UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

DEPARTAMENTO DE ESTADÍSTICA E INFORMÁTICA

"COMPARACIÓN DE LOS MODELOS DE REGRESIÓN POISSON Y BINOMIAL NEGATIVO PARA DATOS DE CONTEO"

Presentado por:

JESÚS EDUARDO GAMBOA UNSIHUAY

TESIS PARA OPTAR POR EL TÍTULO DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

Lima – Perú

2013

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

DEPARTAMENTO DE ESTADÍSTICA E INFORMÁTICA

"COMPARACIÓN DE LOS MODELOS DE REGRESIÓN POISSON Y BINOMIAL NEGATIVO PARA DATOS DE CONTEO"

Presentado por:

JESÚS EDUARDO GAMBOA UNSIHUAY

TESIS PARA OPTAR POR EL TÍTULO DE:

INGENIERO ESTADÍSTICO E INFORMÁTICO

Ms. Víctor Manuel Maehara Oyata
Presidente

Ms. Luz Jeanet Bullón Camarena
Patrocinadora

Mg. Cesar Higinio Menacho Chiok
Miembro

MS. Carlos López de Castilla Vásquez
Miembro

Dedicado a mis padres:

Feliciano Gamboa y Angélica Unsihuay,

Por sus enseñanzas de vida, apoyo incondicional,

constante motivación y comprensión

en cada momento.

AGRADECIMIENTOS

Agradezco a mis maestros, por la formación y enseñanzas brindadas en la escuela y la universidad, lo cual permitió conseguir este logro profesional.

A los profesores del Departamento de Estadística e Informática, en especial a la profesora Luz Bullón por su apoyo y orientación en el desarrollo de la tesis.

Asimismo agradezco a los miembros del jurado, profesores Víctor Maehara, César Menacho, Carlos López de Castilla por sus aportes que fueron de mucho provecho para enriquecer mi investigación.

A mis amigos, compañeros de trabajo y alumnos de la UNALM, por sus sugerencias y opiniones, las cuales fueron valiosas en el proceso de investigación

TABLA DE CONTENIDOS

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El objetivo de esta investigación es presentar y comparar modelos de regresión Poisson y Binomial Negativo, en el contexto de sobredispersión, desde su enfoque clásico, inflado en cero y *hurdle*, para lo cual, en base a la revisión de literatura, se propone una metodología de comparación que se resume en: análisis exploratorio, selección de variables para construir el modelo, interpretación de coeficientes estimados, indicadores de bondad de ajuste y la comparación entre los modelos haciendo uso de la prueba de Vuong y el AIC.

En la primera aplicación se recolectaron variables sobre consumo de cigarros en alumnos ingresantes a la UNALM en el semestre 2012-I, siendo la variable respuesta el número de cigarros consumidos semanalmente. Los modelos con mejor ajuste fueron el modelo binomial negativo, Poisson inflado en cero y *hurdle* Poisson, mediante los cuales se determinó que el consumo regular de bebidas alcohólicas, el entorno de compañeros y la edad del ingresante son los principales factores de riesgo en el consumo de cigarros.

En la segunda aplicación se consideró una data disponible en internet, acerca de la tasa de peces capturados en un lago estatal de Estados Unidos. Los modelos con mejor ajuste fueron los modelos binomial negativo clásico, inflado en cero y *hurdle*, los cuales indicaron que el número de acompañantes y acudir al lago en casa rodante incrementan la tasa de peces capturados por pescador.

En base a los resultados se concluye que la sobredispersión está presente en ambas aplicaciones y el modelo Poisson no resulta adecuado en esos casos, sin embargo no se puede presentar un único mejor modelo alternativo sino que, en la práctica, debe optarse por aquel que brinde un buen ajuste con la menor cantidad de variables predictoras y además de ello, que permita interpretar los resultados según los objetivos del investigador.

Palabras Clave: Datos de conteo, Poisson, binomial negativo, Inflado en cero, hurdle, Comparación, prueba de Vuong, cigarros, pesca.

ABSTRACT

The purpose of this research is to present and compare Poisson and negative binomial regression, in the context of overdispersion, from its classical, zero-inflated and *hurdle* approach, for which, based on literature review, a methodology for comparison is proposed, whose stages are: exploratory analysis, variable selection for building the model, estimated coefficients interpretation, goodness of fit indicators and models comparison using Vuong test and AIC.

For the first application, were collected some variables about cigar consumption in freshmen of the UNALM in the first semester of 2012, being the number of cigarettes consumed by week the response variable. The models with best fit were the negative binomial, zero inflated Poisson and *hurdle* Poisson, by which it was determined that regular consumption of alcohol, the circle of friends who smoke and the age of the freshman are the main risk factors in cigar consumption.

For the second application, it was considered a dataset available on the internet, about rate of caught fishes in a state lake of USA. The models with best fit were the negative binomial, zero inflated negative binomial and *hurdle* negative binomial, which indicated that he number of companions and going to the lake in a camper increase the ratio of caught fishes.

Based on results, it's concluded that overdispersion is present in both applications and Poisson regression is not adequate for these cases, however it's not possible to present a best alternative model, but, in practice, should choose a model with good fit and fewest variables, and besides this model must permit interpret the results according to the purposes of the researcher.

Key Words: Count Data, Poisson, negative binomial, zero inflated, hurdle, comparison, Vuong test, cigars, fishing

INTRODUCCIÓN

La relación entre una variable aleatoria dependiente de naturaleza discreta proveniente de conteo y un conjunto de variables predictoras es usualmente explicada mediante el modelo de regresión Poisson, el cual asume la igualdad entre la media y la varianza de la variable respuesta. Sin embargo, cuando no se cumple este supuesto, siendo la varianza mayor que la media, se dice que existe problema de sobredispersión, cuyo origen se debe a diversos factores, uno de ellos es el exceso de ceros. Si en esta condición se emplea el modelo Poisson, los errores estándar de los coeficientes podrían subestimarse, lo cual invalidaría la inferencia.

El objetivo general de esta investigación es presentar comparativamente modelos de regresión paramétricos alternativos que tengan la capacidad de contemplar el efecto de la sobredispersión por exceso de ceros. Estos modelos realizan la estimación de parámetros desde un enfoque clásico, inflado en cero y *hurdle*: Los modelos clásicos asumen una distribución probabilística (Poisson o Binomial Negativa) en su componente aleatorio, mientras que los modelos inflados en cero asumen la existencia de dos procesos: uno generador de ceros estructurales y otro, de conteos aleatorios (éstos últimos siguen una distribución Poisson o Binomial Negativa); finalmente, los modelos *hurdle* sólo hacen distinción entre ceros y conteos positivos (los cuales siguen una distribución Poisson o Binomial Negativa truncada en cero). En este sentido, los objetivos específicos de la tesis son los siguientes:

(a) Revisar brevemente la literatura acerca de los modelos de regresión para datos de conteo desde su enfoque clásico, inflado en cero y *hurdle*:

En ella se describe, para cada modelo, su origen, la especificación de los componentes, los métodos de estimación de parámetros e interpretación de coeficientes estimados. De manera general, también se realiza una breve revisión sobre indicadores de bondad de ajuste y pruebas estadísticas para la comparación de modelos de regresión para datos de conteo. En el caso específico del modelo Poisson, además se describen las principales pruebas de sobredispersión.

(b) Proponer una metodología de comparación de los modelos de regresión ya mencionados:

Esta propuesta metodológica se resume en cinco pasos: análisis exploratorio, construcción del modelo, pruebas de bondad de ajuste del modelo, comparación del modelo y pruebas de sobredispersión para el caso específico de la regresión Poisson.

(c) Comparar los modelos de regresión para datos de conteo mediante dos aplicaciones:

La primera está referida al número de cigarros consumidos semanalmente en ingresantes a la Universidad Nacional Agraria La Molina. La elección de este grupo de alumnos se dio debido a que la edad de inicio de consumo de tabaco para el 75 por ciento de los fumadores se da a partir de los 17 años (Oficina contra la Droga y el Delito de las Naciones Unidas, 2008) la cual coincide con la edad promedio de ingreso a la universidad en Perú. En la segunda aplicación se utilizó un conjunto de datos disponible en Internet acerca del número de peces capturados por pescadores en un lago estatal de Estados Unidos.

II. REVISIÓN DE LA LITERATURA

2.1 MODELO LINEAL GENERALIZADO

De acuerdo con Agresti (2002), los modelos lineales generalizados (MLG) constituyen una clase de modelos estadísticos caracterizados por la pertenencia a la familia exponencial de la distribución probabilística de los datos a modelar, asimismo permite que la distribución de la variable respuesta sea no normal, pudiendo generarse así una relación no lineal entre ésta y las variables predictoras. Esta relación no lineal se origina al restringir el rango de la variable respuesta, por ello es necesario aplicar transformaciones en ella o lo que es más común, emplear funciones de enlace pertinentes (Hardin & Hilbe, 2012). Además, en los MLG a diferencia de los modelos de regresión lineal (modelo lineal general), las observaciones no necesariamente deben presentar varianza común (Dobson, 2002).

