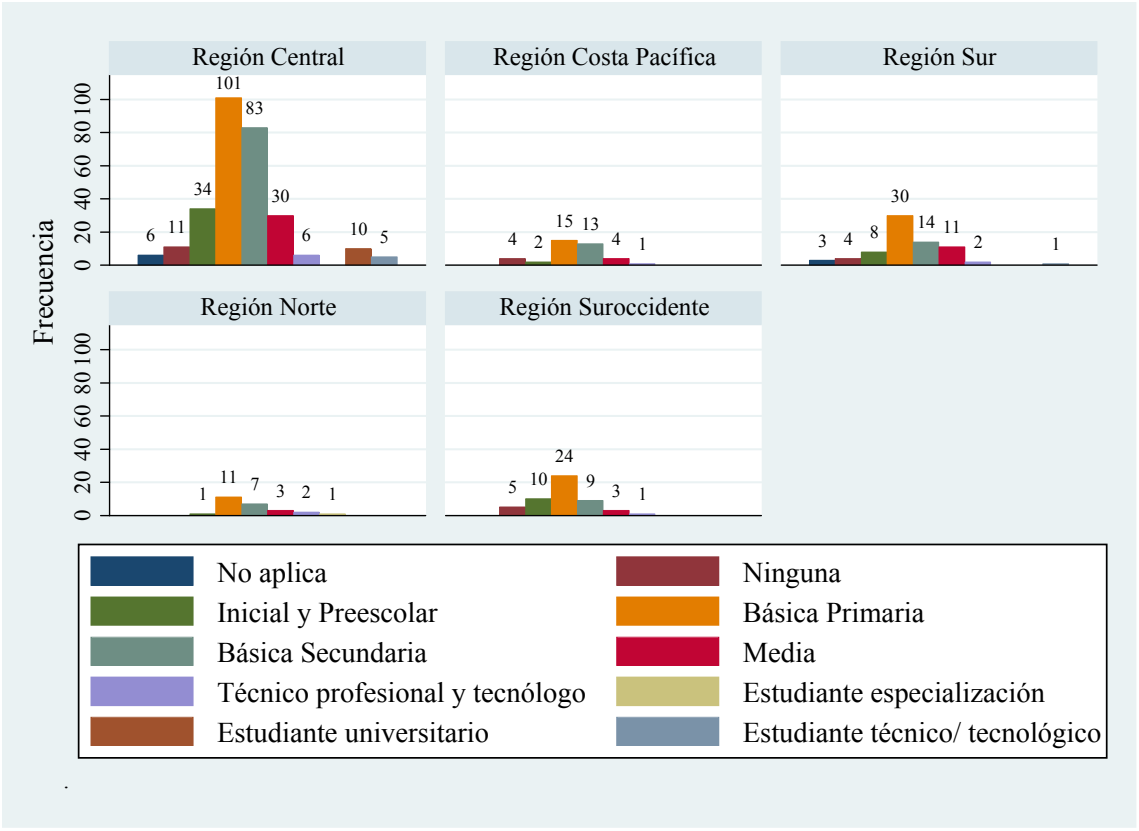

FIGURA 1.5. Delitos sexuales por región de hecho y escolaridad de la víctima.

La tabla 1.4 muestra los conteos de la figura 1.5 en términos de tasas. Observe que la región Centro es la que mayor tasa de denuncias tiene, 24,49 casos por cada 100.000 habitantes. La tasa más baja, de acuerdo a la población a riesgo, se aprecia en la región Costa Pacífica, 4,71 casos por cada 100.000 habitantes, algo similar se muestra en la zona Norte, 5,08 denuncias por cada 100.000 habitantes. Por último, las regiones Sur y Suroccidente presentan tasas de 12,86 y 11,45 respectivamente, por cada 100.000 habitantes.

Para obtener la relación de tasas, se define la **Razón de Tasa** (RT), así:

$$RT = Tasa_i / Tasa_j, \text{ siempre que } i \neq j. \tag{1.2}$$

TABLA 1.4. Tasas de denuncias por región de hecho.

Región	Población a Riesgo	Número de Casos	Tasa
Centro	1167806	286	24,49
Sur	567785	73	12,86
Norte	492273	25	5,08
Costa Pacífica	828233	39	4,71
Suroccidente	454037	52	11,45

De acuerdo a la ecuación (1.2), el riesgo de violencia por delitos sexuales en la región Centro es 1,90 veces superior al riesgo de violencia por delitos sexuales en la región Sur, es decir,

$$RT = \frac{Tasa_{centro}}{Tasa_{sur}} = \frac{24,49}{12,86} = 1,90.$$

De igual manera, el riesgo de violencia por delitos sexuales en la región Centro, dada la escolaridad de la víctima, es: 4,82 veces superior al de la región Norte, 5,2 superior a la región de la Costa Pacífica y 2,14 superior al de la región Suroccidente. Por lo tanto, teniendo en cuenta estas denuncias existe mayor riesgo de que una persona sufra un evento de este tipo en la región Centro, que en las otras regiones del departamento.

Continuando con el análisis, la figura 1.6 representa el número de denuncias de delitos sexuales por región de acuerdo a la edad de la víctima. Nuevamente como en anteriores casos, la región Centro es en donde más denuncias se presentan.

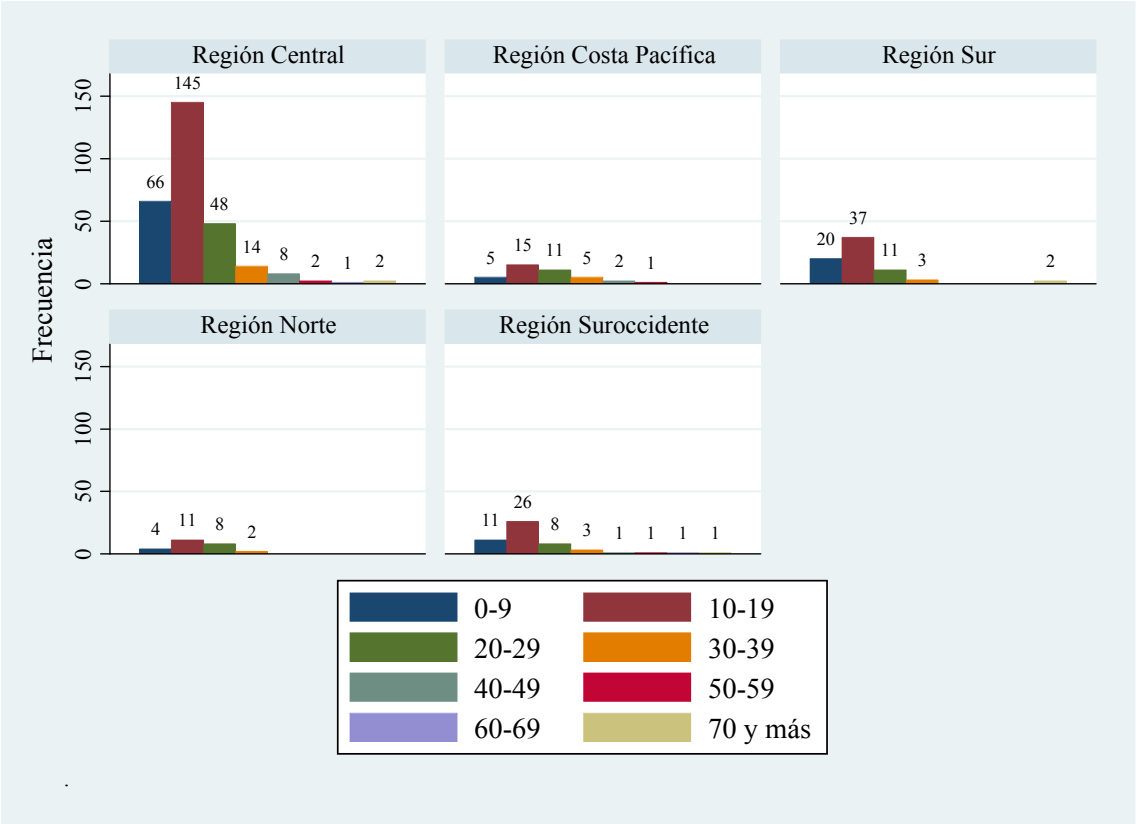

FIGURA 1.6. Delitos sexuales por región de hecho y edad de la víctima.

La región con menor cantidad de denuncias en este caso es la región Norte. La región Centro cuenta con 286 denuncias de un total de 475, es decir que en esta región se presentó el $60,21\,\%$ del total de denuncias sobre delitos sexuales en esta temporada. Le sigue la región Sur con 73 denuncias, es decir, con el $15,36\,\%$ de las ocurrencias. En la región Suroccidente se mostraron 52 casos, lo que equivale al $10,94\,\%$. La región de la Costa Pacífica tiene 39 casos, lo que expresa el $8,21\,\%$ de las denuncias. Finalmente la región Norte fue el lugar donde menos denuncias se interpusieron, solo 25 casos en los dos años, lo que representa un $5,26\,\%$ del total de denuncias sobre delitos sexuales en todo el departamento de Nariño.

De otro lado, la edad con mayor número de casos es la categoría entre 10 y 19 años. En cada una de las cinco regiones, las personas con edades entre 10 y 19 años fueron las que mayor riesgo de sufrir un delito sexual presentaron, esto indica que se debe proteger mucho mas a este tipo de personas, estas posibles víctimas con 234 casos representan el $49,26\,\%$ con respecto al resto de edades, le sigue la edad entre 0 y 9 años, con 106 casos en toda la temporada, es decir, con un porcentaje de $22,31\,\%$. Observe que la edad entre 0 y 19 años suma aproximadamente el $71,57\,\%$ del total de las denuncias presentadas ante Medicina Legal por presunto delito sexual. Por lo anterior, se sigue que la edad es un factor de riesgo a tener en cuenta para establecer políticas sociales que permitan intervenir al máximo este tipo de eventos no deseados.

La tabla 1.5 muestra las tasas de cada grupo de edad y región del evento, con base en la población a riesgo. Con las tasas se puede dar una mejor lectura e interpretación de los resultados presentados ya que éstas se basan en la población susceptible de padecer el evento. Esto significa que entre mayor sea la población a riesgo en una determinada región o grupo de edad, existe mayor probabilidad de ocurrir el hecho en esa población.

TABLA 1.5. Tasas delitos sexuales por edad de la víctima y región de hecho.

Grupos de Edad	R. Centro Tasa	R. Sur Tasa	R. Norte Tasa	R. Pacífica Tasa	R. Suroccidente Tasa
0 a 9 años	37,85	20,77	4,71	2,36	12,58
10 a 19 años	73,50	36,69	12,41	8,02	29,35
20 a 29 años	24,40	12,00	9,74	7,52	10,90
30 a 39 años	7,44	3,54	2,74	4,46	4,99
40 a 49 años	5,24	0,00	0,00	2,73	2,04
50 a 59 años	4,24	0,00	0,00	2,12	2,60
60 a 69 años	1,32	0,00	0,00	0,00	3,40
70 años y más	3,16	6,32	0,00	0,00	3,61
Promedio Región	**19,64**	**9,92**	**3,70**	**3,40**	**8,68**

La tabla anterior muestra, con respecto a la variable región, que el riesgo de presentarse un delito sexual en la región Centro es en promedio superior a todas las cuatro regiones restantes. Por otro lado, teniendo en cuenta la edad, el grupo de edad con mayor riesgo lo tienen aquellas personas que están entre los 10 y 19 años; el menor riesgo se presenta en personas de 40 años en adelante. Observe que el riesgo de ocurrir un delito sexual a personas con edades entre 10 y 19 años es 2,33 veces superior que a personas con edades entre 0 y 9 años. De igual forma, el peligro de ocurrir un posible delito sexual a personas entre 10 a 19 años es 2,7 veces superior que a individuos con rango de edad de 20 a 29 años. También se puede apreciar que en la región Norte, el riesgo de sufrir un delito sexual

en individuos mayores de 40 años es nulo. Algo análogo se muestra en la región Sur, donde el riesgo es cero en personas mayores de 40 años pero menores que 70. En la región Costa Pacífica el riesgo es nulo en población con edad igual o superior a los 60 años.

Sexo de la víctima y año del evento.

La tabla 1.6 presenta el número de denunciados sobre delitos sexuales por sexo de la víctima y año del evento a nivel de todo el departamento. Teniendo en cuenta que la población a riesgo de hombres y mujeres en la temporada *2015* en Nariño, de acuerdo a la página oficial del DANE, fue de 875.449 y 868.779 respectivamente, las tasa sobre denuncias de estos eventos es de 3,54 por 100.000 habitantes para víctimas hombres y 29,0 por cada 100.000 habitantes para víctimas mujeres. Esto significa que en esta temporada y de acuerdo a las denuncias, las mujeres tuvieron un riesgo 8,19 veces superior que los hombres de ser blanco de estos hechos.

En cuanto al año *2016*, la población total masculina en Nariño fue de 886.341 habitantes, luego la tasa de denuncias por delitos sexuales es de 1,13 por cada 100.000 habitantes para víctimas varones. De igual forma, la población de mujeres en *2016* fue de 879.565, así que la tasa de denuncias para víctima mujer es de 20,69 por cada 100.000 habitantes. Por tanto, la razón de tasa (RT) con respecto al sexo para esta temporada indica que las mujeres tuvieron un riesgo 18,31 veces superior que los hombres de ser objeto de un delito sexual.

Ya para terminar el análisis de la tabla 1.6 y conociendo que la población total para el año *2015* y *2016* en el departamento de Nariño fue de, 1.744.228 habitantes y 1.765.906 habitantes, respectivamente, se puede afirmar que la tasa general de denuncias por delitos sexuales en la primera temporada fue de 16,23 casos por cada 100.000 habitantes. En la segunda temporada la tasa fue de 10,87 casos por cada 100.000 habitantes. De aquí se deduce y de acuerdo a la RT, que el riesgo de presentarse un caso sobre delito sexual en el año *2016* estuvo $33,03\,\%$ por debajo de los casos presentados en la temporada inmediatamente anterior. Además, un detalle importante de conocer en los 41 casos presentados en hombres, es que en 22 oportunidades la víctimas fueron niños entre 0 y 9 años de edad.

Tabla 1.6. Delitos sexuales por sexo de la víctima y año del evento.

Sexo de la víctima	**Año del hecho**		
	2015	2016	Total general por sexo
Hombre	31	10	41
Mujer	252	182	434
Total general por año	283	192	475

Escolaridad de la víctima y rango de edad.

Hablando de otro aspecto, la tabla 1.7 muestra la incidencia que tiene la Escolaridad de la víctima y la Edad en la posibilidad de sufrir un delito sexual. Observe que independientemente de la edad de la víctima, la población de estudiantado de Técnico profesional en adelante, solo representa el $6,10\,\%$ de los sucesos presentados de violencia sobre delitos sexuales, algo diferente a lo que se muestra en población de estudiantado de Básica y Media, donde se encuentra el mayor porcentaje, esto sin tener en cuenta los conteos presentados en población con ningún grado de escolaridad, que son 24 y los hechos para los cuales no aplica que son 9. Con respecto a la edad, la personas entre 0 y 29 años, que

corresponde a los grupos de edad I, II y II, recogen aproximadamente el 90 % de las denuncias presentadas para las temporadas *2015* y *2016*. La categoría de edad I corresponde al grupo de 0-9 años, la categoría II corresponde al grupo de edad de 10-19 años; de esta forma hasta la categoría VII, que representa la edad de 70 años en adelante.

Tabla 1.7. Denuncias delitos sexuales por escolaridad de la víctima y rango de edad.

Escolaridad de la víctima	**Rango de edad de la víctima**								
	I	II	III	IV	V	VI	VII	VIII	**Total**
No aplica	9	0	0	0	0	0	0	0	9
Ninguna	13	1	3	0	2	1	0	4	24
Inicial y Preescolar	33	12	3	2	1	2	1	1	55
Básica Primaria	51	100	16	11	1	1	1	0	181
Básica Secundaria	0	96	22	5	3	0	0	0	126
Media	0	24	21	3	3	0	0	0	51
Técnico profesional y tecnólogo	0	1	7	3	1	0	0	0	12
Estudiante especialización	0	0	0	1	0	0	0	0	1
Estudiante universitario	0	0	9	1	0	0	0	0	10
Estudiante técnico/tecnológico	0	0	5	1	0	0	0	0	6
Total	106	234	86	27	11	4	2	5	475

Ocupación de la víctima y rango de edad.