2.1.1 COMPONENTES DE LOS MLG

Los MLG definen la relación entre la variable respuesta y sus covariables a través de tres componentes esenciales: el predictor lineal, el componente aleatorio y la función de enlace (Agresti, 2002).

a. Predictor lineal

También es conocido como componente sistemático (Agresti, 2002), recoge la información que las variables predictoras aportan al modelo para explicar la variabilidad de la variable respuesta (**Y**) (Azen, R, Walker, CM, 2010). Se trata de una combinación lineal de parámetros **β** cuyos coeficientes se encuentran en la matriz de variables predictoras (**X**). Se denotará con la letra griega *Eta* **(η)**, como $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, para $i = 1, 2, ..., n$.

b. Componente aleatorio

Se refiere a la distribución probabilística de la variable respuesta, la cual debe pertenecer a la familia exponencial (Agresti, 2002), entonces debe expresarse de la siguiente manera:

$$f(y_i \mid \theta_i) = h(y_i)\exp\left(\frac{\theta_i T(y_i) - A(\theta_i)}{a(\phi)}\right) \qquad i = 1,2,...,n \qquad (1)$$

Donde:

- θ_i es el parámetro exponencial, especifica una función del (de los) parámetro(s) necesario(s) para la distribución y determina la función de enlace a utilizar (Hilbe, 2011).
- $T(y_i)$ es el estadístico suficiente. Cuando $T(y_i) = y_i \quad \forall i = 1,2,...,n$ se trata de una distribución en su forma canónica (o estándar) y de ese modo θ_i se convierte en el parámetro canónico (Dobson, 2002)
- $A(\theta_i)$ se denomina función de log partición porque es el logaritmo del factor de normalización o función de partición. Asegura que $f(y_i \mid \theta_i)$ resulte una función de probabilidad. Su primera derivada es convexa y la segunda es positiva ya que es la varianza.
- $h(y_i)$ es una cantidad no negativa conocida como constante normalizadora (Cameron y Trivedi, 1998).
- $a(\phi)$ es el parámetro de ruido (Cameron y Trivedi, 1998), también denominado parámetro de escala, el cual toma el valor de uno en los modelos de regresión para datos de conteo (Hilbe, 2011).

La media y varianza de la variable respuesta estarán dadas por (Hilbe, 2011):

$$E[y_i] = \frac{\partial A(\theta_i)}{\partial \theta_i} \qquad V[y_i] = \frac{\partial^2 A(\theta_i)}{\partial \theta_i^2} \qquad i = 1,2,...,n$$

c. Función de enlace

La función de enlace es una función matemática (Dobson, 2002) utilizada para relacionar la media de la variable respuesta con el predictor lineal. En el modelo lineal general se presenta el caso $\mu_i = \eta_i$. Sin embargo en los MLG el valor esperado μ_i y el predictor lineal η_i no se encuentran en la misma escala, por lo que se debe usar una función de enlace, monótona y diferenciable (Agresti, 2002; Hardin & Hilbe, 2012), de la siguiente manera:

$$g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta} \qquad i = 1,2,...,n \qquad (2)$$

De modo que $g^{-1}(\square)$ que es conocida como función respuesta, la cual se obtiene invirtiendo la función $g(\square)$ (Hardin & Hilbe, 2012).

$$E[y_i \mid \mathbf{x}_i] = \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) \qquad i = 1,2,...,n$$

Cuando la función de enlace se deriva directamente de la función de probabilidad de la variable respuesta, se trata de un enlace canónico (Cameron y Trivedi, 1998)

Las principales funciones de enlace empleadas en los MLG (Agresti, 2002) son:

- Identidad: $g(\mu) = \mu$
- Logarítmica: $g(\mu) = \log(\mu)$
- *Logit*: $g(\pi) = \log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$
- C-loglog: $g(\pi) = \log(-\log(1-\pi))$
- Probit: $g(\pi) = \Phi^{-1}(\pi)$, donde Φ^{-1} es la inversa de la función de distribución Normal Estándar

2.1.2 SUPUESTOS

Uno de los principales supuestos en los MLG es la pertenencia de la variable respuesta a la familia exponencial, es decir, la función de densidad o de probabilidad de esta variable debe poder expresarse como en (1); otros supuestos no menos importantes son los siguientes:

a. Independencia estadística

Las observaciones de la variable respuesta deben ser independientes, ello permite que la función de densidad de probabilidad conjunta pueda ser expresada como el producto de las individuales (Cameron y Trivedi, 1998). Debido a que la base para efectuar la estimación mediante máxima verosimilitud es la función de densidad de probabilidad conjunta, los parámetros estimados dependerán en gran medida del cumplimiento de este supuesto. Por el lado de las variables predictoras, las observaciones dentro de cada una de ellas deben ser independientes; asimismo, la correlación que existe entre estas variables no debe ser muy alta con el fin de evitar multicolinealidad (Orme, JG y Orme, TC, 2009).

b. Correcta especificación de la función de enlace

Al momento de presentar la distribución probabilística como parte de la familia exponencial, el parámetro exponencial (θ_i) determina la función de enlace a utilizar (Hilbe, 2011). Ésta debe ser monótona y diferenciable (Hardin & Hilbe, 2012). Breslow (1996) brinda más detalles acerca de este supuesto.

c. Correcta especificación de la función de varianza:

Se debe realizar el modelamiento adecuado de la no equidispersión (por lo general sobredispersión) mediante una función de varianza adecuada. Existen tests que permiten evaluar esta característica. Breslow (1996) brinda más detalles acerca de este supuesto.

2.2 MODELOS CLÁSICOS PARA REGRESIÓN DE DATOS DE CONTEO

Según Kleiber y Zeileis (2008), el conjunto de modelos clásicos para regresión de datos de conteo está compuesto por los modelos Poisson y binomial negativo.

2.2.1 MODELO DE REGRESIÓN POISSON

a. Antecedentes

A mediados del siglo XX, Cochran (1940) fue uno de los primeros en proponer el uso de la regresión Poisson, aplicada en el contexto de los diseños experimentales. Durante la década de 1940 a 1950, los estudios acerca de datos de conteo fueron desarrollados principalmente por G. Beall (1942), M. Bartlett (1947) y F. Anscombe (1949), quienes propusieron transformaciones que estabilicen la heterogeneidad de los datos provenientes de una distribución Poisson.

Entre tanto, Birch (1963) fue el primero en desarrollar un modelo de regresión Poisson de un predictor mediante máxima verosimilitud; luego de esta década, se comenzó a estudiar otros modelos alternativos al modelo Poisson. Uno de los hitos en la historia de los modelos de regresión para datos de conteo se dio cuando Jhon Nelder y R. W. M. Wedderburn (1972) describieron el modelo de regresión Poisson en el marco de los modelos lineales generalizados. Hasta entonces, el principal interés no era el uso de esta metodología sino la relación entre las distribuciones Poisson, binomial negativa y otras como la binomial, chi-cuadrado, gamma, beta, etc; el énfasis cambió desde entonces por estudiar la relación funcional entre la variable de conteo y sus respectivas predictoras.

La distribución Poisson es, actualmente, la distribución probabilística más utilizada para modelar datos de conteo; es además, luego de la distribución normal, una de las más importantes y de mayor uso (Haight, 1967), así como la referente en cuanto a los modelos para datos de conteo (Cameron y Trivedi, 1998).

b. Componentes del modelo de regresión Poisson

El componente aleatorio viene dado por la distribución probabilística de la variable respuesta, que en este caso es Poisson, lo cual puede ser expresado del siguiente modo:

$$f(y_i \mid \mu_i) = \frac{1}{y_i!}\exp\left(-\mu_i + y_i \ln(\mu_i)\right) \qquad i = 1,2,...,n$$

Entonces pertenece a la familia exponencial, ya que según (1) esta última expresión se puede descomponer de la siguiente manera:

$$\begin{aligned} h(y_i) &= \frac{1}{y_i!} \qquad \theta_i = \ln(\mu_i) \\ T(y_i) &= y_i \qquad A(\theta_i) = \exp(\theta_i) = \mu_i \end{aligned} \tag{3}$$

En este caso $T(y_i) = y_i \quad \forall i = 1,2,...,n$, por lo tanto se trata de una distribución en su forma canónica y de ese modo θ_i se convierte en el parámetro canónico.