La ocupación de la víctima es otro factor importante de resaltar en los casos presentados sobre denuncias de delitos sexuales, para esta variable al igual que para la escolaridad, no es posible determinar tasas debido a que no se cuenta con la población a riesgo. Observe en la tabla 1.8 que la categoría estudiante en todas las edades acumula el $62,10\,\%5$ de las ocurrencias presentadas, aunque las edades en la cuales se presentan es solo entre 0 y 29 años. Los estudiantes que cursan educación Básica secundaria y Media son los de mayor riesgo de sufrir un evento de este tipo, con 188 casos; se afirma que son estudiantes de este nivel por el rango de edad en los cuales se encuentran. Ahora bien, las restantes categorías de la Ocupación de la víctima, las más representativas después de estudiantes es: Ninguna, Ama de casa/encargado (a) del hogar, Oficios varios y No aplica, con 123 casos, lo que equivale al $25,89\,\%$.

De igual manera, como se aprecia en la tabla 1.8 y teniendo en cuenta análisis anteriores, existe una relación entre la categoría 10 a 19 años de la variable edad y la categoría estudiante de la variable ocupación. Esto permite describir que la edad entre 10 y 19 anida en su gran mayoría a estudiantes de Básica secundaria y Media que han sido víctima de delitos sexuales, es decir, las denuncias presentadas sobre delitos sexuales en población con grupo de edad de 10 a 19 años, en su gran mayoría son estudiantes de Bachillerato por lo que muestra la tabla.

Por lo anterior, es importante cuidar de la población que se encuentra en las escuelas y colegios para no sean objeto de hechos como los delitos sexuales. Además, no se puede desconocer que en algunas oportunidades el presento agresor de estos casos se encuentra compartiendo el mismo techo que la víctima, en este sentido los responsables del cuidado de los menores son los padres y por ende no se debe descuidarlos o dejarlos con personas que no inspiren confianza o protección de los mismos.

También se aprecia en la tabla 1.8, que en tercer orden decreciente, las empleadas del hogar son quienes más vulnerabilidad tienen de sufrir este tipo de eventos, con 42 casos denunciados, lo que representa un $8,84\,\%$.

Tabla 1.8. Denuncias delitos sexuales por ocupación de la víctima y rango de edad.

Ocupación de la víctima	**Rango de edad de la víctima**								
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70 y más	**Total**
Estudiante	82	188	25	0	0	0	0	0	295
Ninguna	10	19	10	4	2	2	0	0	47
Ama de casa/encargado (a) del hogar	0	10	15	8	1	2	1	5	42
Oficios varios	0	6	8	4	2	0	0	0	20
No aplica	14	0	0	0	0	0	0	0	14
Agricultores y trabajadores pesqueros	0	3	2	0	0	0	1	0	6
Sin ocupación específica	0	2	2	1	1	0	0	0	6
Trabajadores independientes	0	2	1	0	3	0	0	0	6
Soldados de las fuerzas militares	0	1	3	0	0	0	0	0	4
Auxiliares y administrativos y afines	0	0	3	0	0	0	0	0	3
Otras ocupaciones no clasificadas	0	0	2	1	0	0	0	0	3
Personal doméstico, aseadores, lavanderos y afines	0	0	1	1	1	0	0	0	3
Asesores comerciales	0	0	1	1	0	0	0	0	2
Empleados de servicios de información a clientes	0	0	2	0	0	0	0	0	2
Encargado (a) del hogar - rol del hogar	0	1	0	1	0	0	0	0	2
Personal de servicios de restaurante y bebidas	0	0	0	2	0	0	0	0	2
Vendedores de tiendas, almacenes y afines	0	0	2	0	0	0	0	0	2
Agentes comerciales y corredores	0	0	1	0	0	0	0	0	1
Agentes de la administración pública	0	0	0	1	0	0	0	0	1
Arquitectos, ingenieros y afines	0	0	0	0	1	0	0	0	1
Cajeros, taquilleros y afines	0	0	1	0	0	0	0	0	1
Encargados de servicio de apoyo a la producción	0	0	0	1	0	0	0	0	1
Limpiabotas y trabajadores de la calle	0	0	0	1	0	0	0	0	1
Profesionales en ciencias de la salud y afines	0	0	1	0	0	0	0	0	1
Profesionales de la educación	0	0	0	1	0	0	0	0	1
Profesionales del derecho	0	0	1	0	0	0	0	0	1
Representantes comerciales y técnicos de ventas	0	0	1	0	0	0	0	0	1
Técnicos y asistentes de la medicina y la salud	0	0	1	0	0	0	0	0	1
Técnicos no universitarios y asistentes de medicina	0	0	1	0	0	0	0	0	1
Trabajador (a) sexual	0	0	1	0	0	0	0	0	1
Trabajadores de los cuidados personales y afines	0	1	0	0	0	0	0	0	1
Vendedores de kioscos y puestos fijos de mercado	0	1	0	0	0	0	0	0	1
Vigilantes y celadores	0	0	1	0	0	0	0	0	1
Total	106	234	86	27	11	4	2	5	475

Convivencia de la víctima con el agresor.

De los 283 casos presentados en el *2015* sobre delitos sexuales, en 58 oportunidades la víctima convivía con el agresor, lo que equivale al $20,49\,\%$. De estos 58 casos, 45 son población entre 2 y 15 años, lo que indica que es importante conocer a cabalidad con quien se encarga los menores de edad. En *2016*, 30 casos de los 192 hechos presentados sobre delitos sexuales, la víctima convivía con el presunto agresor, lo que equivale al $15,62\,\%$.

De estos 30 casos, 20 son víctimas entre 3 y 15 años de edad. Por tanto, en total en las dos temporadas, de los 475 casos presentados sobre denuncias de delitos sexuales, en 75 oportunidades la víctima convivía con el presunto agresor, lo que representa el $15,78\,\%$.

Presunto agresor de la víctima.

Ya para finalizar este capítulo, la tabla 1.9 muestra la información del presunto agresor junto con el sexo del mismo. Se puede observar que alrededor del $97,68\,\%$ el agresor es hombre. Con respecto a la categoría del agresor, en 114 casos éste tiene un tipo de vínculo familiar con la víctima, es decir, en un porcentaje igual al $30,31\,\%$, el presunto agresor es muy cercano a la víctima. Observe que el $18,95\,\%$ de los agresores no fueron identificados, esto puede indicar entre otras cosas, un desconocimiento total del posible agresor o temor de la víctima a denunciar al víctimario con nombre propio.

TABLA 1.9. Presunto agresor de los casos denunciados sobre delitos sexuales.

Presunto agresor	**Sexo del agresor**			
	Hombre	Mujer	Sin información	**Total**
Familiar consanguíneo	97	0	0	97
No identificado	90	3	0	93
Conocido	83	2	2	87
Conocido sin trato	82	3	0	85
Pareja o expareja	49	0	0	49
Familiares civiles	42	1	0	43
Miembro de grupos al margen de la ley	8	0	0	8
Familiar consanguíneo o civil	5	0	0	5
Encargado del niño (a) o adolescente	3	0	0	3
Bandas criminales	3	0	0	3
Miembros de seguridad privada	1	0	0	1
Miembro de un grupo de la delincuencia organizada	1	0	0	1
Total	464	9	2	475

CAPÍTULO 2

Modelos de Regresión para Datos de Recuento

En la primera sección de este capítulo se dan a conocer algunos conceptos teóricos que serán necesarios para comprender el desarrollo del mismo. Posteriormente se presenta un estudio teórico de los modelos de regresión para datos de recuento, entre ellos: Modelo de Regresión de Poisson (MRP), Modelo de Regresión Binomial Negativo (MRBN) y modelo de regresión Inflado con Ceros, este último, tanto en versión Poisson como Binomial Negativo. Los modelos de regresión para datos de recuento son de naturaleza multiplicativa, es decir, el valor esperado es una combinación productoria de cada parámetro con su variable explicativa, lo que no sucede por ejemplo en los modelos lineales donde la variable respuesta sigue una distribución normal.

2.1. Conceptos básicos

La variable aleatoria observada en este trabajo es de tipo discreto, es decir, variable que solo toma valores enteros positivos y el cero. Las variables explicativas en los modelos de regresión para nuestro caso, son de tipo categórico. Además, se incluye una variable numérica discreta de control N, denominada *exposición* o variable *offset*.

Definición 2.1 (Variable Aleatoria)**.**
Si S un espacio muestral sobre el que se encuentra definida una función de probabilidad y Y es una función de valor real definida sobre S tal que transforma los resultados de S en puntos sobre la recta de los reales, entonces Y es una variable aleatoria.

Definición 2.2 (Variable Aleatoria Discreta)**.**
Una variable aleatoria Y se dice que es discreta si el número de valores que puede tomar es numerable, ya sea finito o infinito. En este caso la observación de la variable se hace por recuento y no por medición.

Cuando la variable aleatoria no toma valores enteros, sino valores reales, entonces la variable aleatoria no es discreta y se denomina variable aleatoria continua.

Un ejemplo de variable aleatoria discreta, es el número de hijos que puede tener una pareja: 0, 1, 2, 3, ..., 10; en este caso, cada valor que toma la variable aleatoria tiene asociada una probabilidad, que no necesariamente es la misma, ya que no es igualmente

probable que una familia tenga un hijo a que tenga diez hijos. La suma de cada una de estas probabilidades debe dar como resultado la unidad.

En el caso anterior, se habla de variable aleatoria numérica, sin embargo, existen también variables aleatorias categóricas, es decir, aquellas variables que no toman valores numéricos sino categorías, por ejemplo, el sexo del número de hijos que tiene una pareja es un tipo de esta variable, ya que puede ser Hombre o Mujer. Las variables aleatorias categóricas tienen propiedades diferentes a las variables aleatorias numéricas, ya que éstas no permiten realizar aritmética sobre ellas, esto es, calcular medias, varianzas, etc.

Las variables de recuento son un tipo particular de variables discretas y solo toman valores enteros no negativos; se refieren al número de sucesos o eventos que ocurren en un intervalo temporal o espacial definido. Autores como Lindsey en [14] definen las variables de recuento como el número de eventos de una misma variable que ocurren en el mismo sujeto o unidad de observación. En este trabajo la unidad de observación es el tamaño N_i de una población en riesgo de ocurrirles el evento. La variable muestral de respuesta tipo recuento es y, donde y/N_i expresa la tasa muestral de denuncias sobre delitos sexuales en función de las las variables explicativas: sexo, edad, región y tiempo.

$$\left(\frac{\text{Número de eventos ocurridos}}{\text{Tamaño población a riesgo}}\right) = \frac{y}{N_i}\,. \tag{2.1}$$

La población a riesgo N_i que se toma en la ecuación (2.1) corresponde al tamaño de una unidad de observación, información que para nuestro trabajo se encuentra proyectada en la página oficial del DANE, de acuerdo al Censo Poblacional que esta misma entidad realizó en el año 2005. El recuento y denota el número de casos (número de denuncias por violencia sexual) en una unidad específica del vector $N = (N_1, N_2, \ldots, N_i, \ldots, N_m)^t$.

De esta forma, el valor esperado de la tasa muestral se escribe de la siguiente forma:

$$E(Y/N) = \frac{1}{N}E(Y) = \frac{\lambda}{N}\,.$$

Luego, para una componente (observación) específica de N se tiene que:

$$E(y/N_i) = \frac{1}{N_i}\lambda_i = \frac{\lambda_i}{N_i}\,.$$

Eventos Múltiples y Eventos Únicos

Es importante anotar que los modelos de regresión de Poisson, Binomial Negativo y sus extensiones de ambos modelos (modelo inflado con ceros), se utilizan para el estudio de eventos múltiples, es decir, aquellos eventos que pueden suceder mas de una vez en la misma unidad de observación. En este trabajo, la unidad de observación es un grupo de personas con características predefinidas probablemente susceptibles de padecer el riesgo de un delito sexual (riego de padecer el evento).

Por otro lado, son ejemplos de modelos para estudio de eventos únicos, el modelo de regresión Logística y el modelo de Supervivencia. La muerte es un evento de naturaleza única cuando la unidad de análisis es el individuo, sin embargo, si este evento se toma en un conjunto de personas entonces se convierte en un evento de naturaleza múltiple debido a que puede suceder más de una vez en la misma unidad de observación.

Definición 2.3 (Función Gamma)**.**
La función gamma es una extensión de la función factorial, con su argumento desplazado por 1, a numeros reales y complejos. Si n es un entero positivo se tiene:

$$\Gamma(n) = (n-1)!$$

La función gamma se define para todos los números complejos excepto los enteros negativos. Para números complejos con una parte real positiva, se define a través de la integral:

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx \,.$$

Definición 2.4 (Verosimilitud-lilkehood)**.**
Sea C un conjunto finito o numerable, $\{P_\theta\}$ una familia de leyes de probabilidad sobre C y m un entero. Llamamos verosimilitud asociada a la familia $\{P_\theta\}$, a la función que para una m-tupla $(y_1, \ldots, y_m)$ de elementos de C y un valor θ del parámetro está definida por:

$$L(\theta) = L(\theta / y_1, \ldots, y_m) = \prod_{i=1}^{m} P_\theta(y_i) \,.$$

La verosimilitud hace referencia a la credibilidad o congruencia de un elemento determinado dentro de un conjunto específico. Observe que para la función de verosimilitud se fijan los datos y se varia el valor del parámetro.

Definición 2.5 (Estimación por máxima verosimilitud-EMV)**.**
El método de estimación por máxima verosimilitud, selecciona como estimador a aquél valor del parámetro que tiene la propiedad de maximizar el valor de la probabilidad de la muestra aleatoria observada, es decir, el método de máxima verosimilitud (MV) consiste en encontrar el valor del parámetro que maximiza la función de verosimilitud. Si $f(y;\theta)$ es la función de probabilidad poblacional y θ el parámetro desconocido, se llama función de verosimilitud de la muestra $(y_1, y_2, \ldots, y_m)$ a la función de probabilidad conjunta de los valores muestrales $L(\theta; y) = f(y; \theta)$, esto es,

$$L(\theta; y_1, y_2, \ldots, y_m) = \prod_{i=1}^{m} f(y_i; \theta) \,.$$

Dado que la función $\ln(L)$, denominada log-verosimilitud, tiene las mismas propiedades que la función de verosimilitud L, el EMV de θ se estima mediante la expresión:

$$\frac{d}{d\theta}[\ln(L)] = 0 \,.$$

Resolviendo la ecuación para θ se estima el valor del parámetro $\hat{\theta}$.

Definición 2.6 (Pseudo R^2 de McFadden - 1974)**.**
Sean ℓ_o y ℓ_1 las funciones log-verosimilitud del modelo nulo (modelo sólo con la constante) y del modelo ajustado, respectivamente. Considérese además que las funciones ℓ_o y ℓ_1 corresponden a los modelos de regresión Poisson o Binomial Negativo. Se define el Pseudo R^2 de McFadden para los modelos mencionados como sigue:

$$Rseudo\, R^2 = 1 - \frac{\ell_1}{\ell_o} \,.$$

Su valor esta en el intervalo $[0, 1]$ y se interpreta como la proporción de reducción de la discrepancia del modelo nulo debido a la inclusión de variables explicativas en el modelo ajustado. En ocasiones este estadístico sirve como prueba de bondad de ajuste (ver [2]).

2.2. Regresiones de Poisson y Binomial Negativa

En esta sección se presenta el modelo de regresión de Poisson y el modelo de regresión Binomial Negativo, junto con las distribuciones de probabilidad para cada uno de los modelos. Se hace mayor énfasis en el modelo de regresión de Poisson por ser uno de los modelos por excelencia para datos de recuento, sin embargo, en ocaciones el supuesto de equidispersión no se cumple y utilizar este modelo para ajustarlo a un conjunto de datos sería estar lejos de una aproximación real, en tal caso es mejor recurrir a otros modelos de regresión o una extensión de éste para corregir fenómenos tales como, exceso de ceros y sobredispersión en la variable respuesta.

2.2.1. Modelo de Regresión de Poisson

Distribución de Poisson

La distribución de Poisson (ley de los sucesos raros), llamada así e honor a Simeon Denis Poisson, probabilista frances del siglo XIX (1781-1840), representa la probabilidad de que un evento aislado (o variable aleatoria discreta) ocurra un número específico de veces en un intervalo de tiempo, espacio o distancia, dado un promedio por unidad de medida. Más específicamente, si Y es una variable aleatoria que representa el número de eventos aleatorios independientes que ocurren a una rapidez constante λ, sobre el tiempo o el espacio, entonces la variable aleatoria Y tiene distribución Poisson con función de probabilidad:

$$f(y;\lambda) = P(Y = y/\lambda) = \begin{cases} \dfrac{\exp(-\lambda)\lambda^y}{y!} & y = 0, 1, 2, 3, \ldots \qquad \lambda > 0 \\ 0 & \text{para cualquier otro valor.} \end{cases}$$

El parámetro de la la distribución de Poisson es λ, representa el número medio de veces que se espera que ocurra el evento en un período determinado, y es el número de veces que ocurre un evento de interés. A diferencia de la distribución normal que tiene dos parámetros, la media μ y la varianza σ^2, la distribución de Poisson solo tiene el parámetro λ. Además, si una variable aleatoria Y sigue una distribución Poisson se cumple que:

$$Var(Y) = E(Y) = \lambda,$$

propiedad conocida como *equidispersión*. Una de las características de esta distribución es que las probabilidades disminuyen a medida que el valor de la variable aleatoria aumenta.