La media y varianza de la distribución Poisson vienen dadas por:

$$E\left[Y_i\right] = \frac{\partial \exp(\theta_i)}{\partial \theta_i} = \exp(\eta_i) = \mu_i \qquad V\left[Y_i\right] = \frac{\partial^2 \exp(\theta_i)}{\partial \theta^2} = \exp(\eta_i) = \mu_i$$

Por otro lado, para relacionar la media μ_i con el predictor lineal $\mathbf{x}_i'\boldsymbol{\beta}$ se debe utilizar una función que asegure que todos los valores predichos de la variable respuesta no tomen valores negativos. Ya que según (3), $\theta_i = \ln(\mu_i)$, entonces la función de enlace conveniente para la regresión Poisson es la logarítmica, la cual a su vez es canónica, entonces: $\eta_i = g(\mu_i) = \ln(\mu_i) = (\mathbf{x}_i'\boldsymbol{\beta})$, para $i = 1,2,...,n$.

c. Estimación

La estimación de los coeficientes $\boldsymbol{\beta}$ se realiza mediante el método de máxima verosimilitud (Cameron y Trivedi, 1998; Hilbe, 2011) utilizando algoritmos computacionales. Las expresiones de la función de log-verosimilitud y sus respectivas derivadas se muestran en el ANEXO 1.

d. Interpretación de los parámetros estimados

La estimación de $\boldsymbol{\beta}$ conlleva a conocer la distribución condicional de Y_i dado que se conoce $\mathbf{x}_i$, por ende es posible estimar la tasa promedio de ocurrencias (μ_i) y probabilidades de la forma $P(Y_i = y_i \mid \mathbf{x}_i)$. Esta podría ser una interpretación indirecta o derivada de los coeficientes de regresión. Sin embargo es más común interpretar los efectos de la variación de las variables predictoras en la variable respuesta. Los coeficientes de regresión en los modelos de regresión para datos de conteo no tienen la misma interpretación que en los modelos de regresión lineal, pues no indican, directamente, la variación en la variable respuesta ante el cambio unitario en la predictora, sino que éstos pueden interpretarse de dos maneras: aditiva o multiplicativa

- **Efecto aditivo**

El efecto marginal es el **efecto aditivo** que indica el cambio (incremento o disminución) de la media condicional al darse el cambio unitario en la j-ésima variable predictora para la i-ésima observación. Se obtiene mediante la derivada del valor esperado condicional respecto a la j-ésima variable predictora en la i-ésima observación (Hilbe, 2011):

$$\frac{\partial\left[E(y_i \mid \mathbf{x}_i)\right]}{\partial x_{ij}} = \hat{\beta}_j \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}\right) = \hat{\beta}_j E(y_i \mid \mathbf{x}_i) \qquad (4)$$

Donde: $i = 1, 2, 3, \ldots n \qquad j = 1, 2, \ldots, p \qquad \mathbf{x}_i = (x_1, \ldots, x_j, \ldots, x_p)$

Esto quiere decir que un cambio unitario en el j-ésimo regresor (siempre y cuando éste sea cuantitativo) conlleva a la variación de la media condicional en una cantidad de $\hat{\beta}_j E\left(y_i \mid \mathbf{x}_i\right)$, manteniendo los demás variables en valores constantes. Sin embargo, si se evalúa la variación en el valor esperado, al variar en una unidad la j-ésima variable predictora $E\left(y_i \mid x_1,...,x_j+1,...,x_p\right)-E\left(y_i \mid x_1,...,x_j,...,x_p\right)$, no se obtiene un valor igual al obtenido mediante la derivada: $\hat{\beta}_j E\left(y_i \mid \mathbf{x}_i\right)$, ya que la derivada por definición no se aplica directamente para variaciones unitarias en x_j , sino para una variación que tiende a cero. Se elige el término "unitario" por practicidad en la interpretación.

Según la definición de derivada: $f'(x)=\underset{h\to 0}{Lím}\dfrac{f(x+h)-f(x)}{(x+h)-x}$. Luego, si $x_j \to x_j+h$, entonces $E\left[y_i \mid \mathbf{x}_i\right] \to E\left[y_i \mid \mathbf{x}_i\right]+h\beta_j E\left[y_i \mid \mathbf{x}_i\right]$. Como ya se mencionó, se utiliza $h=1$ por practicidad en la interpretación ("ante un cambio unitario…")

Para el caso de que la j-ésima variable predictora sea binaria, el efecto marginal es conocido como "cambio discreto" (Hilbe, 2011) y se obtiene mediante:

$$\frac{P\left(Y_i=y_i \mid x_{ij}=1\right)-P\left(Y_i=y_i \mid x_{ij}=0\right)}{\Delta x_{ij}}=\frac{\Delta P\left(Y_i=y_i \mid \left(x_{ij}=1 \mid x_{ij}=0\right)\right)}{\Delta x_{ij}}$$

Sin embargo, *todas* estas expresiones son válidas para la i-ésima observación, debido a ello es común utilizar el efecto marginal evaluado en el promedio de las variables predictoras.

- **Efecto multiplicativo**

Para analizar este efecto, ya no se trabaja con la diferencia (absoluta) del valor esperado ante la variación unitaria en una variable predictora, sino su cociente (a veces también llamado diferencia relativa):

$$\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_j\left(x_{j1}+1\right)+\ldots+\hat{\beta}_p x_{ip}\right)}{\exp\left(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_j x_{j1}+\ldots+\hat{\beta}_p x_{ip}\right)}=\exp\left(\hat{\beta}_j\right) \qquad (5)$$

Este cociente es conocido como *Rate Ratio* o Razón de Tasas, el cual indica que ante la variación unitaria de x_j, el valor esperado condicional es $\exp\left(\hat{\beta}_j\right)$ veces el valor esperado condicional inicial (sin la variación unitaria de x_j), manteniendo las demás variables en valores constantes. Los términos de interacción y sus respectivos coeficientes $\left(\hat{\beta}_{ik} X_i X_k\right)$, deben interpretarse para cada nivel (i,k) como se verá más adelante en los casos de aplicación.

e. Ventajas y desventajas

El supuesto de igualdad entre la media y la varianza de la variable respuesta representa una de las principales ventajas de utilizar la distribución Poisson para el modelamiento de datos de conteo ya que conlleva a la simplicidad en la estimación de un único parámetro. Sin embargo, esta relación de igualdad se transforma en una limitación pues dicha situación rara vez se observa en la realidad. Por lo general, la distribución de los datos de conteo presenta sobredispersión, es decir la varianza es mayor que el valor esperado, en ocasiones debido al exceso de ceros (Mc Cullagh y Nelder, 1989). Aunque es menos común en la realidad, también pueden presentarse casos de infradispersión (Hinde y Demétrio, 2007). Ismael y Jemain (2007) señalan que el uso inadecuado de la distribución Poisson, conlleva a obtener estimadores puntuales de $\boldsymbol{\beta}$ válidos, pero éstos pueden tener errores estándar subestimados, lo

cual exagera la significancia de los parámetros de regresión, en consecuencia la inferencia que se realiza se tornaría inválida.

Existen diversas pruebas para evaluar la equidispersión en un modelo de regresión para datos de conteo. Una de las pruebas más básicas, de carácter descriptivo, consiste en comparar la media y la varianza muestrales. Si el ratio media-varianza es uno y/o como máximo dos, se puede estar frente al caso de una regresión Poisson, de lo contrario debe optarse por otro modelo de regresión que contemple el efecto de ausencia de equidispersión (Cameron y Trivedi, 1998). Con mayores detalles, más adelante se analizarán otros indicadores y pruebas estadísticas para evaluar la equidispersión en un conjunto de datos.

Uno de los primeros métodos para lidiar con la sobredispersión fue el uso de estimadores más robustos obtenidos de la matriz de covarianza *sandwich* o la aplicación de un modelo *QuasiPoisson*, el cual no restringe un parámetro de dispersión de la distribución Poisson (Agresti, 2002). No obstante, mediante este método semiparamétrico, los coeficientes se tornan ineficientes ya que presentan mayor variabilidad muestral que la necesaria (Allison, 1999). Por ello a pesar de su relativa simplicidad respecto al uso de otros modelos, no es del todo recomendable su aplicación. Además de ello, en esta investigación sólo se abordará el enfoque paramétrico.

Una segunda opción para remediar casos sin equidispersión consistiría en aplicar transformaciones para lograr que el conjunto de datos distribuya normalmente y con varianza constante (Slymen, Ayala, Arredondo, & Elder, 2006). La transformación razonable para datos de conteo es emplear la función logaritmo, pero añadiendo un valor pequeño, por ejemplo 0.001, ya que el logaritmo de cero es un valor indefinido (King, 1989). Sin embargo, efectuar esta transformación no garantiza que se elimine la inflación en algún otro valor distinto de cero. Además, según Agresti (1996), muchas veces aplicar transformaciones conllevan al nuevo problema que consiste en la obtención de residuales no normales, por lo que no podría tratar el caso como un modelo lineal general.

Entonces, por lo expuesto, aplicar una transformación logarítmica a los datos para tratar de modelarlos como en el caso de una regresión lineal no es la mejor opción. Por ello, se tiene como alternativa el uso de un conjunto de modelos de regresión que pueden manejar la sobredispersión en el conjunto de datos:

- Utilizar un modelo de regresión binomial negativo, el cual se obtiene como una extensión de la distribución Poisson. (Jang, 2005).

- Emplear un modelo de regresión más flexible como el modelo para datos de conteo inflados en cero, el cual es una extensión para los modelos de regresión Poisson y binomial negativo en los que está presente la sobredispersión debido al exceso de ceros. (Winkelmann, 2008)

- Usar un modelo *hurdle*, que puede considerarse, en cierta medida, como una generalización de los modelos inflados en cero, ya que también considera casos de deflación (porcentaje de ceros menor al esperado). Sin embargo, los modelos *hurdle* e inflados en cero no comparten las mismas características en la generación de datos y planteamiento del modelo. (Winkelmann, 2008)

2.2.2 MODELO DE REGRESIÓN BINOMIAL NEGATIVO

a. Antecedentes

Según Todhunter (1865) fue Pierre de Montmort, en 1713, el primero en utilizar la distribución binomial negativa como el número de fallas antes del k-ésimo evento mediante una serie de experimentos Bernoulli, sin embargo su verdadero potencial no sería estudiado sino hasta inicios del siglo XX (Hilbe, 2011), cuando Eggenberger y Polya (1923) obtuvieron la distribución binomial negativa como una mixtura de distribuciones, es decir una distribución Poisson cuya media presenta distribución Gamma. Por ello es también conocida como distribución Poisson-Gamma. Años más tarde, Evans (1953), desarrolló la distribución NB1, una reparametrización de la binomial negativa clásica.

Poco después, Leroy Simon (1960), un científico actuarial, señaló las diferencias entre los modelos Poisson y binomial negativo. Un año después, propuso por primera vez, un algoritmo de máxima verosimilitud para ajustar los modelos de regresión binomial negativo. Entre tanto, Plackett (1981) fue el pionero en desarrollar un modelo de regresión binomial negativo con un predictor mediante máxima verosimilitud.

En 1982, Nelder y Peter McCullagh publicaron la primera edición de *Generalized Linear Models*, texto en el cual documentan sus hallazgos de la última década. La segunda edición, con mayores detalles, especialmente en la regresión binomial negativa, fue lanzada el año 1989.

b. La distribución binomial negativa

Una variable que presenta distribución binomial negativa, tiene dos reparametrizaciones en el planteamiento original de su función de probabilidad, una en función al número de intentos y otra al número de fallas, ambas antes del k-ésimo éxito.

Cuando la variable aleatoria (v.a.) Y es el número de intentos hasta que el r-ésimo evento sucede, en una serie de eventos Bernoulli independientes:

$$f\left(y_i \mid p,r\right)=\binom{y_i-1}{r-1}p^r\left(1-p\right)^{y_i-r} \qquad i=1,2,3,\ldots n \qquad (6)$$

Donde: $0<p<1$ representa la probabilidad de ocurrencia, $y_i \geq r, y_i \in Z^+, r \in Z^+$

Cuando la v.a. Y es el número de fallas antes del r-ésimo suceso, en una serie de eventos Bernoulli independientes:

$$f\left(y_i \mid p,r\right)=\binom{r+y_i-1}{y_i}p^r\left(1-p\right)^{y_i} \quad i=1,2,3,\ldots n \qquad (7)$$

Donde: $0<p<1$ representa la probabilidad de ocurrencia, $y_i \geq 0, y_i \in Z^+, r \in Z^+$

Es más conveniente trabajar con la segunda parametrización (7), ya que el rango de y_i es positivo y no depende de r como en (6). En este caso r es entero y la distribución toma el nombre de Pascal (Hilbe, 2011), sin embargo para el modelamiento de datos, el dominio de r será el conjunto de números reales.

c. Modelos de regresión binomial negativo

Existe una variedad de modelos de regresión binomial negativo. Boswell-Patil (1970) mencionaron hasta 12 derivaciones o variedades de modelos asociados a la distribución binomial negativa, mientras que Hilbe (2011) enuncia 22, sin tomar en cuenta a la distribución geométrica y la Poisson, que pueden considerarse como casos especiales de la binomial negativa. Los modelos de regresión binomial negativa más usuales son (a) el modelo binomial negativo con función de enlace canónica, o NB-C, (b) el modelo binomial negativo con

función de varianza cuadrática $\mu+\alpha\mu^2$, usualmente denominado NB2 y (c) el modelo binomial negativo con función de varianza lineal $\mu+\alpha\mu$, conocido como modelo NB1. De éstos, el segundo es el más empleado en la actualidad, por ello muchas veces la literatura al mencionar "el modelo de regresión binomial negativo" alude al NB2. No obstante, ésta aun no es una notación estandarizada. (Cameron y Trivedi, 1998)

La principal diferencia entre los modelos NB-C, NB2 y NB1 radica en la función de enlace y función de variancia utilizada, así como la matriz de información empleada para la estimación de los parámetros (Hilbe, 2011). El modelo NB-C considera como base la función de probabilidad de la distribución binomial negativa en su forma original, por lo que la función de enlace empleada es la canónica, mientras que los modelos NB2 y NB1 no son canónicos, sino que emplean la función de enlace logarítmica, tal como en una regresión Poisson. Luego, la diferencia entre los modelos NB2 y NB1 se encuentra en la especificación de sus funciones de variancia, mientras que para el clásico NB2 ésta toma la siguiente estructura: $\mu+\alpha\mu^2$, para el modelo NB1 es $\mu+\alpha\mu$; en ambos casos se observa que $\alpha\mu^i$ con $i=1,2$ es la cantidad que indica dispersión, por lo cual α será conocido como el parámetro de dispersión. Finalmente, ya que el modelo NB-C es canónico puede emplear tanto la matriz de información observada o esperada, ya que se obtendrán los mismos resultados, mientras que los modelos NB2 y NB1 deben realizar una corrección en el algoritmo si es que usan la matriz de información esperada de modo que se obtengan errores estándar similares que cuando se usa la matriz de información observada, sin embargo, Hilbe (2011) señala que para estos modelos no canónicos (NB1, NB2) los errores estándar observados son asintóticamente menos sesgados que los errores estándar esperados, pero que las diferencias son insignificantes siempre y cuando se trabajen con tamaños de muestra grandes.

d. Modelo NB-C

Con la finalidad de determinar su pertenencia a la familia exponencial, la expresión (7) puede formularse del siguiente modo:

$$f\left(y_i \mid p,r\right)=\binom{r+y_i-1}{y_i}\exp\left(y_i\ln\left(1-p_i\right)-\left(-r\ln\left(p_i\right)\right)\right)$$

Por lo tanto, se tiene:

$$h(y_i)=\binom{r+y_i-1}{y_i} \qquad \theta_i=\ln\left(1-p_i\right)$$
$$T(y_i)=y_i \qquad A\left(\theta_i\right)=-r\ln\left(1-\exp\left(\theta_i\right)\right) \tag{8}$$

Entonces, la media y varianza de la distribución BN son iguales a:

$$E\left[Y_i\right]=\frac{\partial A(\theta_i)}{\partial\theta_i}=\frac{r\left(1-p_i\right)}{p_i} \qquad V\left[T\left(Y_i\right)\right]=\frac{\partial^2 A(\theta_i)}{\partial\theta_i^2}=\frac{r\left(1-p_i\right)}{p_i^{\,2}}$$

Expresando convenientemente los parámetros iniciales r y p en función de α y μ_i :

$$r=\frac{1}{\alpha} \qquad p_i=\frac{1}{\alpha\mu_i+1} \qquad \mu_i=\frac{\alpha^{-1}(1-p_i)}{p_i} \tag{9}$$

También se puede obtener la varianza, que a diferencia del modelo de regresión Poisson, presenta una cantidad adicional a μ_i, $\left(\alpha\mu_i^2\right)$, la cual es un indicador de sobredispersión:

$$\sigma_i^2=\frac{r(1-p_i)}{p_i^2}=\frac{\alpha^{-1}\left(\frac{\alpha\mu_i}{\alpha\mu_i+1}\right)}{\left(\frac{1}{\alpha\mu_i+1}\right)^2}=\mu_i\left(1+\alpha\mu_i\right)=\mu_i+\alpha\mu_i^2 \tag{10}$$

Como se observa en (8) y (9), el enlace canónico queda determinado por una expresión que depende del parámetro α; cuyo valor debe ser $\alpha > 0$, lo cual es un indicador de sobredispersión, mientras que si se obtiene una estimación para $\alpha < 0$ sería un indicador del fenómeno opuesto (infradispersión), sin embargo no es adecuado considerar valores negativos para α ya que la varianza sería negativa cuando $\alpha < -1/\mu_i$ (Hilbe, 2011). La estimación de α por los métodos propuestos en la presente investigación no podrán tomar valores negativos ya que el software R estima el parámetro de dispersión α con el siguiente artificio: $\alpha^* = \ln \alpha$, de modo que $\alpha = \exp(\alpha^*)$ toma siempre valores positivos, sin embargo Cameron y Trivedi (1998) presentan una propuesta de estimación de α mediante la cual se puede llegar a estimaciones negativas para este parámetro.