La función de probabilidad de la distribución de Poisson, es una forma límite de la distribución Binomial (ver [8]) cuando se cumplen las siguientes condiciones:

- El número de observaciones se hace grande (n grande).

- La probabilidad se hace pequeña $p < 0,01$.
- La media de la distribución se hace constante ($np = constante$).

La distribución de Poisson tiende a la normal a medida que aumenta su media λ. Así, cuando el fenómeno de estudio no sea un evento raro y los valores de recuento tengan una frecuencia elevada, la distribución de Poisson convergirá a la distribución normal. Cuando λ es un valor pequeño, la distribución de Poisson se encuentra sesgada positivamente, sin embargo, a medida que aumenta el valor de λ la asimetría disminuye, y como se dijo anteriormente, esta distribución se aproxima a una distribución normal.

Algunos ejemplos de eventos aleatorios que se pueden expresar como una variable aleatoria de Poisson son:

- Número de accidentes de tráfico en un tramo de cierta carretera en un mes.
- Número de accidentes laborales durante un periodo de tiempo.
- Número de artículos publicados por una revista.
- Número de denuncias sobre violencia en una población de tamaño N.
- Número de árboles infectados por hectárea en un bosque.

Los dos primeros conteos son en el tiempo, los tres siguientes son conteos en el espacio.

Estimación del Parámetro

Sea $y_1, y_2, \ldots, y_m$ una muestra de m observaciones. Para encontrar el parámetro λ que maximice la función de verosimilitud para todos los valores observados y_i de Poisson, se utiliza el método de máxima verosimilitud (MV). Este procedimiento se realiza aplicando logaritmo a la función de verosimilitud $L(\lambda)$, donde $\ell(\lambda) = \ln(L(\lambda))$. La función ℓ comúnmente se la conoce con el nombre de, log-verosimilitud (en inglés log-likelihood).

$$\ell(\lambda) = \ln \prod f(\lambda) = \sum_{i=1}^{m} \ln \left(\frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \right) = -m\lambda + \ln(\lambda) \left(\sum_{i=1}^{m} y_i \right) - \sum_{i=1}^{m} y_i! \,.$$

Luego, el estimador de maxima verosimilitud de λ es el valor que haga maxima la función de verosimilitud de la muestra, por tanto se resuelve la ecuación $\ell'(\lambda) = 0$ con respecto a λ para encontrar el valor del parámetro estimado. Esto es,

$$\frac{d\ell(\lambda)}{d\lambda} = 0 \qquad \text{si y sólo si} \qquad \lambda = \left(\sum_{i=1}^{m} y_i \right) / m \,.$$

Regresión de Poisson

El Modelo de Regresión de Poisson (MRP) es un tipo de Modelo Lineal Generalizado (MLG). Los modelos lineales generalizados propuestos por Nelder y Wedderburn en 1972 (ver [20]), permiten incluir distintas relaciones entre las medidas condicionales de las variables respuesta y explicativas. Estos modelos son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (Binomiales, Poisson, Gamma, Binomial Negativa) y varianzas no constantes. Un requisito de los MLG es que

la distribución de la variable respuesta Y pertenezca a la familia exponencial. Para ello, la función de probabilidad si la variable aleatoria es discreta o la función de densidad si la variable es continua, se debe poder expresar como en (2.2).

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.2}$$

En donde, ϕ es un parámetro conocido y θ es un parámetro desconocido de la familia exponencial también llamado parámetro canónico de la función, el cual sirve en el MLG como función de enlace η. Fácilmente se puede probar que la distribución de Poisson satisface la ecuación (2.2).

El modelo de regresión de Poisson se utiliza para datos de conteo y es adecuado cuando $Var(Y) = E(Y) = \lambda$ (ver [24]). En este modelo la media λ se explica en términos de covariables mediante el siguiente enlace:

$$g(\lambda) = \ln(\lambda). \tag{2.3}$$

El MRP se caracteriza por:

- Es un modelo heterocedástico, es decir, las varianzas de las perturbaciones no son constantes, por tanto, la variabilidad es diferente para cada observación.
- Tiene la propiedad de equidispersión, esto es, $E(Y) = Var(Y)$.

El MRP se presenta cuando la variable de respuesta es una cantidad discreta que se puede modelizar con una Poisson y se quiere estudiar si ciertas variables explicativas influyen en la variable de respuesta y en que forma lo hacen. Este tipo de variable suele representar el recuento de sucesos o hechos que se presentan en un determinado fenómeno, por ejemplo, número de denuncias por delitos sexuales que se presentan en un conjunto de personas con unas características definidas. En el modelo de regresión de Poisson se analiza la variable respuesta Y a través de otras variables explicativas X mediante un análisis de regresión. Se pretende construir un modelo para

$$\lambda(x) = E(Y/X = x),$$

es decir, para la media de Y condicionada a cada valor de la variable explicativa.

La matriz de variables explicativas o matriz de diseño se presenta en la ecuación (2.4). Para incluir el intercepto en el modelo de regresión se considera la primera variable explicativa X_1 como un vector columna de unos, es decir, $X_1 \equiv 1$.

$$X = (X_1 \quad X_2 \quad \dots \quad X_p) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mp} \end{pmatrix} \tag{2.4}$$

Los modelos lineales generalizados (ver [16, 20]) presentan tres componentes que son:

Componente sitemático; Componente aleatorio y *Función de enlace.*

- **Componente sistemático:** El componente sistemático resume como la variabilidad en la respuesta es explicada por los valores de ciertas variables explicativas (independientes, predictoras o covariables) y es descrita, generalmente, mediante un modelo de regresión. Esta componente recoge la variabilidad de Y expresada a través de p variables explicativas $X_1, X_2, X_3, \ldots, X_p$, que denotaremos por X y de sus correspondientes parámetros $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^t$. La componente sistemática, también denominado predictor lineal, se simboliza con η y se expresa:

$$\eta = \ln(\lambda(X, \beta)) = X\beta, \tag{2.5}$$

donde el predictor lineal es el presentado en la ecuación (2.3), es decir,

$$\eta = g(\lambda) = \ln(\lambda),\ \lambda \in (0, +\infty).$$

En la ecuación (2.5), la expresión $X\beta$ representa el producto de la matriz X de variables explicativas por el vector columna de parámetros desconocidos β, esto significa que $X\beta$ es un vector columna de dimensión $m \times 1$. Por lo tanto, la función de regresión para un modelo lineal generalizado es

$$\lambda = E(Y) = \exp(X\beta). \tag{2.6}$$

Teniendo en cuenta las expresiones (2.4) y (2.6), para un índice fijo i en el conjunto $i \in \{1, 2, \ldots, m\}$, se tiene que:

$$\lambda_i(X_i) = \lambda_i(x_{i1}, \ldots, x_{ip}) = \exp(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip}),$$

donde $x_{i1} = 1$ es una componente del vector X_i. Linealizando la expresión anterior,

$$\ln(\lambda_i) = \beta_1 + \sum_{j=2}^{p} x_{ij}\beta_j \tag{2.7}$$

- **Componente aleatorio:** Identifica la variable respuesta (variable que se desea explicar) y su distribución de probabilidad. Este componente, también denominado error, constituye la variabilidad no explicada por el componente sistemático y describe, mediante una distribución de probabilidad, la discrepancia entre la respuesta observada y la respuesta esperada o predicha por la parte sistemática del modelo.

$$Y \sim DP(\theta),$$

esto es, la variable Y sigue una distribución de probabilidad (DP) de parámetro θ.

- **Función de enlace:** Este elemento especifica cómo el valor esperado del componente aleatorio $E(Y)$ está vinculado al componente sistemático. Para el modelo de regresión de Poisson o Binomial Negativo, se tiene que

$$g(\lambda) = \eta = \ln(\lambda) = X\beta\,, \tag{2.8}$$

donde $E(Y) = \lambda$, y la transformación (ln) es la función de enlace. En el MRP, la función de enlace transforma el valor esperado a la escala del predictor lineal.

La ecuación (2.8) es la más indicada para la función de enlace en la regresión de Poisson, debido a que en los modelos de regresión tipo recuento la variable de respuesta no toma valores negativos. De (2.8) se tiene que $g^{-1}(\eta) = \exp(\eta)$.

Variable Offset o de Exposición

La regresión de Poisson también puede ser apropiada para los datos de tasas, donde la tasa es un recuento de eventos dividido por alguna medida de la exposición de esa unidad (una unidad particular de observación). Por ejemplo, los biólogos pueden contar el número de especies arbóreas en un bosque: los eventos serían observaciones de árboles, la exposición sería un área unitaria y la tasa sería el número de especies por unidad de área. Los demógrafos pueden modelar las tasas de mortalidad en áreas geográficas como el recuento de muertes dividido por persona-años. Más generalmente, las tasas de eventos pueden calcularse como eventos por unidad de tiempo, lo que permite que la ventana de observación varíe para cada unidad. En estos ejemplos, la exposición es, respectivamente, unidad de área, persona-año y unidad de tiempo. En la regresión de Poisson esto se maneja como un desplazamiento, donde la variable de exposición entra en el lado derecho de la ecuación, pero con una estimación de parámetro (para log-exposición) limitada a 1. Luego el parámetro en la distribución de Poisson será la tasa media de un evento por unidad de exposición.

En este trabajo, la variable de exposición (control) se define como el número de individuos en riesgo de ocurrirles el evento, en cada subcategoría de las variables explicativas, es decir, N es un vector columna con tamaños de poblaciones, tales poblaciones son extraídas de la página oficial del DANE. Cuando no se de ningún valor de exposición se asume que su valor es uno (puede ser un año, un individuo, un semestre, etc). Cada valor de exposición es un número positivo, no tiene sentido que tome valor cero o negativo. La variable de exposición también puede ser una cantidad positiva constante diferente de uno, esto ocurre cuando la unidad de observación es una cantidad fija de elementos para todos los conteos.

Por lo tanto, usando esta notación la distribución de Poisson se escribe como:

$$f(y_i; N_i, \lambda_i) = P(Y = y_i/N_i, \lambda_i) = \begin{cases} \dfrac{\exp\left[N_i(-\lambda_i)\right](N_i\lambda_i)^{y_i}}{y_i!} & y_i = 0, 1, 2, 3, \ldots \qquad \lambda_i > 0 \\ 0 & \text{para cualquier otro valor.} \end{cases}$$

Observe que si $N \equiv 1$, la distribución de Poisson queda reducida a su forma tradicional.

Luego, para un $i \in \{1, 2, \ldots, m\}$, el modelo de regresión que permite obtener los valores esperados de las tasas es:

$$\lambda_i/N_i = \exp(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \tag{2.9}$$

Observe que la expresión 2.9 no modela el conteo, sino la tasa.

La ecuación 2.9 se puede escribir de la siguiente forma:

$$\ln(\lambda_i) = \ln(N_i) + \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}\,.$$

La variable *offset*, es el logaritmo de la variable de exposición, es decir,

$$Offset = \ln(N)\,.$$

En una regresión múltiple (Modelos Lineales), cuando la variable de respuesta sigue una distribución normal, $Y \sim Normal(\mu, \sigma^2)$, la función de enlace que se utiliza es el operador identidad, es decir,

$$g(\mu) = \mu.$$

Por tal motivo, en la regresión múltiple el valor esperado μ y el valor predicho η tienen la misma escala y por ende la función de enlace que se utiliza es la función identidad. Sin embargo, esto no sucede en la regresión de Poisson, donde el valor esperado $E(Y) = \lambda$ y el valor predicho η se encuentran en escalas de medidas diferentes, por ende se necesita una función de enlace como la que se presentó en (2.8).

Solución por estimación de máxima verosimilitud

Los coeficientes de regresión β se estiman utilizando el método de máxima verosimilitud (MV). El logaritmo de la función de verosimilitud es $\ell(\beta) = \ln[L(\beta)]$ esta dada por:

$$\ell(\beta) = \ln[L(\beta)] = \sum_{i=1}^{m} y_i \ln[N_i\lambda(X_i\beta)] - \sum_{i=1}^{m} N_i\lambda(X_i\beta) - \sum_{i=1}^{m} \ln(y_i!) \qquad (2.10)$$

Las ecuaciones de verosimilitud pueden formarse tomando las derivadas con respecto a cada coeficiente de regresión y estableciendo el resultado igual a cero. Hacer esto conduce a un conjunto de ecuaciones no lineales que no admiten una solución analítica. Por lo tanto, se debe usar un algoritmo iterativo para encontrar el conjunto de coeficientes de regresión que maximicen la log-verosimilitud. Por fortuna, con el uso de los computadores de hoy en día, los parámetros β se pueden estimar de forma iterativa convirtiéndose cada vez menos en un problema. Algunos paquetes estadísticos también tienen integrado para la estimación de los parámetros el algoritmo iterativo de Newton Raphson.

2.2.2. Modelo de Regresión Binomial Negativa

Distribución Binomial Negativa

Como se ha visto en el apartado anterior, a pesar de que el modelo de referencia en estudios de variables de recuento es el MRP, éste presenta varios problemas a la hora de tratar con datos en los que la media y la varianza condicionadas no coinciden, en concreto, con datos que presentan sobredispersión. La sobredispersión es común en el modelado de conteos. Cuando el modelo para la media es correcto pero la distribución verdadera no es Poisson, las estimaciones de máxima verosimilitud de los parámetros del modelo siguen siendo consistentes, pero los errores estándar son incorrectos. Una forma de relajar la restricción de igualdad media-varianza del MRP, es especificar una distribución que permita un modelado más flexible, esa distribución se denomina binomial negativa (BN). Esta distribución se usa para predecir el número de fallas antes de que ocurra un éxito en una secuencia de ensayos de Bernoulli. Además, se usa para modelar una distribución que muestra una variación excesiva porque algunas categorías muestran un exceso y otras una deficiencia de los recuentos. La distribución binomial negativa es una mezcla de Poisson

y la distribución Gamma. En este sentido, el modelo paramétrico estándar para datos de recuento con presencia de sobredispersión es el Modelo de Regresión Binomial Negativo.

La sobredispersión no es un problema en la regresión ordinaria con Y distribuido normalmente, porque esa distribución tiene un parámetro separado (la varianza) para describir la variabilidad (ver [2]).

Supóngase que hay una secuencia $n = y + r$ de ensayos independientes de Bernoulli, donde cada ensayo tiene dos resultados, *éxito* y *fracaso*. En cada ensayo la probabilidad de éxito es π y de fracaso es $(1 - \pi)$. Se observa esta secuencia hasta que se ha producido un número r de fallos predefinidos (no aleatorios). Entonces el número aleatorio de éxitos que se han observado, $Y = y$, seguirá una distribución binomial negativa $Y \sim BN(r, \pi)$, donde la función de probabilidad f es:

$$f(y; r, \pi) = P(Y = y/r, \pi) = \begin{cases} \binom{y + r - 1}{r - 1} \pi^y (1 - \pi)^r & y = 0, 1, \ldots \quad r > 0 \\ \\ 0 & \text{para cualquier otro valor.} \end{cases}$$

El valor esperado y la varianza vienen dados por:

$$E(Y) = \frac{\pi r}{1 - \pi} \qquad \text{y} \qquad Var(Y) = \frac{\pi r}{(1 - \pi)^2}, \tag{2.11}$$

estableciéndose entre las dos expresiones la relación,

$$Var(Y) = \frac{1}{1 - \pi} E(Y).$$

Los parámetros de la distribución binomial negativa son r y π, donde $0 < \pi < 1$ y $r > 0$. Luego, como $0 < \pi < 1$ se verifica que $Var(Y) > E(Y)$. Esto justifica la aptitud natural de esta distribución para modelar datos que se caracterizan por la existencia de sobredispersión.