La función de enlace para el modelo NB-C es:

$$g(\mu_i) = \theta_i = \ln(1 - p_i) = \ln\left(\frac{\alpha\mu_i}{\alpha\mu_i + 1}\right) = \mathbf{x}_i'\boldsymbol{\beta}$$

Luego, la inversa de la función de enlace, es decir el valor esperado, resulta:

$$g^{-1}(\theta_i) = \mu_i = \frac{1}{\alpha^{-1}\left(\exp(-\theta_i) - 1\right)} = \frac{1}{\alpha^{-1}\left(\exp(-\mathbf{x}_i'\boldsymbol{\beta}) - 1\right)}$$

Dadas las observaciones independientes, la función de verosimilitud para el modelo de regresión NB-C, necesaria para la estimación de los coeficientes $\boldsymbol{\beta}$, así como la gradiente y matriz Hessiana se presentan en el ANEXO 1.

Para el modelo NB-C es posible emplear la matriz de información observada como la esperada, ya que al tratarse de un modelo canónico, los resultados son los mismos (Hilbe, 2011). La convergencia puede resultar tediosa si se emplea el algoritmo de Newton Raphson

para estimar los parámetros mediante máxima verosimilitud (usando la matriz de información observada) por lo que es preferible emplear el scoring de Fisher y su algoritmo de mínimos cuadrados iterativamente ponderados (MCIP) para las estimaciones de los parámetros.

A pesar de que pueden darse situaciones en las que las el modelo NB-C se ajusta mejor a los datos que otros modelos como el NB2, el primero de éstos no es apropiado para lidiar con la sobredispersión pues si bien también es un modelo de regresión para datos de conteo, su interpretación mediante efectos aditivos y/o multiplicativos es compleja ya que la relación entre la media y los coeficientes del modelo también incluye al parámetro de dispersión como se observa en $g^{-1}(\theta_i)$. A veces este parámetro es introducido como una constante para simplificar la estimación (Cameron y Trivedi, 1998), sin embargo en la actualidad *softwar*es como R o *SPSS Statistics* ya cuentan con rutinas para estimarlo sin mayores problemas.

e. Modelo NB2

El uso de la distribución binomial negativa en la regresión para datos de conteo se origina debido al principal limitante de trabajar con la distribución Poisson: no recoge la heterogeneidad (latente, no observable) de los datos, ya que asume que la media es idéntica a la varianza. Para hacer frente a ello, se debe especificar $y_i \mid \mathbf{x}_i, \tau_i$ donde τ_i es la heterogeneidad no observada en el i-ésimo individuo. Entonces, la media se encuentra definida como: $\tilde{\mu}_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \omega_i\right) = \mu_i \tau_i$, donde $\exp\left(\omega_i\right) = \tau_i$ es el factor que recoge la heterogeneidad latente, el cual es independiente de las variables predictoras y su valor esperado es uno, de modo que el valor de la media no se ve alterado al añadir dicha heterogeneidad latente (Gurmu y Trivedi, 1996). Además, se debe asegurar que el rango de la distribución de τ_i no tome valores negativos. Por lo tanto:

$$E\left[\tilde{\mu}_i\right] = E\left[\exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \omega_i\right)\right] = E\left[\mu_i\right]E\left[\tau_i\right] = \left(\mu_i\right)\left(1\right) = \mu_i$$

Luego, asumiendo que la variable respuesta presenta distribución fundamental Poisson con parámetro $\mu\tau$, y que el factor que recoge la heterogeneidad, τ, debe ser una variable aleatoria continua con media uno, entonces una posible distribución para τ es Gamma (Gurmu y Trivedi, 1996) con parámetros (ν,ν); por ello a la distribución binomial negativa también se le conoce como distribución Poisson-Gamma (McCullagh&Nelder, 1989). Entonces, si $Y \mid \mu,\tau \sim Poisson(\mu\tau), \tau \sim Gamma(\nu,\nu)$ se obtiene que $Y_i \mid \mu_i \sim BinNeg\left(r=\nu, p=\frac{\mu}{\mu+\nu}\right)$. Para esta distribución se tiene que:

$$E[Y \mid \mu] = \mu \qquad V[Y \mid \mu] = \mu\left(1+\mu\nu^{-1}\right) \tag{11}$$

Se define $\alpha = \nu^{-1}$ como el parámetro de dispersión a estimar. Entonces (11) queda expresado como $\sigma_i^2 = \mu_i + \alpha\mu_i^2$, de modo que permite observar que $\alpha\mu_i^2$ es la cantidad que indica dispersión, del mismo modo que en el modelo NB-C, ver (10).

La principal distinción entre NB-C y NB2 radica en la función de enlace utilizada, la cual en este caso (NB2) no es canónica, sino logarítmica, al igual que en un modelo Poisson. A diferencia del modelo canónico, la función de enlace no depende del parámetro α.

Para efectos de estimación, la ecuación de log verosimilitud es la misma que en el caso de NB-C, con el respectivo cambio en $\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$. La expresión algebraica de la log verosimilitud, la gradiente y matriz Hessiana se presentan en el ANEXO 1. Respecto al parámetro de dispersión α, la metodología básica de los MLG indica que este parámetro debe ser ingresado como una constante (Hinde y Demetrio, 2007) o estimado de algún modo independiente de $\boldsymbol{\beta}$ (Hilbe, 2011). Sin embargo en la actualidad, se pueden emplear diversos métodos para su estimación, entre ellos se tiene: (a) Máxima Verosimilitud o Mínimos Cuadrados Iterativamente Ponderados (Hilbe, 2011), (b) Algoritmo de estimación puntual (Hilbe, 2011) o (c) Método de los Momentos (Ismail & Aziz, 2007).

Cuando la estimación se realiza por el método de máxima verosimilitud, son necesarias la gradiente y Hessiana para α, sin embargo las expresiones obtenidas en este modelo NB2 son más complejas que en el modelo NB-C, por lo que puede no ser conveniente trabajar el Método de Máxima Verosimilitud. Hilbe (2011) propone este otro método de estimación para el parámetro de dispersión α:

1. Modelar usando regresión Poisson.
2. Para este primer modelo, obtener el estadístico chi cuadrado de Pearson: $\chi^2_{Pois} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$. Los detalles de este estadístico se mostrarán más adelante al tratar el tema de bondad de ajuste.
3. Calcular el estadístico de dispersión: $\delta_0 = \frac{\chi^2_{Pois}}{gl}$
4. Calcular la inversa de este estadístico: $\phi_0 = \delta_0^{-1}$ y considerarla como valor inicial para α es decir $\hat{\alpha}_1$
5. Fijar un valor de tolerancia pequeño: $t \to 0$
6. Inicializa un bucle en $j = 1$

6.1. Correr un modelo binomial negativo usando $\hat{\alpha}_1 = \phi_0$

6.2. Calcular el estadístico chi cuadrado para este nuevo modelo $\chi^2_{NB2} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + \hat{\alpha}_j \hat{\mu}_i^2}$

6.3. Luego, la dispersión $\delta_j = \frac{\chi^2_{NB2}}{gl}$

6.4. Calcular el nuevo valor de phi (dispersión): $\phi_j = \delta_j \delta_{j-1}$

6.5. $j = j + 1$ mientras que $|\delta_j - \delta_{j-1}| > t$

El valor de α estimado mediante este algoritmo es similar al obtenido mediante máxima verosimilitud, sin embargo fuerza al estadístico de Pearson a tomar el valor de uno, es decir, este algoritmo es bastante útil solo para el caso de modelos NB2 sin sobredispersión.

También es posible emplear el algoritmo IRLS, el cual solo permite estimar $\boldsymbol{\beta}$, introduciendo el parámetro ancilar α como una constante, sin embargo programas como R han añadido subrutinas que permiten la estimación de éste último. Otra manera de poder estimar α es mediante el método de los momentos, comparando el momento poblacional con el muestral, ambos de segundo orden, este procedimiento se detalla más adelante.

En el modelo NB2, los errores estándar se tornan inconsistentes si se especifica incorrectamente la distribución probabilística para la variable respuesta, lo cual conlleva a utilizar una función de varianza inadecuada.

f. Interpretación de los parámetros estimados

En el modelo NB2 se utiliza la misma función de enlace (logarítmica) que en el modelo Poisson, por lo que la interpretación de los coeficientes de regresión, mediante el uso de probabilidades y/o *Rate Ratios* es la misma. Por otro lado, el parámetro α es un indicador de dispersión. Si toma el valor de 0, significa que no existe sobredispersión en los datos, por lo tanto la media y la varianza son iguales, es decir, la variable respuesta seguiría una distribución Poisson, si toma valores mayores a cero es indicador de Sobredispersión, si toma el valor de uno se trataría del caso de un modelo de regresión geométrico (Gurmu y Trivedi, 1996).

g. Modelo de regresión binomial negativo generalizado

Cameron y Trivedi (1986) consideraron una generalización para el caso anterior de la binomial negativa. Si se observa la expresión (10) y se formula como $\sigma^2 = \mu + \alpha^{-1}\mu^m$, donde $m = 1, 2, \ldots$, es posible añadir el parámetro adicional m en la estimación. Esta flexibilidad en la formulación de la distribución binomial negativa es un punto a favor de ella ya que conlleva a generar una amplia gama de modelos BN, sin embargo en esta investigación solo se abordará el NB2, ya que puede reducirse a un modelo Poisson, posibilitando su comparación directa. Winkelmann (2008) y Hilbe (2011) discuten sobre este modelo generalizado.

h. Ventajas y desventajas de los modelos BN

Su principal ventaja es la flexibilidad en la dispersión de los datos respecto a un modelo Poisson, asumiendo una distribución para la media, y de ese modo ya no es necesario que la media sea igual que la varianza (tal como sucede casi siempre en la realidad).