Ahora, si r es entero la distribución anterior se conoce como distribución de Pascal, pero si r no es entero, la función de probabilidad se escribe de manera que involucre la función Gamma (ver [8]). Aunque es imposible visualizar un número no entero de fallas, la distribución BN se puede parametrizar en términos de la función Gamma, para ello se extiende el coeficiente binomial como:

$$\binom{y + r - 1}{r - 1} = \frac{(y + r - 1)!}{y!(r - 1)!} = \frac{(y + r - 1)(y + r - 2) \cdots (r)}{y!} = \frac{\Gamma(y + r)}{y! \Gamma(r)}. \tag{2.12}$$

De esta forma, utilizando la expresión (2.11) se tiene que:

$$E(Y) = \frac{r\pi}{1 - \pi} = \lambda,$$

de aquí se se obtiene,

$$\pi = \frac{\lambda}{r + \lambda} \qquad \text{y} \qquad 1 - \pi = \frac{r}{r + \lambda}. \tag{2.13}$$

Luego, expresando la varianza de Y en términos de r y λ tenemos:

$$Var(Y) = \lambda + \frac{\lambda^2}{r},$$

donde λ es la varianza de Poisson y λ^2/r es la varianza de Gamma. Claramente se observa que cuando r tiende hacia ∞, la distribución binomial negativa se aproxima a la distribución de Poisson. Luego, empleando las ecuaciones (2.12) y (2.13) se puede volver a escribir la distribución binomial negativa tal que (ver [2]):

$$f(y; r, \lambda) = P(Y = y/r, \lambda) = \frac{\Gamma(y+r)}{y!\Gamma(r)} \left(\frac{\lambda}{r+\lambda}\right)^y \left(\frac{r}{r+\lambda}\right)^r \tag{2.14}$$

Haciendo $\alpha = 1/r$, la varianza de la distribución BN es

$$Var(Y) = \lambda + \alpha\lambda^2,$$

donde α es denominado parámetro de dispersión. Observe que si $\alpha = 0$, se cumple el supuesto de equidispersión y el modelo se reduciría a la distribución de Poison. La expresión (2.14) representa lo que se denomina distribución binomial negativa tradicional, ésta es una combinación de distribuciones de Poisson donde la frecuencia aleatoria de ocurrencia tiene una distribución Gamma y cuya media es igual a la media de Poisson.

La aplicación de la distribución binomial negativa es una alternativa adecuada para el modelo de Poisson cuando la frecuencia de ocurrencia no es constante sobre el tiempo, el espacio, el tamaño de una población, etc. Un evento que se distribuye como una variable aleatoria en un modelo Binomial Negativo es: número de infracciones recibidas por un automovilista en un periodo de 5 años. Aquí se dice que la variable aleatoria Y sigue una distribución Binomial Negativa y no una de Poisson, porque los eventos no ocurren con una velocidad constante, puesto que si una persona recibe un comparendo por infringir una norma de tránsito, es razonable que conducirá luego con más cuidado. Esto hace que se genere sobredispersión en la variable de respuesta.

Ahora bien, desde el punto de vista de este trabajo, la distribución binomial negativa no debe tomarse en términos de cuantos ensayos se necesitan para alcanzar un determinado número de éxitos, sino mas bien, como el número de denuncias por delitos sexuales que ocurren en una población cuando la frecuencia de éstos no es constante.

Estimación de Parámetros

La estimación de los parámetros α y λ se realiza por el método de máxima verosimilitud (MV). El estimador de máxima verosimilitud solo existe para muestras en las cuales la varianza es mayor que el valor esperado. La función de verosimilitud para una muestra de m observaciones corresponde a:

$$L(\alpha, \lambda) = \prod f(\alpha, \lambda)\,.$$

A partir de la cual se calcula la función $\ell = \ln(L)$, denominada log-verosimilitud, la cual alcanza su valor máximo en los mismos puntos que la función L.

$$\ell = \sum_{i=1}^{m} \ln[\Gamma(y_i+\alpha^{-1})] - \sum_{i=1}^{m} \ln(y_i!) - m\ln[\Gamma(\alpha^{-1})] + \sum_{i=1}^{m} y_i \ln\left(\frac{\lambda}{\alpha^{-1}+\lambda}\right) + m\alpha^{-1}\ln\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right).$$

Luego, para hallar el máximo se toman las derivadas parciales con respecto a los parámetros α y λ, posteriormente se igualan a cero, es decir:

$$\frac{\partial \ell(\alpha,\lambda)}{\partial \lambda} = 0 \qquad \text{y} \qquad \frac{\partial \ell(\alpha,\lambda)}{\partial \alpha} = 0$$

La segunda ecuación no se puede resolver para α en forma analítica. Si se desea una solución numérica, se puede usar una técnica iterativa como el método de Newton Raphson.

Regresión Binomial Negativa

El Modelo de Regresión Binomial Negativo (MRBN) también pertenece a la familia de los MLG (ver [16]), es adecuado cuando los datos presentan sobredispersión, es decir, cuando la varianza es mayor que el valor esperado. El MRBN modela la heterogeneidad de Poisson con una distribución Gamma. De esta forma, cuando el supuesto de equidispersión en el MRP no se cumple, se presentan dos casos: primero $Var(Y) > E(Y) = \lambda$, fenómeno denominado como sobredispersión y siendo éste el que se presenta con mayor frecuencia; segundo caso $Var(Y) < E(Y) = \lambda$, fenómeno denominado como infradispersión. En este sentido, si $Var(Y) > E(Y) = \lambda$, la variable respuesta tipo recuento sigue mejor una distribución Binomial Negativa ($Y \sim BN$) que una Poisson (ver [11]).

Según Boswell y Patil (*1970*), se han identificado 13 tipos de derivaciones para la distribución binomial negativa, otros estadísticos argumentan que existen más. En este informe, dado que se toma la distribución binomial negativa tradicional, entonces se trabaja con el modelo tradicional comúnmente conocido como $NB2$, en el cual la distribución Gamma es usada para ajustar los datos Poisson que presentan sobredispersión (ver [6, 11]).

Por lo tanto, teniendo en cuenta lo anterior y la matriz (2.4) de la sección 2.2.1, para todo $i \in \{1, 2, \ldots, m\}$, el modelo de regresión binomial negativo es:

$$\ln(\lambda_i) = \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij},$$

esto implica que la media de Y está determinada por

$$\lambda_i = \exp\left(\beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}\right).$$

Solución por estimación de máxima verosimilitud

Los coeficientes de regresión β y α se estiman utilizando el método de máxima verosimilitud (MV). El logaritmo de la función de verosimilitud $\ln L(\beta,\alpha) = \ell(\beta,\alpha)$ es:

$$\begin{aligned}\ell(\beta,\alpha) = &\sum_{i=1}^{m}\{\ln[\Gamma(y_i+\alpha^{-1})] - \ln[\Gamma(\alpha^{-1})] - \ln[\Gamma(y_i+1)] - \alpha^{-1}\ln(1+\alpha\lambda(X_i\beta))\} \\ &+ \sum_{i=1}^{m}\{-y_i\ln[1+\alpha\lambda(X_i\beta)] + y_i\ln(\alpha) + y_i\ln[\lambda(X_i\beta)]\}.\end{aligned}$$

Los valores de α y β que maximizan $\ell(\alpha,\beta)$ son las estimaciones de verosimilitud que se buscan. Los estimadores de α y β están incorrelacionados, es decir, $Cov(\hat{\beta},\hat{\alpha}) = \vec{0}$.

2.3. Modelos de Regresión Inflados con Ceros

Cuando la variable respuesta es un conteo los datos observados se deben modelar estadísticamente con distribuciones discretas como la Poisson y la binomial negativa. Sin embargo, no es raro que el número de ceros observados en la variable respuesta exceda a la frecuencia que se espera observar bajo la distribución que se ajusta. En este caso se dice que los datos presentan exceso de ceros o están inflados con ceros. Este fenómeno no debe ignorarse ni tampoco debe ajustarse a distribuciones que no consideren el exceso de ceros, pues el hecho de no tener en cuenta los ceros adicionales puede dar lugar a estimaciones de parámetros sesgados e inferencias engañosas.

En este sentido, se han desarrollado modelos estadísticos que describen este fenómeno permitendo derivar conclusiones realistas y confiables a partir de las inferencias. Para modelar el exceso de ceros es fundamental entender la naturaleza de su origen. De acuerdo a lo anterior, los ceros se clasifican en dos tipos: *ceros estructurales* y *ceros muestrales* [10]. A manera de ejemplo, suponga que se observa el número de denuncias por delitos sexuales en una determinada población. Esta población podría no presentar denuncias por tener factores protectores a la ocurrencia del evento, luego en esta población siempre observaremos cero denuncias, lo que significa que se tiene un cero estructural ($Y = 0$). Por otro lado, considérese que la población observada si es vulnerable a un posible delito sexual, pero al momento de observarla no sucedió el evento o la víctima no denunció el caso, así que esta población no presenta riesgo y observamos un cero ($Y = 0$). Luego en esta situación se presenta un cero muestral. Así, un cero inevitable es un cero estructural, mientras que un cero que ocurre debido al mecanismo de muestreo es un cero muestral.

Por tal motivo, se han desarrollado modelos con inflado de ceros que consideran diversos escenarios, es decir, tanto ceros muestrales como estructurales en la variable respuesta. Entre estos modelos, el Modelo de Regresión de Poisson Inflado con Ceros, conocido como modelo ZIP y propuesto por Lamber en [12], el cual presenta un mejor ajuste a los datos que el MRP cuando la variable explicada presenta un número elevado de ceros. Sin embargo, el modelo ZIP no es apropiado cuando la parte no nula de la distribución esta sobredispersa con respecto a la distribución de Poisson. Entonces cuando la variable respuesta en un modelado presenta exceso de ceros y sobredispersión, el modelo mas apropiado que recomiendan algunos autores es el Modelo de Regresión Binomial Negativo Inflado con Ceros, denominado en esta literatura como modelo ZINB (ver [11, 15, 19]). A pesar de ello, se ha encontrado que el modelo ZINB en algunas ocasiones no converge y por ende su uso es restringido [10]. De acuerdo con esto, en el año *2006* los autores de [10] proponen un Modelo de Regresión de Poisson Generalizado Inflado con Ceros, llamado ZIGP, el cual es un buen competidor del modelo ZINB. Recientemente Rodríguez y Jiménez en [22] desarrollaron un modelo de regresión para datos sobredispersos con un número reducido de ceros en la variable respuesta, con base en la distribución biparamétrica compleja de Pearson (CBP).

Los Modelos de Regresión ZIP y ZINB no hacen parte de los MLG.

2.3.1. Regresión de Poisson Inflado con Ceros

La distribución de Poisson es un modelo de probabilidad ampliamente utilizado en la descripción de los datos relacionados con el recuento que surgen con frecuencia en disciplinas tales como Seguros, Salud pública, Epidemiología, Psicología y muchas otras áreas de

investigación. Cuando hay una gran cantidad de recuentos cero en los datos, hay evidencia de dispersión excesiva y la distribución ordinaria de Poisson no es un modelo apropiado porque a menudo subestima la dispersión observada. Generalmente se determinan dos fuentes de sobredispersión: la heterogeneidad de la población y el exceso de ceros. La heterogeneidad se observa cuando la población se puede dividir en muchas subpoblaciones homogéneas. El exceso de ceros se detecta cuando el número de ceros observados excede en gran medida el número de ceros reproducidos por la distribución de Poisson.

En este sentido, la distribución de probabilidad del modelo de Poisson inflado con ceros ZIP (Zero-Inflated Poisson), dada por Lambert en [12] se usa para datos de recuento que muestran un exceso de ceros. La distribución ZIP combina la distribución de Poisson y la distribución logit. Como ya se había comentado anteriormente, los valores posibles de Y son enteros no negativos: 0, 1, 2, 3 y así sucesivamente. Supóngase que para cada observación hay dos casos posibles, caso I y caso II. Si ocurre el caso I (probabilidad de éxito), el recuento es cero y su probabilidad es π. Sin embargo, si ocurre el caso II, los recuentos (incluidos los ceros) se generan de acuerdo con el modelo de Poisson estándar. Luego, las ocurrencias de la variable aleatoria Y sigue las siguientes distribuciones:

$$Y = \begin{cases} 0 & \text{con probabilidad} \quad \pi + (1-\pi)\exp(-\lambda) \\ \\ k \in \mathbb{Z}^+ & \text{con probabilidad} \quad (1-\pi)\dfrac{\lambda^y \exp(-\lambda)}{y!} \end{cases}$$

Por lo tanto, la distribución de probabilidad ZIP de la variable aleatoria Y se escribe:

$$f(y;\lambda) = P(Y=y) = \begin{cases} \pi + (1-\pi)\exp(-\lambda) & \text{si} \quad y = 0 \\ \\ (1-\pi)\dfrac{\lambda^y \exp(-\lambda)}{y!} & \text{si} \quad y > 0 \end{cases}$$

donde π es el enlace de la función logit tal que $0 < \pi < 1$ y dada por

$$\pi = \frac{\varphi}{1+\varphi} = \frac{\exp(Z\gamma)}{1+\exp(Z\gamma)} \tag{2.15}$$

El logit de un número π entre 0 y 1 se escribe como:

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

La media y la varianza de la distribución ZIP son, respectivamente,

$$E(Y) = (1-\pi)\lambda \qquad \text{y} \qquad Var(Y) = \lambda(1-\pi)(1+\lambda\pi).$$

Note que si $\pi \longrightarrow 0$ la distribución ZIP se acerca a la distribución de Poisson.

Incluyendo la variable de exposición, para un conjunto de p variables explicativas, la componente de Poisson (observación i) es:

$$\lambda_i = N_i \exp\left(\beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}\right) \tag{2.16}$$

De igual forma, la componente logit para un conjunto de n variables explicativas donde se incluya la variable de exposición es:

$$\varphi_i = N_i \exp\left(\gamma_1 + \sum_{j=2}^{n} \gamma_j z_{ij}\right) \tag{2.17}$$

El vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^t$ son los parámetros desconocidos del modelo que se van a estimar. Por otro lado, es importante notar que los modelos logísticos no calculan el valor de Y, sino que modelan la probabilidad de que Y asuma una categoría. En el caso de dos categorías por ejemplo, la variable respuesta es la probabilidad de que Y asuma la categoría de éxito. Un cambio en la variable X_j genera un cambio en la probabilidad.

Las variables explicativas de las componentes (2.16) y (2.17) corresponden a las matrices X y Z, estas son las variables explicativas del modelo Poisson y logit respectivamente. En algunas ocasiones, de acuerdo a las necesidades del investigador y la estructura de los datos, estas matrices pueden ser idénticas o tener algunas variables explicativas comunes. Sin embargo, puede suceder que el modelo logit no necesite de todas las variables que se incluyeron en el modelo de Poisson y por ende la matriz Z tenga un número menor de variables explicativas. Es decir, la parte con inflado de ceros puede incluir o no el total de variables explicativas del modelo. Además, se debe tener en cuenta que $X_i = (1, x_{i2}, \ldots, x_{ip})$ y $Z_i = (1, z_{i2}, \ldots, z_{in})$ pueden o no incluir términos en común. Estos vectores son las i-ésimas filas de la matriz de covariables X y Z respectivamente.

La función de verosimilitud para la distribución ZIP es:

$$L(\beta, \gamma) = \prod \left[\left(\frac{\exp(Z_i\gamma)}{1 + \exp(Z_i\gamma)} + \frac{\exp(-\exp(X_i\beta))}{1 + \exp(Z_i\gamma)} \right) \left(\frac{\exp[-\exp(X_i\beta)][\exp(X_i\beta)]^y}{[1 + \exp(Z_i\gamma)]y!} \right) \right]$$

Solución por estimación de máxima verosimilitud

La estimación de los parámetros se puede llevar a cabo empleando el método de máxima verosimilitud o el algoritmo de Newton-Raphson. La función log-verosimilitud $\ln(L)$ que se usa para estimar los vectores de parámetros β y γ es:

$$\ell = \ln[L(\beta, \gamma)] = \ell_1 + \ell_2 - \ell_3 \,,$$

donde

$$\begin{aligned} \ell_1 &= \sum_{y_i=0} \ln[\exp(Z_i\gamma) + \exp(-X_i\beta)] \\ \ell_2 &= \sum_{y_i>0} [y_i X_i\beta - \exp(X_i\beta)] - \sum_{y_i>0} \ln(y_i!) \\ \ell_3 &= \sum_{i=1}^{m} \ln[1 + \exp(Z_i\gamma)] \end{aligned} \tag{2.18}$$

2.3.2. Regresión Binomial Negativa Inflado con Ceros

La Regresión Binomial Negativa Inflado con Cero se emplea para el modelado de las variables de conteo con ceros excesivos y, por lo general, para las variables de recuento

sobredispersas. Además, la teoría sugiere que el exceso de ceros se genera mediante un proceso separado de los valores de conteo. La distribución de datos combina la distribución binomial negativa y la distribución logit. Los valores posibles de Y son los enteros no negativos: 0, 1, 2, 3,...

Nuevamente, supóngase que para cada observación existen dos casos posibles. Consideremos que ocurre el caso I, luego el recuento es cero y su probabilidad es π. Si acontece el caso II, los recuentos (incluidos los ceros) se generan de acuerdo con el modelo binomial negativo. Asumiendo que ocurre el caso I con probabilidad π y el caso II con probabilidad $1-\pi$, la distribución de probabilidad ZINB de la variable aleatoria Y se escribe como:

$$f(y;\lambda,\alpha) = P(Y=y) = \begin{cases} \pi + (1-\pi)f(y;\alpha,\lambda) & \text{si} \quad y=0 \\ \\ (1-\pi)f(y;\alpha,\lambda) & \text{si} \quad y>0 \end{cases}$$

donde π es la función logit definida en (2.15) y $f(y;\alpha,\lambda)$ es la distribución binomial negativa definida en la ecuación (2.14), sección 2.2.2. Por otro lado, la media y la varianza de la distribución ZINB, son respectivamente:

$$E(Y) = (1-\pi)\lambda \qquad \text{y} \qquad V(Y) = (1-\pi)\lambda(1+\pi\lambda+\alpha\lambda)\,.$$

Observe que la distribución ZINB se acerca a la distribución ZIP y a la distribución binomial negativa cuando $\alpha \longrightarrow 0$ y $\pi \longrightarrow 0$, respectivamente. Si α y π se aproximan a cero, entonces la distribución ZINB se reduce a la distribución de Poisson.

Si la componente binomial negativa incluye la variable de exposición, entonces para un conjunto de p variables explicativas tal que $i \in \{1,2,\ldots,m\}$ se tiene que:

$$\lambda_i = N_i \exp\left(\beta_1 + \sum_{j=2}^{p} \beta_j x_{ij}\right) \tag{2.19}$$

Así mismo, la componente logit que incluye la variable de exposición se escribe como:

$$\varphi_i = N_i \exp\left(\gamma_1 + \sum_{j=2}^{n} \gamma_j z_{ij}\right) \tag{2.20}$$

El vector columna $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)^t$ corresponde a los parámetros desconocidos del modelo logit.

La función de verosimilitud de la distribución ZINB es la siguiente:

$$L(\alpha,\beta,\gamma) = \prod f(\alpha,\beta,\gamma)\,.$$

Solución por estimación de máxima verosimilitud

Los coeficientes de regresión se estiman utilizando el método de máxima verosimilitud (ver [11, 14, 19]). La log-verosimilitud del modelo ZINB dados los datos observados es:

$$\begin{aligned}\ell(\alpha,\beta,\gamma) = &\sum_{i=1}^{m}\ln[1+\exp(Z_i\gamma)] - \sum_{i=1:y_i=0}^{m}\ln\left[\exp(Z_i\gamma)+\left(\frac{\exp(X_i\beta)+\alpha}{\alpha}\right)^{-\alpha}\right] \\ &+ \sum_{i=1:y_i>0}^{m}\left[\alpha\ln\left(\frac{\exp(X_i\beta)+\alpha}{\alpha}\right)+y_i\ln[1+\exp(-X_i\beta)\alpha]\right] \\ &+ \sum_{i=1:y_i>0}^{m}[\ln\Gamma(\alpha)+\ln\Gamma(1+y_i)-\ln\Gamma(\alpha+y_i)]\end{aligned} \tag{2.21}$$

Por último, se describe el estadístico de Wald utilizado en las pruebas de hipótesis de los betas y el Residual de Pearson (ver capítulo 3).

Definición 2.7 (Estadístico de Wald - 1943)**.**
Se presenta el estadístico para una prueba de significacia de una hipótesis nula $H_0 : \beta = \beta_o$, tal que con error estándar (SE) de $\hat{\beta}$, el estadístico de prueba

$$Z = \frac{\hat{\beta}-\beta_o}{SE}$$

tiene una distribución normal estándar aproximada cuando $\beta = \beta_o$.

La extensión multivariada para la prueba de Wald $H_0 : \beta = \beta_o$ tiene como estadístico

$$W = (\hat{\beta}-\beta_o)^t[cov(\hat{\beta})]^{-1}(\hat{\beta}-\beta_o)\,.$$

La distribución normal multivariada asintótica para $\hat{\beta}$ implica una distribución asintótica chi-cuadrado para W. Los grados de libertad son igual al rango de $cov(\hat{\beta})$, que es el número de parámetros no redundantes en β. Otros métodos de uso general se pueden ver en [2].

Definición 2.8 (Residual de Pearson)**.**
El residual de Pearson, definido por

$$r_p = \frac{y-\hat{\lambda}}{\sqrt{\widehat{Var}(y)}}\,,$$

es justo el residuo escalado por la estimación de la desviación estándar de Y. El nombre proviene del hecho de que para la distribución de Poisson, el residual de Pearson es solo la raíz cuadrada del estadístico de bondad de ajuste de Pearson χ^2_P, así que

$$\sum r_P^2 = \chi_P^2\,.$$

Sin embargo, el estadístico de Pearson en este trabajo se utiliza no tanto como medida de bondad de ajuste, sino mejor como una medida de la variación residual (ver capítulo 3).

CAPÍTULO 3

Modelación Número de Eventos Vía Regresión

En este capítulo se presenta un análisis de modelado vía regresión para datos de recuento, en el número de denuncias sobre delitos sexuales en Nariño, en las temporadas *2015* y *2016* segregadas por semestre. Los modelos que se evalúan son los que se presentaron en el capítulo 2. Inicialmente se muestran las variables objeto de estudio, posteriormente las pruebas de sobredispersión y bondad de ajuste, mas adelante la evaluación de los modelos y finalmente los criterios de selección de los mismos. Algunas tablas exportadas de STATA con información que no se incluyen en este capítulo se encuentran en el apéndice A.

3.1. Variables del modelo y población a riesgo

Para modelar un fenómeno como el de denuncias sobre delitos sexuales, aplicando un modelo de regresión para datos de recuento, se debe considerar la población expuesta N formada en conjuntos disjuntos, los cuales están bien definidos para cada variable explicativa elegida. Esto implica que al interior de estos conjuntos se conozca con anticipación la población a riesgo, es decir, el número de personas en cada conjunto que pueden ser potenciales víctimas de delitos sexuales. Algunas variables explicativas que podrían resultar interesantes, tales como: la orientación sexual de la víctima, su actividad u ocupación, su etnia, factor de vulnerabilidad, entre otras, no se incluyen como variables predictoras porque no contar con una base de datos que pueda proporcionar la población de control en estos grupos.

Por otra parte, las bases de datos suministradas poseían grandes diferencias en cuanto a la denominación de las variables y la respuesta de cada una de ellas, por cuanto fue necesario realizar una armonización sobre las mismas; en este proceso se pierde de cierta forma algunos pequeños detalles de información pero se conserva la generalidad.

Las variables con las cuales se evaluaron los modelos aparecen en la tabla 3.1. Las variables explicativas son: *Sexo*, *Periodo de tiempo*, *Región* y *Rango de edad*, debajo de cada una de éstas aparecen sus categorías y al frente la etiqueta asignada. La variable *Recuento* corresponde al valor observado de la variable dependiente Y. La *Población* es la variable de exposición N (población a riesgo) que se obtuvo a partir de la página oficial del DANE `http://www.dane.gov.co`, este vector columna se compone de valores positivos.

En este caso cada componente del vector N es una unidad de observación. Si $N \equiv 1$ entonces la unidad de observación será un solo individuo.

TABLA 3.1. Variables explicativas, de exposición y de respuesta.

Variables en el modelo	**Código en el análisis**	**Categóricas**		**Numérica**
		Nominal	**Ordinal**	**y discreta**
Sexo	Sexo_vic			
Hombre		1		
Mujer		2		
Semestre	Semestre			
I semestre 2015			1	
II semestre 2015			2	
I semestre 2016			3	
II semestre 2016			4	
Región	region_hc			
Región Central		1		
Región Costa Pacífica		2		
Región Norte		3		
Región Sur		4		
Región Suroccidente		5		
Rango de edad	rango_edad			
0-9 años			1	
10-19 años			2	
20-29 años			3	
30-39 años			4	
40-49 años			5	
50-59 años			6	
60-69 años			7	
70 años y más			8	
Recuento	_Conteo			Natural
Población (Exposición N)	Poblacion			Entero positivo

Observe que para construir la tabla de frecuencias se combinan las categorías de las variables explicativas y se generan $2 \times 4 \times 5 \times 8 = 320$ observaciones o subcategorías. La tabla de frecuencias se muestra en el apéndice A, tabla A.2.

Las proyecciones de la población aparecen en la página del DANE desde el año *2005* hasta *2020*, por área, sexo y grupos quinquenales, estimados a cada 30 de junio. En estos años la curva de la población tiene un comportamiento decreciente más o menos lineal.

Luego, como se desea conocer la población para cada semestre, se realiza lo siguiente:

- Se transforma los datos suministrados por el DANE de modo que la población quede clasificada según los grupos de edad (los establecidos en la tabla 3.1) y región geográfica a la que pertenecen.
- Sobre la base de datos creada se ejecuta una regresión de orden dos, esto se consigue por medio de un `script` creado en STATA, el cual facilita obtener los coeficientes de la misma, con estos coeficientes se crea una estimación de la población a 30 de marzo para el primer semestre de cada año y a 30 de septiembre para el segundo

semestre de cada año, en la mayoría de las regresiones efectuadas el coeficiente de determinación R^2 fue superior a 95.9 %.

Por otro lado, en la figura 3.1 se observa la frecuencia de cada valor posible que toma la variable observada. A simple vista se muestra una gran cantidad de ceros que toma la variable de recuento, el $67,81\,\%$ corresponde a conteos nulos. Sin embargo, por tratarse de eventos raros, generalmente en los conteos se espera que aparezca una gran cantidad de ceros como los que se observan en la gráfica.

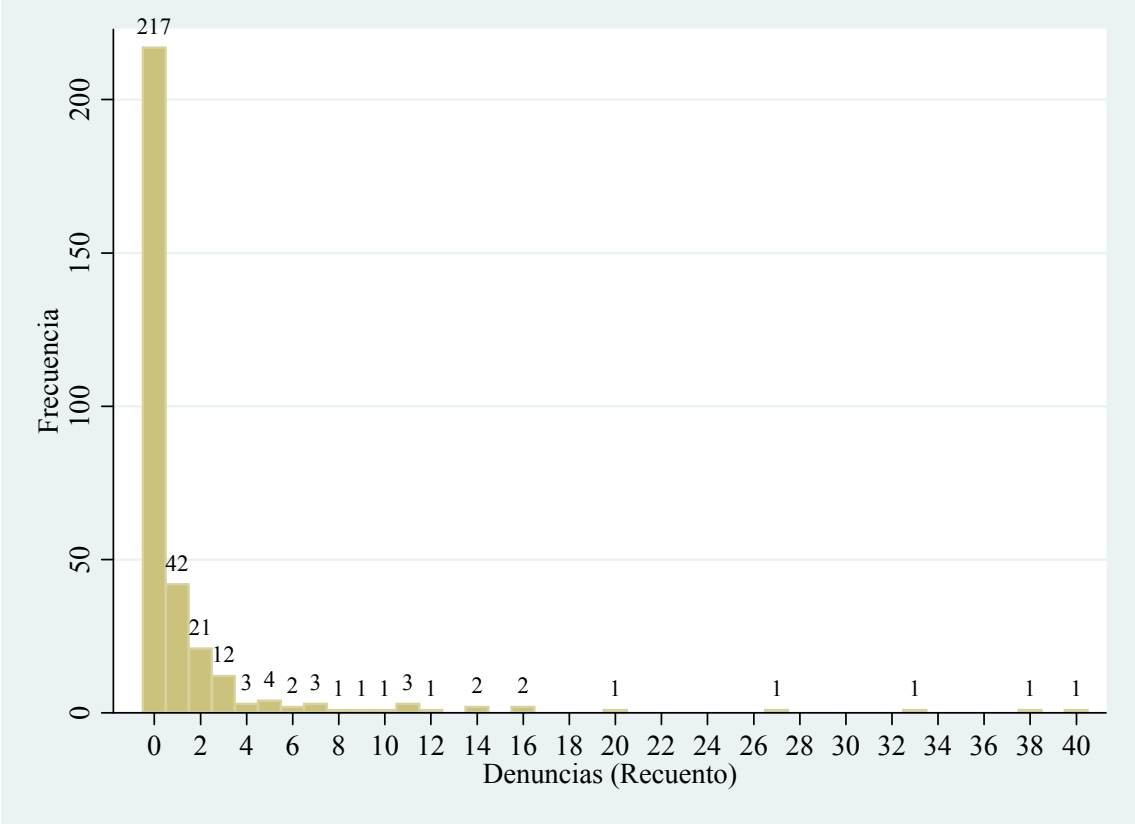

FIGURA 3.1. Número de denuncias.

3.2. Sobredispersión y bondad de ajuste

3.2.1. Sobredispersión

Un primer aspecto a tener en cuenta en este tipo de datos es la sobredispersión, es decir, cuando se cumple la desigualdad $Var(Y) > E(Y)$. Este fenómeno no lo corrige el MRP, luego utilizar un modelo no adecuado puede llevar a interpretaciones falsas o ajustes muy pobres. Generalmente se determinan dos fuentes de sobredispersión: la heterogeneidad de la población y el exceso de ceros. La heterogeneidad se observa cuando la población se puede dividir en muchas subpoblaciones homogéneas. El exceso de ceros se detecta cuando el número de ceros observados excede en gran medida el número de ceros reproducidos por la distribución Poisson (ver [11]).

Se mencionan algunas relaciones que permiten evaluar sobredispersión de los datos en los MLG (ver [2, 9, 16]). Estos estadísticos se obtienen tanto en STATA como en *R*.

I. Razón entre el estadístico chi-cuadrado de Pearson y los grados de libertad χ^2_P/gl. Si este valor es mayor que 1, indica sobredispersión. El estadístico χ^2_P se distribuye aproximadamente en chi-cuadrado con $m-p$ grados de libertad (m número de observaciones y p número de parámetros estimados incluyendo el intercepto).

$$\chi^2_P = \sum_{i=1}^{m} \frac{(y_i - \hat{\lambda}_i)^2}{\widehat{Var}(y_i)}\,,$$

donde y_i es la cantidad de observaciones en cada subcategoría $i \in \{1, 2, \ldots, m\}$. El estadístico χ^2_P también sirve para probar bondad de ajuste cuando la variable que se analiza es de frecuencia, es decir, cuando los datos son agrupados. Si los datos no están agrupados entonces χ^2_P no sigue una distribución chi-cuadrado.

II. Razón entre la función desvianza y los grados de libertad D/gl. Un valor de este cociente mayor que 1 indica sobredispersión. De acuerdo a McCullagh y Nelder en [16], el estadístico D así construido tiene distribución asintótica chi-cuadrado con $m-p$ grados de libertad, es decir, si el modelo es correcto entonces $D(y;\hat{\lambda}) \sim \chi^2_{m-p}$.

$$D(y;\hat{\lambda}) = -2[\ln[L(\hat{\lambda};y)] - \ln[L(y;y)]]\,.$$

La desvianza es el doble de la diferencia entre la máxima log-verosimilitud alcanzable del modelo saturado (m parámetros, uno por cada observación) y la log-verosimilitud del modelo ajustado. Este estadístico se utiliza para modelos Poisson y Binomial, generalizando los métodos de análisis de varianza para modelos lineales normales.

III. Otro diagnóstico se basa en una prueba de Razón de Verosimilitud (Likelihood-ratio test) apoyada en las distribuciones Poisson y Binomial Negativa tradicional.

- ✠ Para la distribución de Poisson $Var(Y) = \lambda$.
- ✠ Para la distribución binomial negativa $Var(Y) = \lambda + \alpha\lambda^2$.

Si $\alpha = 0$, la distribución Binomial Negativa se reducirá a una Poisson. Luego las hipótesis que se plantean son las siguientes:

$$H_0 :\ \alpha = 0 \qquad\qquad H_a :\ \alpha > 0$$

Ahora, para llevar a cabo esta prueba se deben ajustar los dos modelos: Poisson y Binomial Negativo. Con cada modelo se obtiene su respectiva función de log-verosimilitud $\ell = \ln(L)$. El estadístico que se propone es:

$$RV = -2\ln\left(\frac{L(Poisson)}{L(BN)}\right) = 2\ell_{BN} - 2\ell_{Poisson} \tag{3.1}$$

De acuerdo a Cameron y Trivedi 1998, este estadístico tine una distribución asintótica $RV \sim \chi^2_{(k)}$, con k grados de libertar ($k = 1$ en el caso de una prueba de MRBN contra MRP). Por tanto, se rechaza H_0 si el estadístico de prueba RV es mayor que $\chi^2_{(1)}$, al nivel de significancia escogido. En tal caso, será más conveniente modelar el número de ocurrencias a través de una regresión binomial negativa. La interpretación de los resultados es la misma que en el caso de la regresión Poisson.

3.2.2. Bondad de ajuste

Se presentan a continuación los estadísticos empleados para probar bondad de ajuste en los modelos. Los estadísticos LR y de Vuong se utilizan como prueba de bondad de ajuste de una hipótesis. Los criterios AIC y BIC se emplean como estadísticos para la selección del mejor modelo que representa los datos.