Una desventaja es la inconsistencia en las sumas de valores observados y esperados, es decir estos pueden no coincidir (ligeramente) (Levine, Lord y Park, 2010). Por otro lado se tiene una mayor complejidad en la estimación debido al parámetro adicional (el de sobredispersión). Además de ello, al comparar un modelo Poisson con un NB, si bien este último se puede ajustar mejor a los datos, es necesario revisar la parsimonia (un mejor ajuste pero con una menor cantidad de parámetros estimados).

Hilbe (2011) señala que el uso de un parámetro de sobredispersión en el modelo Poisson, o el modelamiento vía la regresión binomial negativa no remedia las principales causas de variabilidad (sobredispersión) en los conjuntos de datos. Estas causas son: (a) variables independientes faltantes, (b) interacciones no incluidas o (c) exceso de ceros. Esta última causa es la que se tratará de remediar mediante los modelos inflados y *hurdle*. Las dos primeras se solucionan desde un enfoque más teórico acerca del fenómeno en estudio.

2.3 MODELO DE REGRESIÓN INFLADO EN CERO

Cuando el número observado de ceros difiere significativamente del esperado bajo cierta distribución probabilística, es necesario proponer modificaciones a los clásicos modelos de regresión para datos de conteo, como el Poisson o el binomial negativo. Una de estas es la regresión inflada en cero, adecuada para el modelamiento de datos sobredispersos debido al exceso de ceros en el conjunto de datos. Min y Agresti (2004) definen la inflación en cero como aquellos "datos para los que un modelo lineal generalizado presenta falta de ajuste debido a la presencia desproporcionada de muchos ceros". Si bien Lambert desarrolla esta teoría a profundidad en 1992, Singh (1963) fue el primero en describirla.

La lógica que subyace a este tipo de modelos es la siguiente: "Cuando un equipo de manufactura está correctamente alineado, los defectos son casi imposibles. Pero cuando está desalineado, los defectos pueden ocurrir de acuerdo a una distribución Poisson (…) Una interpretación (para los modelos de regresión inflados en cero) es que pequeños e inobservables cambios en el ambiente causan el movimiento aleatorio del proceso entre el estado de perfección en el cual los defectos son extremadamente raros y el estado de imperfección en el cual los defectos son posibles pero no inevitables" (Lambert, 1992). Dicho de otra manera, para cada registro existen dos posibles procesos de generación de datos, los cuales son determinados por un evento Bernoulli. El primero de los procesos es aquel que sólo genera ceros, y el segundo aquel que genera valores acorde a una distribución, por lo general una Poisson o Binomial Negativa. Se puede asumir que la estructura de generación de ceros, los cuales se conocen como ceros estructurales, sigue una distribución Bernoulli con parámetro π.

Por ejemplo, si se estudia la variable "Número de reclamos resueltos a favor del cliente mensualmente", ésta puede tomar valores enteros mayores o iguales a cero. Esta variable puede tomar el valor cero en dos situaciones: La empresa no recibió ningún reclamo en un mes (cero estructural) o la empresa recibió n reclamos en un mes pero ninguno fue resuelto a favor del cliente (cero aleatorio). Otro ejemplo puede ser el "número de patentes de una

empresa". Si esta variable toma el valor de cero se puede deber a que no ha realizado ninguna invención por lo tanto no tiene qué patentar (cero estructural) o puede que haya realizado n invenciones de las cuales ninguna ha sido patentada (cero aleatorio).

Un modelo de regresión inflado en cero se formula como se muestra a continuación:

$$f(y_i;\mu_i,\pi_i)=\left[\pi_i+(1-\pi_i)\,f(y_i=0\mid\mu_i)\right]I_{\{0\}}(y_i)+\left[(1-\pi_i)\,f(y_i\mid\mu_i)\right]I_{\{1,2,\ldots\}}(y_i) \tag{12}$$

Donde $f(y_i;\mu_i,\pi_i)$ es la función de densidad de la variable respuesta, $f(y_i\mid\mu_i)$ es la función de probabilidad generadora de conteos, μ_i es su parámetro y $0\le\pi_i\le 1$ es el parámetro de la distribución binomial generadora de ceros estructurales (proporción de ceros).

Los modelos inflados en cero que se estudiarán son el modelo Poisson inflado en cero (ZIP) y el modelo binomial negativo inflado en cero (ZIBN). Éstos se caracterizan por trabajar con mixturas, cada una con su propio componente aleatorio, sistemático y función de enlace, como se verá en detalle en cada uno de estos modelos.

2.3.1 MODELO DE REGRESIÓN POISSON INFLADO EN CERO

Luego de Lambert (1992), Cameron y Trivedi (1998) presentaron una mayor discusión acerca del modelo de regresión Poisson inflado en cero. Tal como se expresó en (31), este tipo de modelos cuenta con una proporción π de ceros estructurales y la restante $1-\pi$ para el componente aleatorio, el cual en este caso se trata de la distribución Poisson, por lo que se reemplaza en $f(y \mid \mu)$ en la expresión (12). Luego, la distribución Poisson también puede tomar valor de cero, pero éste no sería estructural sino un cero aleatorio.

a. Componentes del modelo

Para determinar el componente aleatorio, la variable respuesta puede expresarse como:

$$f(y_i;\mu_i,\pi_i)=\left[\pi_i+(1-\pi_i)e^{-\mu_i}\right]I_{\{0\}}(y_i)+\left[(1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right]I_{\{1,2,3,\ldots\}}(y_i)$$

Donde $0\leq\pi_i\leq 1$ es la proporción de ceros estructurales, $\mu_i\geq 0$ es el parámetro de la distribución Poisson, $I_{\{\square\}}(y_i)$ es la función indicadora de y.

El componente aleatorio viene dado por la distribución binomial (Bernoulli) para modelar π_i y la distribución Poisson para modelar $\boldsymbol{\mu}_i$. El parámetro π_i se expresará de la siguiente manera: $\pi_i = f_{zero}(0;\mathbf{z}_i,\boldsymbol{\gamma})=g^{-1}(\mathbf{z}_i'\boldsymbol{\gamma})$, de modo que la función de enlace comúnmente empleada para este parámetro es *logit*. Como se trata de una regresión logística, también es posible emplear la función de enlace probit o c-log log, brindando resultados similares. Por otro lado, para el parámetro μ_i se utilizará la función de enlace logarítmica como se hizo para el modelo de regresión Poisson.

El componente sistemático para modelar la proporción de ceros estructurales será $\mathbf{z}_i'\boldsymbol{\gamma}$ y para el estado de ocurrencia del evento (distribución Poisson), $\mathbf{x}_i'\boldsymbol{\beta}$. Los vectores de covariables $\mathbf{z}_i$ y $\mathbf{x}_i$ no tienen necesariamente que ser distintos.

Para el caso de este modelo, cabe recalcar que μ_i es la media del componente aleatorio asociado a la distribución Poisson, no el valor esperado de la variable aleatoria inflada en cero. El valor esperado y la varianza de la variable respuesta en el modelo de regresión Poisson inflado en cero (ZIP) vienen dados por las expresiones:

$$E\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right] = \left(1-\pi_i\right)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}$$

$$V\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right] = \left(1-\pi_i\right)\left(\mu_i + \pi_i\mu_i^2\right) = \frac{1}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}\left(\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + \frac{\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)^2\right)$$

b. Estimación de los parámetros

Para efectos de estimación, también se hace uso de la función de log-verosimilitud, la cual se muestra en el ANEXO 1. La estimación de los parámetros $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ es recomendable realizarla mediante el algoritmo de maximización EM (Lambert, 1992).

c. Interpretación de los parámetros estimados

Cuando se tienen modelos que involucran más de una distribución probabilística como los modelos inflados en cero o los modelos *hurdle*, se dice que se trabaja con una mixtura de modelos. En estos casos, los parámetros pueden interpretarse independientemente para cada modelo que compone la mixtura (Hilbe, 2011), dependiendo de la distribución probabilística y función de enlace elegida.

Para el caso particular de los modelos inflados en cero, los efectos marginales obtenidos para cada componente reciben el nombre de efectos marginales extensivos e intensivos (R. Winkelmann, 2008). Los primeros hacen referencia al efecto marginal obtenido para el componente de conteo aleatorio, mientras que los intensivos se refieren al efecto marginal para el componente cero estructural. Comúnmente esta opción es la más utilizada, ya que la media general $\left(E\left[y_i \mid \mathbf{x}_i, \mathbf{z}_i\right]\right)$, comúnmente, no es de mayor interés en estos modelos sino la proporción de ceros estructurales, así como la media para los valores aleatorios.