- El estadístico LR es la prueba chi-cuadrado de razón de verosimilitud de que al menos uno de los coeficientes de regresión de los predictores no es igual a cero. Bajo la hipótesis nula H_0, LR se distribuye $LR \sim \chi^2_{p-1}$, donde p es el número de parámetros estimados en el modelo incluyendo el intercepto.

$$H_0: \ \beta_j = 0 \ \text{ para todo } j \qquad H_a: \ \beta_j \neq j \ \text{ para algún } j$$

$$LR = -2[\ln(L_0) - \ln(L_1)]\,,$$

  aquí $\ln(L_0)$ es la función log-verosimilitud del modelo nulo (modelo sin variables explicativas) y $\ln(L_1)$ es el valor de log-verosimilitud del modelo ajustado completo.

- La prueba de Vuong 1989 (ver [25]), es una extensión de la prueba de razón de verosimilitud para evaluar modelos no anidados. Por ejemplo, el MRP y el modelo de regresión ZIP son modelos no anidados, ocurre lo mismo con el MRBN y el modelo de regresión ZINB. Los modelos anteriores se denominan no anidados debido a que los parámetros del modelo inflado con ceros y los parámetros del modelo estándar no se pueden comparar por utilizar distintas funciones de enlace. Por tanto, la prueba de Vuong es una prueba del modelo inflado con ceros versus el modelo estándar. En STATA una prueba `z` significativa indica que se prefiere el modelo con exceso de ceros, en vez del modelo tradicional. La expresión analítica de este estadístico es:

$$V = \frac{\sqrt{m}\left[\frac{1}{m}\sum_{i=1}^{m} n_i\right]}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(n_i - \bar{n})^2}}\,,$$

  siendo $n_i = \ln[P_1(Y = y_i)/P_2(Y = y_i)]$, $P_1(Y = y_i)$ y $P_2(Y = y_1)$ las funciones de probabilidad para el modelo inflado de ceros y el tradicional, respectivamente, y $\bar{n}$ la media de n_i, $i \in \{1, 2, \ldots, m\}$. Bajo la hipótesis nula, V es un estadístico bidireccional con una distribución asintótica normal tipificada. Cuando $|V| < 1.96$ el contraste no permite decantarse por uno u otro modelo; en caso contrario, si $V > 1.96$, es una evidencia a favor del modelo inflado de ceros, mientras que un valor tal que $V < -1.96$ favorece el modelo tradicional (ver [18]).

- El Criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto de datos. Como tal, el AIC proporciona un medio para la selección del modelo. AIC no proporciona una prueba de un modelo en el sentido de probar una hipótesis nula , es decir AIC no puede decir nada acerca de la calidad del modelo en un sentido absoluto. Si todos los modelos candidatos encajan mal, AIC no dará ningún aviso de ello. Su formulación matemática corresponde a:

$$AIC = 2[p - \ln(L)]\,,$$

donde p es el número de parámetros en el modelo estadístico, y L es el máximo valor de la función verosimilitud para el modelo estimado. Dado un conjunto de modelos candidatos para los datos, el mejor modelo es el que tiene el valor mínimo en el AIC.

- Criterio de Información Bayesiano (BIC) como otra estadística de ajuste común. Aunque este estadístico tiene tres formulaciones, solo se considerará la que se muestra a continuación.

$$BIC = -2\ln(L) + p\ln(m)\,,$$

donde p es el número de parámetros libres a ser estimados, m es el tamaño de la muestra y L es el máximo valor de la función de verosimilitud. El BIC está estrechamente relacionado con el AIC y al igual que éste resuelve el problema de selección de modelos mediante la introducción de un término de penalización para el número de parámetros en el modelo, el término de penalización es mayor en el BIC que el en AIC. Se selecciona aquel modelo con menor valor BIC.

3.3. Ajuste y selección del modelo

En esta sección se presenta la evaluación de los modelos. A cada modelo se le aplican pruebas de sobredispersión y de bondad de ajuste para verificar si el modelo representa o no la información y de que forma lo hace. Luego, cada modelo se compara con los demás con el propósito de tratar de establecer cual se ajusta mejor a los datos. Los criterios empleados para inferir sobre el mejor modelo son el AIC y BIC.

La siguiente información exportada del sotfware STATA muestra que *(1/df) Deviance* y *(1/df) Pearson* son valores mayores que 1, esto significa que los datos presentan sobredispersión. De igual manera, observe que los estadísticos $D(y;\hat{\lambda}) = 584.1649355$ y $\chi^2_P = 829.0075744$ no siguen una distribución chi-cuadrado con $m - p = 315$ grados de libertad. Esto indica que posiblemente el MRP no es mejor representante de los datos.

```
Generalized linear models                         No. of obs      =        320
Optimization     : ML                             Residual df     =        315
                                                  Scale parameter =          1
Deviance         =  584.1649355                   (1/df) Deviance =   1.854492
Pearson          =  829.0075744                   (1/df) Pearson  =    2.63177

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   2.766256
Log likelihood   = -437.6009346                   BIC             =  -1232.856
```

_Conteo	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo_vic	2.389164	.163389	14.62	0.000	2.068927	2.7094
semestre	-.2272188	.0419195	-5.42	0.000	-.3093796	-.145058
region hc	-.2501383	.0371691	-6.73	0.000	-.3229884	-.1772881
rango_edad	-.4274009	.0315862	-13.53	0.000	-.4893088	-.365493
_cons	-11.44508	.3462574	-33.05	0.000	-12.12373	-10.76643
ln(Poblac~n)	1	(exposure)				

Por otro lado, siguiendo a Cameron y Trivedi en [7], quienes recomiendan errores robustos para modelos Poisson, se mostrará que los errores estándar de los coeficientes individuales aumentan al utilizar este tipo de errores. Es decir, los errores estándar pueden presentar estimaciones sesgadas de los coeficientes de regresión en estos modelos, por lo que es importante entonces usar errores robustos.

Los errores robustos se basan en la función `Log pseudolikelihood` y en el estadístico `Wald chi-cuadrado`, en cambio los errores estándar utilizan la función `Log likelihood` y el estadístico `LR chi-cuadrado`.

El valor `P>|Z|` es la probabilidad de que el estadístico de prueba `Z` se observe bajo la hipótesis nula de que el coeficiente de regresión de un predictor particular es cero, dado que el resto de los predictores están en el modelo. Para un nivel crítico dado, `P>|Z|` determina si la hipótesis nula puede ser rechazada o no. Si `P>|Z|` es menor que 0.05, entonces la hipótesis nula puede rechazarse y la estimación del parámetro se considera significativa en ese nivel crítico. Para el ejemplo anterior, a pesar de que el modelo probablemente no sea el mejor representante de los datos, dado que éstos presentan sobredispersión, el valor `P>|Z|` $(p-value)$ muestra que a un nivel crítico del $5\,\% = 0.05$, el estadístico de prueba `Z` cae fuera de la región de aceptación porque todos los p-valor son menores que 0.05. Luego se rechaza la hipótesis nula de que los coeficientes de regresión betas son iguales a cero.

La siguiente información obtenida de STATA muestra el modelo anterior pero con errores robustos de los coeficientes, observe que toda la información que se muestra es la misma que la anterior, excepto, el tamaño de los errores de los coeficientes y la variable **semestre** que ya no es significativa. Además, note que como se mencionó en el capítulo 2, el coeficiente de la variable de exposición se puede considerar igual a 1.

En todos los modelos, los intervalos de confianza se determinan con $\hat{\beta}_j \pm (1.96) SE_j$.

```
Generalized linear models                         No. of obs      =        320
Optimization     : ML                             Residual df     =        315
                                                  Scale parameter =          1
Deviance         =  584.1649355                   (1/df) Deviance =   1.854492
Pearson          =  829.0075744                   (1/df) Pearson  =    2.63177

Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   2.766256
Log pseudolikelihood = -437.6009346               BIC             =  -1232.856
```

_Conteo	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo_vic	2.389164	.2599432	9.19	0.000	1.879685	2.898643
semestre	-.2272188	.1181292	-1.92	0.054	-.4587477	.0043101
region_hc	-.2501383	.1004728	-2.49	0.013	-.4470614	-.0532151
rango_edad	-.4274009	.050577	-8.45	0.000	-.5265299	-.3282719
_cons	-11.44508	.5730018	-19.97	0.000	-12.56814	-10.32202
ln(Poblac~n)	1	(exposure)				

Ahora, se mostrará el ajuste del MRBN con errores estándar. Para este modelo los errores robustos y estándar muestran resultados muy similares, además como se dijo an-

teriormente, Cameron y Trivedi en [7] sugieren errores robustos para modelos Poisson, ya sea para el MRP o el modelo de regresión ZIP.

Observe que el coeficiente $\alpha = 0,8709182$ es diferente de cero, lo que indica que existe sobresipersión en los datos, algo que ya se había evidenciado en el ajuste del MRP.

La prueba `Likelihood-ratio test of alpha=0` es una prueba de la BN contra la distribución de Poisson. El valor del estadístico `chibar2(01) = 246.54` se obtiene como en la ecuación (3.1), es decir, $RV = 246.54$. Luego como el p-valor para esta prueba es `0.000`, entonces se rechaza la hipótesis nula de que $\alpha = 0$ y por lo tanto el MRBN presenta un mejor ajuste a los datos que el MRP. Observe también que en el MRP, tanto para errores robustos como para errores estándar, la función log-verosimilitud es `-437.6009346`, valor más pequeño que en el MRBN, donde su valor es `-314.3312`, es decir, el valor de la función log-verosimilitud es mayor en el MRBN que en el MRP. De igual forma, el valor del estadístico desvianza es `D=201.4302456`, valor que sigue una distribución $\chi^2_{(315)}$.

Finalmente, el estadístico $LR = -2[\ln(L_0) - \ln(L_1)] = 170.20$ es una prueba del modelo nulo (sin predictores) versus el modelo completo (con variables explicativas). Una prueba significativa indica que los factores independientes inciden en la variable respuesta, y por tanto se rechaza la hipótesis nula que asume que las variables independientes no explican la variable dependiente. En este caso, dado que el p-valor es `Prob>chi2=0.0000`, entonces el modelo es estadísticamente significativo e indica que por lo menos un factor independiente explica la ocurrencia de delitos sexuales con un 95 % de confiabilidad. Por todo lo descrito anteriormente, el MRBN parece ser un buen representante de los datos.

```
Negative binomial regression                    Number of obs   =        320
                                                LR chi2(4)      =     170.20
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -314.3312                      Pseudo R2       =     0.2131
```

_Conteo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo_vic	2.474253	.2357272	10.50	0.000	2.012236	2.936269
semestre	-.2457265	.0860038	-2.86	0.004	-.4142908	-.0771623
region_hc	-.1631433	.0632675	-2.58	0.010	-.2871453	-.0391413
rango_edad	-.5292587	.0614906	-8.61	0.000	-.6497781	-.4087393
_cons	-11.51644	.4785456	-24.07	0.000	-12.45437	-10.57851
ln(Poblac~n)	1	(exposure)				
/lnalpha	-.1382073	.1943456			-.5191177	.2427031
alpha	.8709182	.1692591			.5950454	1.27469

```
Likelihood-ratio test of alpha=0:  chibar2(01) =  246.54 Prob>=chibar2 = 0.000
```

Como se mencionó en el capítulo 2, el MRBN corrige los dos fenómenos presentados con mayor frecuencia en datos tipo recuento, sobredispersión y exceso de ceros. Sin embargo, cuando el exceso de ceros supera el 80 % de los datos y la muestra es grande, se conjetura que en algunos casos los modelos inflados con ceros ZINB y ZIP subsanan mejor este fenómeno, aunque en este contexto la idea de muestra grande no esta bien definida.

Mostraremos ahora el ajuste para el modelo de regresión ZINB. En este caso se debe ingresar variables explicativas para el MRBN y el modelo inflado. Las variables que se ingresan a los modelos *log* y *logit* no necesariamente deben de ser las mismas. Realizando todas las posibles combinaciones de variables que pueden entrar al modelo inflado (ver apéndice A, tabla A.3), se selecciono aquel modelo con menor AIC. Los resultados exportados de STATA se muestran a continuación. Observe que se rechaza la hipótesis nula $H_o : \alpha = 0$, al 95 % de confianza, y se acepta la hipótesis alternativa $H_a : \alpha > 0$, puesto que la probabilidad de obtener un valor mayor que $RV = 221.83$ es 0.0000. Esto constata que el modelo ZINB es preferible al modelo ZIP (razón por la cual no se coloca información sobre el modelo ZIP). Por otro lado, el estadístico de Vuong muestra que no existe una mejora significativa para determinar que el modelo ZINB representa mejor los datos que el MRBN. Además, se puede apreciar que la variable *sexo* que ingresa al modelo inflado, de acuerdo al p-valor, no es estadísticamente significativa, esto sugiere que se podría extraer del modelo sin peder información importante.

Algunas combinaciones de variables ingresadas al modelo inflado de ZINB no permiten realizar inferencias por cuando el modelo completo no converge. Además, las funciones de verosimilitud de los modelos inflado y binomial negativo no comparten la misma función de enlace, por lo que los parámetros se estiman de forma independiente.

```
Zero-inflated negative binomial regression        Number of obs   =        320
                                                   Nonzero obs     =        103
                                                   Zero obs        =        217

Inflation model = logit                           LR chi2(4)      =     118.95
Log likelihood  = -314.1203                       Prob > chi2     =     0.0000
```

_Conteo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Conteo						
sexo_vic	**2.31284**	**.3394813**	**6.81**	**0.000**	**1.647469**	**2.978211**
semestre	**-.2417251**	**.0862774**	**-2.80**	**0.005**	**-.4108258**	**-.0726244**
region_hc	**-.1605072**	**.0630341**	**-2.55**	**0.011**	**-.2840518**	**-.0369626**
rango_edad	**-.527187**	**.061337**	**-8.59**	**0.000**	**-.6474054**	**-.4069686**
_cons	**-11.21637**	**.6581887**	**-17.04**	**0.000**	**-12.5064**	**-9.926346**
ln(Poblac~n)	**1**	(exposure)				
inflate						
sexo_vic	**-21.49036**	**23956.15**	**-0.00**	**0.999**	**-46974.67**	**46931.69**
_cons	**19.77312**	**23956.15**	**0.00**	**0.999**	**-46933.41**	**46972.96**
/lnalpha	**-.1774743**	**.2024527**	**-0.88**	**0.381**	**-.5742743**	**.2193258**
alpha	**.8373826**	**.1695304**			**.5631134**	**1.245237**

```
Likelihood-ratio test of alpha=0: chibar2(01) =   221.83 Pr>=chibar2 =  0.0000
Vuong test of zinb vs. standard negative binomial: z =      0.37  Pr>z = 0.3570
```

Por otro lado, la tabla 3.2 muestra las estadísticas de los datos observados, mediante la opción recuento y los valores esperados por parte de los cuatro modelos evaluados. Note que los modelos que mejor predicen son MRBN y ZINB. Esto da a entender que estos modelos corrigen mejor el fenómeno de sobredispersión. Sin embargo, es importante evidenciar que los valores máximos de los datos esperados no alcanzan el máximo valor observado, esto se debe a que la varianza de los datos observados supera la varianza esperada.

TABLA 3.2. Estadísticas valores observados y esperados.

Recuento-Modelos	Observaciones	Media	Des. Est.	Mínimo	Máximo
Medidas muestrales	320	1.48438	4.62172	0.00000	40.00000
MRP	320	1.48438	3.18978	0.00291	22.16215
MRBN	320	1.51286	3.38274	0.00188	23.98144
ZIP	320	1.40144	2.87812	0.00349	18.96675
ZINB	320	1.506695	3.35564	0.0019116	23.70935

En la figura 3.2 se muestra el número de eventos predichos por los modelos presentados en la tabla 3.2. Comparando los datos observados de la figura 3.1, se aprecia que la variable explicada sigue las distribuciones de probabilidad reveladas en el capítulo 2. En cuanto a los ceros en los recuentos, el MRBN y el modelo ZINB son los que mejor se ajustan a este fenómeno.

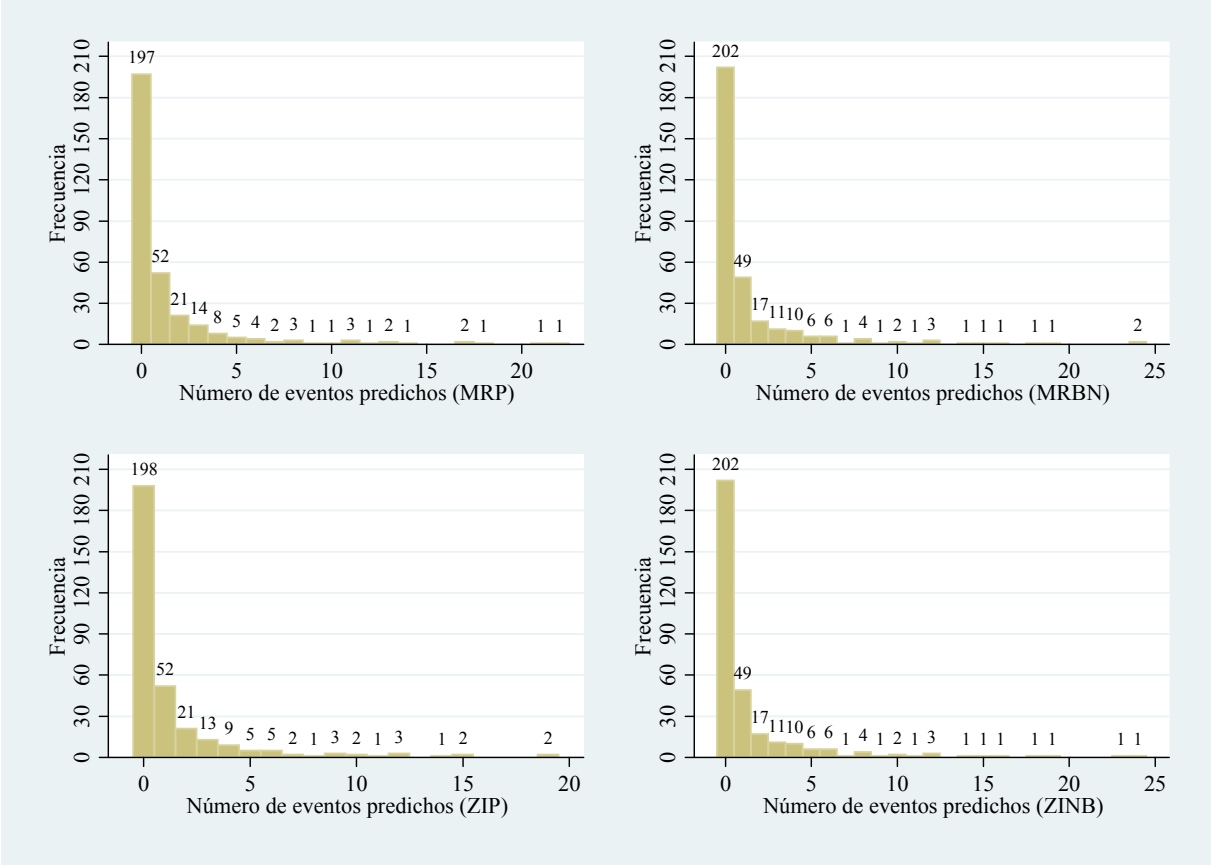

FIGURA 3.2. Recuentos predichos.

A continuación, la tabla 3.3 muestra los AIC y BIC para los cuatro modelos en cuestión. De cada modelo se selecciono aquel con mejor valor AIC y BIC. Los grados de libertad (gl) de cada modelo corresponden a los parámetros estimados en cada uno de éstos. Los valores debajo del modelo nulo y del modelo completo son las funciones de log-verosimilitud de cada uno. El modelo *nulo* hace referencia a un modelo sin variables explicativas, es decir, contiene como único parámetro el valor esperado λ para todas las observaciones, este modelo se utiliza como referencia. Por su parte, el modelo *completo* representa el modelo

con predictores. Se puede mirar que los modelos que contienen variables explicativas son aquellos con mayor valor en la función de log-verosimilitud. Esto significa que las variables independientes dentro de los modelos inciden en la variable respuesta.

Tabla 3.3. Modelos mejor ajustados en base a los criterios AIC y BIC.

Modelos	Obs.	Modelo(nulo)	Modelo(completo)	gl	AIC	BIC
MRP	320	-786.9912	-437.6009	5	885.2019	904.0435
MRBN	320	-399.4337	-314.3312	6	640.6624	663.2723
ZIP	320	-565.973	-425.0354	7	864.0709	890.4491
ZINB	320	-373.5959	-314.1203	8	644.2407	674.3872

Por último, la tabla 3.4 representa los coeficientes de regresión estimados por los modelos y su respectivo error robusto que se comtempló. Los errores robustos, como se había mencionado anteriormente [7], son recomendables para modelos de Poisson ya que permiten corregir estimaciones sesgadas. Sin embargo, para modelos binomiales los resultados no se diferencian y es criterio del investigador utilizar errores que mejor convengan.

También en la tabla 3.4, el error en la estimación del coeficiente es grande en el modelo ZINB de su parte inflada, esto evidencia que la variable, *sexo*, considerada en este caso no explica el exceso de ceros en la variable respuesta. No obstante, en el modelo ZIP, la variable *sexo* incluida en el modelo inflado es significativa.

En el apéndice A se muestra el modelo ZIP, no incluido en este capítulo, donde se observa además que la prueba de Vuong sugiere el modelo ZIP en lugar del MRP.

Por lo tanto, teniendo en cuenta las pruebas de bondad de ajuste realizadas, se tomarán el MRBN y el modelo ZINB como los "mejores modelos"que representa los datos, de acuerdo a los criterios de selección AIC y BIC mostrados en la tabla 3.3. Es importante tener en cuenta que el concepto de *mejor modelo* lo determinan los datos. Para la interpretación de resultados solo se mostrará la información exportada de STATA del MRBN, esto debido a que este modelo tiene un mejor valor de AIC, aproximadamente cuatro puntos por debajo del valor AIC para el modelo ZINB. Además, fue decisión de los autores tomar un modelo que considere un menor número de parámetros a estimar, obteniendo una ganancia en los grados de libertad.

TABLA 3.4. Coeficientes de los modelos asociados con su error.

Variables	(MRP)	(MRBN)		(ZIP)		(ZINB)		
explicativas	_Conteo	_Conteo	$\ln \alpha$	_Conteo	inflado	_Conteo	inflado	$\ln \alpha$
sexo_vic	2.389***	2.474***		1.882***	-1.849***	2.313***	-21.49	
	(0.260)	(0.236)		(0.290)	(0.656)	(0.339)	(23,956)	
semestre	-0.227*	-0.246***		-0.206*		-0.242***		
	(0.118)	(0.0860)		(0.124)		(0.0863)		
region_hc	-0.250**	-0.163***		-0.194**		-0.161**		
	(0.100)	(0.0633)		(0.0963)		(0.0630)		
rango_edad	-0.427***	-0.529***		-0.420***		-0.527***		
	(0.0506)	(0.0615)		(0.0547)		(0.0613)		
Constante	-11.45***	-11.52***	-0.138	-10.54***	1.735	-11.22***	19.77	-0.177
	(0.573)	(0.479)	(0.194)	(0.556)	(1.062)	(0.658)	(23,956)	(0.202)
Observaciones	320	320	320	320	320	320	320	320

Errores estándar robustos entre paréntesis.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

3.4. Interpretación del modelo

En esta sección se analizan los resultados con base en las tasas de incidencia o razones de tasa. Las razones de tasa generalmente en los softwares estadísticos se simbolizan con IRR (Incidence Rate Ratio). En este informe se adoptará la misma simbología.

Considérese un MLG con una variable explicativa x tal que $\lambda(x) = \exp(\beta_1 + \beta_2 x)$ y otro modelo con un incremento de una unidad en la variable independiente, o cambio de categoría si la variable es cualitativa, $\lambda(x+1) = \exp(\beta_1 + \beta_2(x+1))$. Se define la razón de tasa IRR de las medias en $x+1$ y en x así:

$$IRR = \frac{\lambda(x+1)}{\lambda(x)} = \frac{\exp(\beta_1 + \beta_2(x+1))}{\exp(\beta_1 + \beta_2 x)} = \exp(\beta_2)\,.$$

Ahora, para cualquier subíndice $i \in \{1, 2, 3, \ldots, m\}$ se tiene que

$$\begin{aligned} \lambda_i = \exp(X_i\beta_i) &= \exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \\ &= \exp(\beta_1)\exp(\beta_2 x_{i2}) \cdots \exp(\beta_p x_{ip}) \end{aligned} \tag{3.2}$$

donde $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ es la i-ésima fila de la matriz X y p es el número de variables explicativas. De acuerdo con (3.2) se observa que los modelos para datos de recuento suponen efectos multiplicativos, es decir, si la componente explicativa unidimensional x_{ij} aumenta c unidades, la media para la variable del modelo se multiplica por la potencia c-ésima de $\exp(\beta_j)$, supuestas las demás variables independientes constantes. En este trabajo, puesto que las variables predictoras son categóricas, lo que se hace es pasar a la categoría correspondiente. Así, la relación de las medias en $X_i + c$ y en X_i es:

$$\frac{\exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_j(x_{ij} + c) + \ldots + \beta_p x_{ip})}{\exp(\beta_1 + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \ldots + \beta_p x_{ip})} = \exp(c\beta_j) = [\exp(\beta_j)]^c \tag{3.3}$$

La ecuación (3.3) indica que la razón de tasa IRR se puede expresar como el exponencial de los coeficientes de regresión del modelo. Además, cuando hay términos de interacción o transformaciones que involucran la variable explicativa de interés, la relación de las medias es más complicada, pero puede derivarse de manera similar (ver [9]).

Un aspecto a tener en cuenta en el análisis de estos datos, es que se esta trabajando con variables explicativas de tipo categórico. Luego, todas la variables a excepción de *sexo vic*, que solo tiene dos categorías, se deben transformar a variables `Dummy`. Esto significa

Tabla 3.5. Transformación a variables Dummy.

Categorías	$R =$ region hc	Z_1	Z_2	Z_3	Z_4
C_1	Centro	1	0	0	0
C_2	C. Pacífica	0	1	0	0
C_3	Norte	0	0	1	0
C_4	Sur	0	0	0	1
C_5	Suroccidente	0	0	0	0

que cada variable se convierte en variables de dos categorías. Por ejemplo, una variable con k categorías, se reemplaza por $k-1$ variables de dos categorías. Las categorías de

las nuevas variables serán ceros y unos, 1 significa que la variable actúa en la variable respuesta y 0 que la variable no actúa. En el caso de la variable *region hc* que tiene cinco categorías, se transforma en 4 variables dummy Z_i, como se muestra en la tabla 3.5. Esto implica que para cualquier Z_i, $Z_i = 1$ si $R = C_i$ y 0 en cualquier otro caso. Observe que la categoría *Suroccidente* de la variable *region hc* es la que se toma como referencia, motivo por el cual en la tabla 3.5 solo aparecen ceros en la fila C_5.

La categoría que tiene los cuatro ceros se llama categoría de referencia o categoría base (se conoce generalmente como "grupo no expuesto"). En este caso la categoría base es la C_5, sin embargo, se puede tomar cualquier categoría como referencia, STATA por ejemplo toma como base la primera categoría. En el modelo ajustado que se presenta mas adelante, la variable *region hc* toma como categoría de referencia la etiqueta *Centro*.

A continuación se muestra la transformación de las cuatro variables de estudio a variables dummy. En cada variable, la primera categoría es tomada como categoría de referencia. Como se observa, el número de variables tipo dummy que tiene el nuevo modelo es de 15.

Sexo víctima
$M = 1$ si es mujer, 0 en otro caso.

Semestre
$S_2 = 1$ si es semestre dos, 0 en otro caso.
$S_3 = 1$ si es semestre tres, 0 en otro caso.
$S_4 = 1$ si es semestre cuatro, 0 en otro caso.

Región de hecho
$R_2 = 1$ si es región pacífica, 0 en otro caso.
$R_3 = 1$ si es región norte, 0 en otro caso.
$R_4 = 1$ si es región sur, 0 en otro caso.
$R_5 = 1$ si es región suroccidente, 0 en otro caso.

Rango de edad
$E_1 = 1$ si la edad esta entre 10 y 19 años, 0 en otro caso.
$E_2 = 1$ si la edad esta entre 20 y 29 años, 0 en otro caso.
$E_3 = 1$ si la edad esta entre 30 y 39 años, 0 en otro caso.
$E_4 = 1$ si la edad esta entre 40 y 49 años, 0 en otro caso.
$E_5 = 1$ si la edad esta entre 50 y 59 años, 0 en otro caso.
$E_6 = 1$ si la edad esta entre 60 y 69 años, 0 en otro caso.
$E_7 = 1$ si la edad es mayor de 70 años, 0 en otro caso.

De esta forma, el modelo teórico se escribe como:

$$\ln(\lambda/N) = \beta_1 + \beta_2 M + \beta_3 S_2 + \ldots + \beta_6 R_2 + \ldots + \beta_{10} E_1 + \ldots + \beta_{16} E_7 \,.$$

La tabla exportada de STATA muestra el modelo de regresión binomial negativo ajustado a los datos. El modelo se expresa en versión extendida con todas las variables dummy. Note que como se menciono anteriormente, en cada variable original no aparece la categoría de referencia. Lo mismo sucede en el modelo con razones de tasa IRR que se presenta mas adelante, donde la primera categoría es tomada como categoría base.

En la variable *sexo vic* por ejemplo, la categoría no expuesta es la categoría hombre, nivel que se encuentra en primer lugar. Luego, para comparar el riesgo de las mujeres con respecto a los hombres se calcula $IRR = \exp(2.355793) = 10.54649$. Esto significa que las mujeres tienen aproximadamente 10.5 veces mas riesgo que los hombres de padecer un delito sexual. De forma análoga se procede para las restantes variables en el modelo.

```
Generalized linear models                         No. of obs      =        320
Optimization     : ML                             Residual df     =        304
                                                  Scale parameter =          1
Deviance         =   240.101584                   (1/df) Deviance =   .7898078
Pearson          =  608.2467758                   (1/df) Pearson  =   2.000812

Variance function: V(u) = u+(.06)u^2              [Neg. Binomial]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   1.827107
Log likelihood   = -276.3371844                   BIC             =  -1513.468
```

_Conteo	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo vic						
Mujer	2.355793	.1740804	13.53	0.000	2.014601	2.696984
semestre						
2	-.449374	.1533097	-2.93	0.003	-.7498554	-.1488926
3	-.5026923	.1553266	-3.24	0.001	-.8071269	-.1982578
4	-.7816736	.1659289	-4.71	0.000	-1.106888	-.456459
region hc						
C. Pacífica	-1.825244	.1931622	-9.45	0.000	-2.203835	-1.446653
Norte	-1.565826	.2263976	-6.92	0.000	-2.009557	-1.122095
Sur	-.7080458	.1571184	-4.51	0.000	-1.015992	-.4000993
Suroccidente	-.8312191	.1747042	-4.76	0.000	-1.173633	-.4888051
rango_edad						
10-19	.6619506	.1468997	4.51	0.000	.3740326	.9498686
20-29	-.1886402	.1731888	-1.09	0.276	-.528084	.1508037
30-39	-1.296187	.2384616	-5.44	0.000	-1.763564	-.8288111
40-49	-2.028898	.3337631	-6.08	0.000	-2.683062	-1.374735
50-59	-2.774065	.5205825	-5.33	0.000	-3.794388	-1.753742
60-69	-3.053444	.7216077	-4.23	0.000	-4.467769	-1.639119
70 y más	-2.00801	.4693481	-4.28	0.000	-2.927915	-1.088105
_cons	-10.01763	.2120145	-47.25	0.000	-10.43317	-9.602087
ln(Poblacion)	1	(exposure)				

En la siguiente tabla se presenta las razones de tasa del MRBN. Los intervalos de confianza para los IRR se construyen con

$$\exp[\hat{\beta}_j \pm 1.96(SE_j).$$

Observe que las mujeres son 10.54649 veces mas propensas que los hombres de ser víctima de un delito sexual (por cada delito sexual ocurrido en un hombre se presentan 10.54649 casos en mujeres), tendencia que de acuerdo a las denuncias siempre se ha presentado. Con respecto al semestre, note que las denuncias presentadas en el primer semestre del *2015*, están cerca del 40 % por encima de las ocurrencias presentadas en los semestres dos y tres; y 55 % por encima de los casos mostrados en el segundo semestre de *2016*. Esto indica, de acuerdo a las denuncias, que la tendencia en el tiempo es a disminuir el riesgo de ocurrencia de estos sucesos. Todas las categorías de la variable semestre son estadísticamente significativas. Por otro lado, en mención a la región del hecho donde

ocurrió el delito sexual, la región Central (región de referencia) fue donde más sucesos se presentaron, seguido por la región Sur y Suroccidente. En la región Sur se presenta el 49.26 % de las denuncias dadas en la región centro, esto señala que la presencia del factor se asocia a menor ocurrencia del evento. La región Suroccidente representa el 43, 55 % de los casos de la zona Centro. La región norte y la región Costa Pacífica fueron las de menor riesgo, la norte con el 20.89 % con respecto a la región de referencia y la región Pacífica con el 16.12 % con respecto a la misma región. Finalmente, teniendo en cuenta el rango de edad, la población entre 10 y 19 años tiene aproximadamente 2 veces más riesgo de sufrir un delito sexual que la población de referencia entre 0 y 9 años. Observe que con respecto a la edad entre 20 y 29 años, el modelo no permite decidir, dado que esta categoría no es estadísticamente significativa, puesto que su p-valor asociado es mayor que 0,05. Las edades con menor riesgo de padecer un delito sexual se presenta en población adulta entre 50 y 69 años. Finalmente y de acuerdo al ajuste del modelo, se puede manifestar que la población con mayor riesgo de padecer un delito sexual son mujeres de la región centro entre 10 y 19 años de edad.

```
Generalized linear models                         No. of obs      =        320
Optimization     : ML                             Residual df     =        304
                                                  Scale parameter =          1
Deviance         =   240.101584                   (1/df) Deviance =   .7898078
Pearson          =  608.2467758                   (1/df) Pearson  =   2.000812

Variance function: V(u) = u+(.06)u^2              [Neg. Binomial]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =   1.827107
Log likelihood   = -276.3371844                   BIC             =  -1513.468
```

_Conteo	IRR	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexo vic						
Mujer	10.54649	1.835937	13.53	0.000	7.497738	14.83492
semestre						
2	.6380274	.0978158	-2.93	0.003	.4724348	.8616616
3	.6048999	.093957	-3.24	0.001	.446138	.8201584
4	.4576395	.0759356	-4.71	0.000	.3305861	.633523
region hc						
C. Pacífica	.1611782	.0311335	-9.45	0.000	.110379	.2353566
Norte	.2089154	.0472979	-6.92	0.000	.134048	.325597
Sur	.4926059	.0773975	-4.51	0.000	.362043	.6702535
Suroccidente	.435518	.0760868	-4.76	0.000	.3092414	.6133588
rango edad						
10-19	1.93857	.2847753	4.51	0.000	1.453584	2.58537
20-29	.8280844	.143415	-1.09	0.276	.5897338	1.162768
30-39	.2735729	.0652366	-5.44	0.000	.1714329	.436568
40-49	.1314803	.0438833	-6.08	0.000	.0683535	.2529067
50-59	.0624078	.0324884	-5.33	0.000	.0224967	.1731249
60-69	.0471961	.0340571	-4.23	0.000	.0114729	.194151
70 y más	.1342556	.0630126	-4.28	0.000	.0535085	.3368543
cons	.0000446	9.46e-06	-47.25	0.000	.0000294	.0000676
ln(Poblacion)	1	(exposure)				

Finalmente, en la figura 3.3 se exponen las IRR encontradas para cada una de las cinco regiones tomando como referencia la región Central, la cual es la de mayor riesgo. En el gráfico se puede comparar la mayor o menor incidencia de denuncias.