Los coeficientes $\hat{\boldsymbol{\gamma}}$ se interpretarán mediante los efectos marginales intensivos, como en el caso de una regresión logística. También se pueden interpretar mediante sus Odds Ratio (Razón de ventajas, razón de chances u *Odds Ratios*) de ceros estructurales.

Los coeficientes $\hat{\boldsymbol{\beta}}$ no tienen una interpretación directa ya que provienen de una regresión Poisson, entonces debe optarse por los efectos marginales extensivos o las razones de tasa (*Rate Ratios*) como en (4) y (5).

2.3.2 MODELO DE REGRESIÓN BINOMIAL NEGATIVO INFLADO EN CERO

Tomando como base el trabajo de Lambert, Greene (1994) enuncia el modelo de regresión binomial negativo inflado en cero como una alternativa al modelo de regresión Poisson inflado en cero cuando éste último es incapaz de modelar la sobredispersión. Este modelo combina la distribución binomial para la estructura de ceros y la distribución binomial negativa para la distribución de los datos aleatorios.

a. Componentes del modelo

La distribución probabilística para la variable respuesta en este modelo se expresa a continuación:

$$f\left(y_i \mid \pi_i, \mu_i, \phi\right) = \left[\pi_i + \left(1-\pi_i\right)\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right] I_{\{y_i=0\}} + \left[\left(1-\pi_i\right)\binom{y_i+\phi-1}{y_i}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}\right] I_{\{y_i>0\}}$$

Donde $0 \le \pi_i \le 1, \mu_i \ge 0$, $\phi > 0,$ $\phi^{-1} = \alpha$ es el parámetro de dispersión y $\Gamma(.)$ es la función gamma. Si $\pi_i = 0$ se obtiene el modelo de regresión binomial negativo clásico.

El componente aleatorio viene dado por la distribución binomial (Bernoulli) para π_i y la distribución binomial negativa para $\boldsymbol{\mu_i}$. El componente sistemático para los ceros estructurales será $\mathbf{z}_i'\boldsymbol{\gamma}$ y para el estado de ocurrencia del evento (distribución binomial negativa), $\mathbf{x}_i'\boldsymbol{\beta}$. Los vectores de covariables $\mathbf{z}_i$ y $\mathbf{x}_i$ no tienen necesariamente que ser distintos. Las funciones de enlace a utilizarse son similares a las del modelo ZIP, para el caso del componente de conteos aleatorios se tomará como función de enlace la logarítmica, mientras que para el componente binario de ceros estructurales, utilizar *logit*. Sin embargo, utilizando otras funciones de enlace, como la función probit, se obtienen resultados similares (Garay M., 2010).

Por otro lado, se tiene que la media y la varianza de la variable respuesta son, respectivamente:

$$E[Y_i] = (1-\pi_i)\,\mu_i \qquad\qquad V[Y_i] = (1-\pi_i)\left(\mu_i + \pi_i\mu_i^2 + \alpha\mu_i^2\right)$$

b. Estimación de los parámetros

Para estimar los parámetros de este modelo, también se emplea la función log-verosimilitud, la cual se detalla en el ANEXO 1. Una primera alternativa de estimación consiste en maximizar la función de log-verosimilitud sin embargo esta puede no converger a una solución si no se usan valores correctos de inicio, por ello se sugiere utilizar el algoritmo EM (Lambert, 1992), el cual no necesita una segunda derivada, es más estable y los valores estimados no varían según los valores iniciales (Garay, Hashimoto, Ortega y Lachos, 2010).

c. Interpretación de los parámetros estimados

Respecto a la interpretación de parámetros estimados, ésta se realiza mediante los efectos marginales extensivos (tasa de conteos) e intensivos (proporción de ceros estructurales) como en el modelo de regresión ZIP, con la diferencia de que los conteos aleatorios (que incluyen al cero), siguen una distribución binomial negativa. El parámetro de dispersión estimado, $\hat{\alpha}$ mantiene la interpretación de un modelo NB2.

2.4 MODELO *HURDLE*

"La idea subyacente a la formulación de un modelo *hurdle* es un modelo de probabilidad binomial que gobierna el resultado binario sobre si una variable de conteo es cero o una realización positiva. Si la realización es positiva, el "obstáculo" (*hurdle*) es cruzado, y la distribución condicional de los (valores) positivos es gobernada por un modelo para datos de conteo truncado en cero" (Mullahy, 1986).

Mullahy propuso este modelo en la década de los ochenta. Involucra mixturas discretas las cuales tienen una interpretación de un modelo de dos partes, el cual es determinado por una valla u obstáculo que debe ser cruzado. La primera parte corresponde a un modelo binario que brinda la probabilidad de que el ya mencionado obstáculo sea cruzado; y la segunda, un modelo de datos de conteo truncado en determinado valor entero positivo. De acuerdo con el dominio del problema (exceso de ceros), en este caso se fijará en cero, pudiendo tomar cualquier otro valor entero. Incluso, es posible establecer un modelo *hurdle* múltiple, es decir con más de un obstáculo a cruzar.

Zorn (1996) menciona la principal característica diferenciadora de los modelos *hurdle*, los cuales no permiten la ocurrencia de valores iguales a cero una vez que el *hurdle* es cruzado, lo cual sí puede ocurrir en los modelos de regresión inflados en cero, caso en el que los valores iguales a cero pueden darse en cualquier etapa, dicho de otro modo, los modelos *hurdle* separan los valores cero de los positivos, sin marcar diferencia entre ceros estructurales y aleatorios. Otra importante diferencia es que el modelo *hurdle* no solo es útil para casos de inflación sino también en los de deflación.

Los modelos inflados en cero y *hurdle* no son robustos al identificar un falso proceso de generación de datos, por ello, para no caer en inferencias fallidas se debe identificar correctamente dicho proceso (Miller, 2008).

La distribución de un modelo *hurdle*, se presenta a continuación:

$$f(y_i) = f_{cero}(0) I_{\{0\}}(y_i) + \left(1 - f_{cero}(0)\right)\left(\frac{f_{conteo}(y_i)}{1 - f_{conteo}(0)}\right) I_{\{1,2,\ldots\}}(y_i)$$

La cual equivale a un modelo de regresión (para datos de conteo) clásico si $f_{cero}(0) = f_{conteo}(0)$ (Cameron y Trivedi, 1998).

Las funciones de enlace comúnmente empleadas en los modelos *hurdle* son las siguientes:

Cuadro N° 1: Funciones de enlace en los modelos *hurdle*

Componente Cero		Componente de Conteo	
Distribución Probabilística	Función(es) de enlace	Distribución Probabilística	Función(es) de enlace
Binomial	*Logit*, Probit, c-loglog	Poisson (truncada en cero)	Logarítmica
Poisson (censurada a la derecha en 1)	Logarítmica	Binomial negativa (truncada en cero)	Logarítmica
Binomial negativa (censurada a la derecha en 1)	Logarítmica		

FUENTE: Cameron y Trivedi (1998), Winkelman (2008), Hilbe (2011)

Puede ser posible cualquier combinación para los componentes cero y de conteo (Hilbe, 2011). Se asume que para un modelo *hurdle*, no existen observaciones "censuradas", sino sólo para su componente cero, de la siguiente manera: Se define el estado de la variable aleatoria Y mediante la variable binaria C: si $C = 0$ significa que $Y = 0$, mientras que si $C = 1$, significa que el obstáculo (*hurdle*) fue cruzado, es decir $Y > 0$. A diferencia del modelo inflado en cero, C es una variable observable ya que si la variable respuesta presenta un valor positivo entonces $C = 1$, de lo contrario $C = 0$, no existiendo de ese modo algún valor perdido o desconocido para C.

Por otro lado, el valor esperado y la varianza de los modelos *hurdle*:

$$E\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right] = \left(1 - f_{cero}\left(0\right)\right) \sum_{y_i=1}^{\infty} \frac{y_i f_{conteo}\left(y_i\right)}{1 - f_{conteo}\left(0\right)}$$

$$V\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right] = E\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right] + \left(\frac{f_{cero}\left(0\right) - f_{conteo}\left(0\right)}{1 - f_{cero}\left(0\right)} \right) \left(E\left[Y_i \mid \mathbf{x}_i, \mathbf{z}_i\right]\right)^2$$

La varianza tiene una forma similar a la función de varianza del modelo de regresión binomial negativo, la cual se muestra en (10) y (11), con la diferencia que el término análogo a α ahora varía de individuo en individuo.