FIGURA 3.3. Mapa de riesgo por región sobre delitos sexuales en Nariño.

APÉNDICE A

Anexos

En este apéndice se presentan algunos complementos con respecto a tablas que no fueron incluidas en la parte central del trabajo.

Inicialmente se muestra las cinco regiones por Municipios en las que se dividió el departamento de Nariño.

Tabla A.1. Regiones por Municipio departamento de Nariño.

Central	**Sur**	**Norte**	**Costa Pacífica**	**Suroccidente**
Ancuya	Aldana	Albán	Barbacoas	Guaitarilla
Buesaco	Contadero	Arboleda	El Charco	Imués
Chachagui	Córdoba	Belén	Francisco Pizarro	La Llanada
Consaca	Cuaspud	Colón Génova	La Tola	Linares
El Peñol	Cumbal	Cumbitara	Magüi	Los Andes
El Tambo	Guachucal	El Rosario	Mosquera	Mallama
Funes	Gualmatan	El Tablón de Gómez	Olaya Herrera	Ospina
La Florida	Iles	La Cruz	Roberto Payán	Providencia
Nariño	Ipiales	La Unión	Santa Bárbara	Ricaurte
Pasto	Potosí	Leiva	Tumaco	Samaniego
Sandona	Puerres	Policarpa		Santacruz
Tangua	Pupiales	San Bernardo		Sapuyes
Yacuanquer		San Lorenzo		Túquerres
		San Pablo		
		San Pedro de Cartago		
		Taminango		

La tabla A.2 muestra los recuentos con su respectiva frecuencia. Estas observaciones se obtuvieron de la matriz de datos original que contaba con 475 registros.

TABLA A.2. Recuentos por frecuencia.

Recuento	**Frecuencia**	**(Frecuecia)·(Recuento)**
0	217	0
1	42	42
2	21	42
3	12	36
4	3	12
5	4	20
6	2	12
7	3	21
8	1	8
9	1	9
10	1	10
11	3	33
12	1	12
14	2	28
16	2	32
20	1	20
27	1	27
33	1	33
38	1	38
40	1	40
Total	**320**	**475**

La prueba de Vuong muestra que el modelo ZIP presenta una mejora ante el MRP. Notese que la variable sexo presente en el modelo inflado de ZIP es estadísticamente significativa, esto indica que la probabilidad que la variable respuesta asuma la categoría de éxito es alta, es decir, existe una alta posibilidad que al observar la variable muestral en la variable sexo se encuentre un cero. La información se presenta en la siguiente página.

Zero-inflated Poisson regression — Number of obs = **320**; Nonzero obs = **103**; Zero obs = **217**

Inflation model = **logit** — LR chi2(**4**) = **281.88**; Log likelihood = **-425.0354** — Prob > chi2 = **0.0000**

_Conteo	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Conteo						
sexo_vic	**1.881587**	**.2125026**	**8.85**	**0.000**	**1.46509**	**2.298085**
semestre	**-.2064393**	**.0433878**	**-4.76**	**0.000**	**-.2914779**	**-.1214008**
region_hc	**-.193979**	**.0381867**	**-5.08**	**0.000**	**-.2688236**	**-.1191345**
rango_edad	**-.4203038**	**.0342979**	**-12.25**	**0.000**	**-.4875265**	**-.3530811**
_cons	**-10.53818**	**.4297917**	**-24.52**	**0.000**	**-11.38056**	**-9.695808**
ln(Poblac~n)	**1**	(exposure)				
inflate						
sexo_vic	**-1.848828**	**.6199502**	**-2.98**	**0.003**	**-3.063908**	**-.6337475**
_cons	**1.735202**	**.9985215**	**1.74**	**0.082**	**-.2218645**	**3.692268**

Vuong test of zip vs. standard Poisson: z = **1.74** Pr>z = **0.0407**

De igual forma, la tabla A.3 muestra todas las posibles combinaciones convergentes, de variables ingresadas a los modelos inflados de ZIP y ZINB. Los mejores modelos son aquellos que presentan menor valor en los criterios AIC y BIC.

Tabla A.3. Criterios de selección para modelos inflados ZIP y ZINB.

Modelos evaluados								1: variable presente y 0: variable ausente			
Id	Modelo	Obs	Modelo nulo	Modelo completo	gl	AIC	BIC	Sexo	Semestre	Región	Rango de edad
1	ZIP	320	-526.8773	-423.3965	10	866.7929	904.4762	1	1	1	1
2	ZIP	320	-560.525	-423.9651	9	865.9302	899.8451	1	1	1	0
3	ZIP	320	-532.8889	-424.1771	9	866.3543	900.2692	1	1	0	1
4	ZIP	320	-528.9311	-423.7757	9	865.5513	899.4662	1	0	1	1
5	ZIP	320	-564.4728	-427.3581	9	872.7161	906.631	0	1	1	1
6	ZIP	320	-564.1977	-424.6432	8	865.2865	895.433	1	1	0	0
7	ZIP	320	-562.2452	-424.3083	8	864.6166	894.7631	1	0	1	0
8	ZIP	320	-534.8775	-424.6083	8	865.2166	895.3632	1	0	0	1
9	ZIP	320	-590.398	-427.4455	8	870.891	901.0375	0	1	1	0
10	ZIP	320	-568.8669	-427.7216	8	871.4432	901.5898	0	1	0	1
11	ZIP	320	-566.0284	-427.4923	8	870.9846	901.1312	0	0	1	1
12	ZIP	320	-565.973	-425.0354	7	864.0709	890.4491	1	0	0	0
13	ZIP	320	-591.7952	-427.5622	7	869.1245	895.5027	0	0	1	0
14	ZIP	320	-570.3842	-427.9135	7	869.827	896.2052	0	0	0	1
15	ZINB	320	-352.5312	-312.2933	11	646.5865	688.0381	1	1	1	1
16	ZINB	320	-371.1427	-313.8269	10	647.6538	685.337	1	1	1	0
17	ZINB	320	-354.1928	-314.0429	10	648.0858	685.769	1	0	1	1
18	ZINB	320	-382.8855	-314.3312	10	648.6624	686.3456	0	1	1	1
19	ZINB	320	-372.7205	-313.8836	9	645.7673	679.6822	1	1	0	0
20	ZINB	320	-372.0437	-314.045	9	646.09	680.0049	1	0	1	0
21	ZINB	320	-357.2755	-314.1184	9	646.2367	680.1516	1	0	0	1
22	ZINB	320	-398.7232	-314.3312	9	646.6624	680.5773	0	1	1	0
23	ZINB	320	-384.2158	-314.3312	9	646.6624	680.5773	0	1	0	1
24	ZINB	320	-373.5959	-314.1203	8	644.2407	674.3872	1	0	0	0
25	ZINB	320	-398.7275	-314.3312	8	644.6624	674.809	0	1	0	0
26	ZINB	320	-399.3666	-314.3312	8	644.6624	674.809	0	0	1	0
27	ZINB	320	-385.4227	-314.3312	8	644.6624	674.809	0	0	0	1

Los casos que no aparecen en la tabla corresponden a modelos no convergentes.

Conclusiones y recomendaciones

Conclusiones

Se muestran algunas conclusiones que responden al propósito del trabajo desarrollado.

- Las estadísticas realizadas en el primer capítulo permiten describir que las variables: sexo, rango de edad y región del hecho, son los factores que mejor se asocian con el riesgo de ocurrencias de delitos sexuales en una determinada población.
- La prueba LR permitió establecer que las variables independientes: sexo, rango de edad, región de hecho y periodo de tiempo, son estadísticamente significativas, es decir, explican la variable respuesta.
- Las variables independientes que mejor explican la variable respuesta en la evaluación de los modelos son: sexo, rango de edad y región del hecho. Esto significa que el riesgo de ocurrencia de un delito sexual en una determinada población de N, depende en su mayor parte de estos tres factores.
- Con base en las razón de tasas IRR presentadas en el capítulo tres, la tendencia en el tiempo es a disminuir levemente el riesgo de violencia sobre delitos sexuales.
- Después de realizar pruebas exhaustivas con los modelos se puede inferir que aquellos modelos que mejor representan los datos son el MRBN y el modelo ZINB.
- La prueba de razón de verosimilitud (RV) permite evidenciar en gran parte, que los datos presentan sobredispersión, por ende lo mas recomendable es utilizar un modelo que contemple este fenómeno.
- Al evaluar el modelo ZINB en su parte inflada, se pudo revelar que la variable *sexo* ingresada al modelo no explica el exceso de ceros en la variable respuesta, esto porque el p-valor del coeficiente de regresión es superior al punto crítico 0.05.
- La prueba de Vuong mostró una mejora significativa del modelo ZIP con respecto al MRP. Sin embargo, para los modelos binomiales negativos la prueba no decide.
- La prueba de razón de verosimilitud (RV) permitió establecer estadísticamente que los modelos binomiales negativos se ajustan mejor a los datos que los Poisson.
- En datos sobredispersos, los modelos de Poisson pueden conducir a conclusiones erróneas dado que algunas veces parecen ajustarse bien cuando no es cierto.

- Los criterios AIC y BIC permitieron elegir como "mejor modelo"el MRBN.
- Las mujeres son aproximadamente 10 veces más vulnerables que los hombres en ser objeto de un delito sexual.
- El rango de edad con mayor tasa de incidencia se encuentra entre 10 y 19 años.

Recomendaciones

Se presentan una serie de aspectos para emprender investigaciones similares o fortalecer el estudio realizado.

- Estudiar la teoría del modelo de regresión de Poisson Generalizado Inflado con Ceros (ZIGP), para conocer con más detalle si este modelo es una buena alternativa a datos que contemplen una gran cantidad de ceros y sobredispersión en la variable respuesta.
- Modelar el número de consultas que realiza un paciente al médico durante un determinado tiempo mediante un modelo de regresión para datos de recuento.
- Considerar en la evaluación del modelo de regresión binomial negativo, variables explicativas adicionales como: tipo de agresor y convivencia o no de la víctima con el presunto agresor para observar detalles adicionales a los presentados en este trabajo.
- Evaluar los modelos de regresión ZIP y ZINB en una muestra relativamente grande, o en su lugar, en una muestra que supere la considerada en este trabajo para observar si al aumentar el tamaño de individuos los modelos convergen.

Bibliografía

[1] A.P. Acero Alvarez, *Informes periciales sexológicos: Violencia sexual contra la pareja*, (2009), Pages 161–167.

[2] A. Agresti, *Categorical data analysis*, New York, Wiley, 2002.

[3] A. Almasi, M.R. Eshraghian, A. Moghimbeigi, A. Rahimi, K. Mohammad, and S. Fallahigilan, *Multilevel zero-inflated generalized poisson regression modeling for dispersed correlated count data*, Statistical Methodology **30** (2016), 1–14, cited By 0.

[4] G. Baetschmann and R. Winkelmann, *Modeling zero-inflated count data when exposure varies: With an application to tumor counts*, Biometrical Journal **55** (2013), no. 5, Pages 679–686, Cited By :2.

[5] S. Boira, P. Carbajosa, and C. Marcuello, *Partner violence from three perspectives: Victims, abusers, and professionals*, **22** (2013), no. 2, Pages 125–133, Cited By :6.

[6] A.C. Cameron and P.K. Trivedi, *Econometric models based on count data. comparisons and applications of some estimators and tests*, Journal of Applied Econometrics **1** (1986), no. 1, Pages 29–53, Cited By :862.

[7] A.C. Cameron and P.K. Trivedi, *Microeconometrics using stata*, Stata Press, 2009.

[8] G.C. Canavos, *Applied probability and statistical methods:*, Little, Brown, 1984.

[9] R.B. Christopher and M.L. Thomas, *Analysis of categorical data with r*, Texts in Statistical Science, 2015.

[10] F. Felix and S.P. Karan, *Zero-inflated generalized poisson regression model with an application to domestic violence data*, Journal of Data Science **4** (2006), no. 1, Pages 117–130.

[11] J.M. Hilbe, *Negative binomial regression*, Cambridge University Press, 2011.

[12] D. Lambert, *Zero-inflated poisson regression, with an application to defects in manufacturing*, Technometrics **34** (1992), no. 1, Pages 1–14, cited By 1414.

[13] A.H. Lee, K. Wang, J.A. Scott, K.K.W. Yau, and G.J. McLachlan, *Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros*, Statistical methods in medical research **15** (2006), no. 1, Pages 47–61, Cited By :92.

[14] J.K. Lindsey, *Modelling frequency and count data*, Oxford science publications, Clarendon Press, 1995.

[15] D. Lord and S.R. Geedipally, *The negative binomial-lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros*, Accident Analysis and Prevention **43** (2011), no. 5, Pages 1738–1742, Cited By :21.

[16] P. McCullagh and J.A. Nelder, *Generalized linear models*, Chapman and Hall, 1989.

[17] M.C. Melgar Hiraldo and J.A. Ordaz Sanz, *Aplicación de los modelos inflados de ceros en el análisis de la siniestralidad y el componente de culpabilidad en el seguro de automóviles*, **4** (2010), no. 1, Pages 44–63.

[18] Y. Mouatassim and E.H. Ezzahid, *Poisson regression and zero-inflated poisson regression: application to private health insurance data*, European Actuarial Journal **2** (2012), no. 2, 187–204, cited By 9.

[19] S.M. Mwalili, E. Lesaffre, and D. Declerck, *The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research*, Statistical methods in medical research **17** (2008), no. 2, Pages 123–139, Cited By :44.

[20] J.A. Nelder and R.W.M. Wedderburn, *Generalized linear models*, Journal of the Royal Statistical Society **135** (1972), no. 3, Pages 370–384.

[21] Jana Petrzelová, *El abuso sexual de menores y el silencio que los rodea*, Plaza y Valdés, S.A. de C.V, México, D.F, 2013.

[22] J. Rodríguez-Avi and M. J. Olmo-Jiménez, *A regression model for overdispersed data without too many zeros*, Statistical Papers **58** (2017), no. 3, Pages 749–773.

[23] J.M. Rodríguez Hernandez, *Riesgo de muerte por suicidio en población colombiana 2000-2013*, (2017), Pages1–8.

[24] A. Takahashi and T. Kurosawa, *Regression correlation coefficient for a poisson regression model*, Computational Statistics and Data Analysis **98** (2016), Pages 71–78, Cited By :1.

[25] Q.H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica **57** (1989), no. 2, Pages 307–333, cited By 2188.

[26] O.B. Yusuf, T. Bello, and O. Gureje, *Zero inflated poisson and zero inflated negative binomial models with application to number of falls in the elderly*, **1** (2017), no. 4, Pages 1–7.