La log-verosimilitud en los modelos *hurdle*, necesaria para la estimación de parámetros se muestra a continuación:

$$\ln L = \sum_{i \in \Omega_0} \ln f_{cero}\left(y_i = 0\right) + \sum_{i \in \Omega_1} \ln\left(1 - f_{cero}\left(y_i = 0\right)\right) + \sum_{i \in \Omega_1} \ln f_{conteo}\left(y_i \mid y_i > 0\right) \tag{13}$$

Donde $\Omega_0 = \left\{i / y_i = 0\right\}$ y $\Omega_1 = \left\{i / y_i > 0\right\}$. Además $\Omega = \Omega_0 \cup \Omega_1 = \left\{1, 2, ..., n\right\}$

La estimación en los modelos *hurdle* implica maximizar por separado los términos de la verosimilitud, el correspondiente a los ceros $\left(l_1(\boldsymbol{\beta})\right)$ y el otro término referente a los valores positivos $\left(l_2(\boldsymbol{\beta})\right)$. (Min y Agresti, 2004; Min, 2003, Cameron &Trivedi, 1998; Mullahy, 1986). Esta separación de la log verosimilitud para cada modelo se muestra en el ANEXO 1.

2.4.1 MODELO *HURDLE* POISSON

En un artículo publicado en la revista *Journal of Econometrics*, Mullahy (1986) propone el modelo *hurdle* Poisson como una modificación a los modelos para datos de conteo ya existentes.

a. Componentes del modelo

De acuerdo a la estructura planteada de manera general en los modelos *hurdle*, la distribución probabilística para la variable respuesta en los modelos *hurdle* Poisson resulta:

$$f\left(y_i\right)=\left(1-\pi_i\right)I_{\{0\}}\left(y_i\right)+\pi_i\left(\frac{\exp\left(-\mu_i\right)\mu_i^{\,y_i}}{\left(1-\exp\left(-\mu_i\right)\right)y_i\,!}\right)I_{\{1,2,\ldots\}}\left(y_i\right)$$

Donde π_i se puede entender como la probabilidad de cruzar "la valla" o "el obstáculo" (traducción literal de *hurdle*) que en este caso es cero.

El componente aleatorio para la proporción de ceros está dado por la distribución binomial, y su función de enlace es *logit*. Luego, π_i no tiene la misma interpretación que en los modelos inflados en cero ya que en este caso representa la probabilidad de "cruzar la valla" (el cero). El uso de esta función de enlace conlleva a obtener el modelo *hurdle logit* – Poisson, no obstante también se puede modelar con la función de enlace probit (modelo *hurdle* probit – Poisson). Sin embargo, si se opta por modelar el conjunto de ceros con una distribución Poisson, es decir si la función de enlace para el componente cero es logarítmica, se obtiene el modelo *hurdle* Poisson – Poisson. Luego, el componente aleatorio para los datos positivos (truncados en cero) viene dado por la distribución Poisson y su función de enlace es la logarítmica, como en el modelo de regresión Poisson.

El valor esperado de la variable aleatoria Y es: $E\left[Y_i \mid \mathbf{x}_i, \pi_i\right] = \dfrac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}$

La varianza para la variable aleatoria Y es:

$$V\left[Y_i \mid \mathbf{x}_i, \pi_i\right] = \left(\frac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}\right) + \left(\frac{1-\pi_i-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}{\pi_i}\right)\left(\frac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}\right)^2$$

Si $\boldsymbol{\gamma} = \boldsymbol{\beta}$ para el mismo conjunto de variables predictoras $\mathbf{z} = \mathbf{x}$ entonces, el modelo *hurdle* Poisson - Poisson se reduce a un modelo Poisson clásico.

b. Estimación de los parámetros

El modo más usual de conseguir dicha maximización es a través de métodos numéricos como Newton Raphson o el algoritmo EM. Las funciones de log-verosimilitud y su descomposición para la estimación se muestran en el ANEXO 1.

c. Interpretación de los parámetros estimados:

Los coeficientes pertenecientes al vector $\hat{\boldsymbol{\gamma}}$ provienen de una regresión logística, en ese caso su interpretación será en términos de *Odds Ratios* o efectos marginales. No obstante también cuando se emplea una regresión Poisson o binomial negativa, censurada en uno, en ese caso es factible emplear los efectos marginales o multiplicativos (razón de tasas). Los coeficientes pertenecientes al vector $\hat{\boldsymbol{\beta}}$, se interpretarán como *Rate Ratios* o Razón de tasas.

2.4.2 MODELO *HURDLE* BINOMIAL NEGATIVO

Los economistas W. Pohlmeier y V. Ulrich (1995), en su artículo publicado en la revista *Journal of Human Resources*, propusieron el modelo *hurdle* binomial negativo como un caso general al modelo *hurdle* Poisson ya presentado años antes por Mullahy. Este modelo combina la flexibilidad de la distribución binomial negativa en los conteos truncados en cero y la posibilidad de estimar la proporción de ceros.

a. Componentes del modelo

Debido a que el modelo binomial negativo clásico más empleado es el NB2, es el que se tomará como base para este modelo *hurdle*. Análogamente al modelo *hurdle* Poisson, se plantea la distribución probabilística para la variable respuesta:

$$f\left(y_i\right)=\left(1-\pi_i\right)I_{\{0\}}\left(y_i\right)+\pi_i\left(\frac{\binom{\alpha^{-1}+y_i-1}{y_i}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}}{1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}}\right)I_{\{1,2,\ldots\}}\left(y_i\right)$$

El componente aleatorio para la proporción de ceros está dado por la distribución binomial mientras que la función de enlace usual es la misma que en el modelo *hurdle* Poisson, entonces se considera la función de enlace *logit*, obteniéndose de esta manera el modelo *hurdle logit* – NB2. No obstante, también se puede emplear la distribución Poisson "censurada" en uno para el conjunto de ceros, siendo entonces la función de enlace empleada la logarítmica. De esta manera se obtiene el modelo *hurdle* Poisson – NB2.

Para el conjunto de datos provenientes del conteo, se emplea la función de enlace logarítmica como en el modelo NB2.

El valor esperado de la variable aleatoria Y es: $E\left[Y_i \mid \mathbf{x}_i, \pi_i\right] = \dfrac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\left(\dfrac{1}{1+\alpha \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^{\alpha^{-1}}}$

La varianza de la variable aleatoria Y es:

$$V\left[Y_i \mid \mathbf{x}_i, \pi_i\right] = \left(\frac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\left(\dfrac{1}{1+\alpha \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^{\alpha^{-1}}}\right) + \left(\frac{1-\pi_i-\left(\dfrac{1}{1+\alpha \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^{\alpha^{-1}}}{\pi_i}\right)\left(\frac{\pi_i \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1-\left(\dfrac{1}{1+\alpha \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)^{\alpha^{-1}}}\right)^2$$

Indicando los respectivos valores para π_i según la función de enlace elegida.

b. Estimación

La log-verosimilitud a maximizar se obtiene segmentándola en dos partes, al igual que en el modelo *hurdle* Poisson, y puede ser obtenida de acuerdo a la expresión (13). Dicha maximización es comúnmente realizada utilizando métodos numéricos como el algoritmo Newton-Raphson; la descomposición de la log-verosimilitud simplifica en parte este proceso y aparece en el ANEXO 1.

c. Interpretación de los parámetros:

Los parámetros se interpretan independientemente en cada modelo que compone la mixtura, dependiendo de la distribución probabilística y función de enlace elegida, de manera similar al modelo *hurdle* Poisson (con *Odds* o Razón de tasas)

2.5 MÉTODOS DE ESTIMACIÓN DE MODELOS

2.5.1 MÉTODO DE MÁXIMA VEROSIMILITUD

Una vez especificado el modelo a estudiar, se debe estimar su(s) parámetro(s).El objetivo de la estimación por Máxima Verosimilitud (MV) es encontrar un punto máximo en la función de verosimilitud, o lo que es más común, en la de log verosimilitud ya que la convexidad de la función de log partición garantiza un máximo global para esta última función (Winkelmann, 2008). La estimación por MV puede realizarse mediante el algoritmo de Newton Raphson, o mediante una modificación de éste, el scoring de Fisher. No obstante, los pasos iniciales en el proceso de estimación son similares; su única diferencia radica en la matriz de información a emplear (Hilbe, 2011), como se verá más adelante.

La primera derivada de la función de log-verosimilitud es conocida como la **gradiente**, mientras que la segunda, como **Hessiana**. Expresadas de modo matricial, se tiene que la gradiente será representada por **U** y la Hessiana por **H** (Matriz observada para la estimación por MV y matriz esperada para el caso de MCIP). Por lo tanto la fórmula iterativa para estimar $\boldsymbol{\beta}$ es:

$$\boldsymbol{\beta}_r = \boldsymbol{\beta}_{r-1} - \mathbf{H}_{r-1}^{-1}\mathbf{U}_{r-1}$$

Donde $\boldsymbol{\beta}_{r-1}$ y $\boldsymbol{\beta}_r$ son las estimaciones de $\boldsymbol{\beta}$ en las iteraciones $r-1$ y r respectivamente, $\mathbf{H}$ es la matriz Hessiana de la función de log verosimilitud y $\mathbf{U}$ es la derivada de primer orden de la función de log verosimilitud respecto a $\boldsymbol{\beta}$, es decir la gradiente, también conocida como *score* (Winkelmann, 2008)

El valor inicial para $\boldsymbol{\beta}$ comúnmente se fija como un vector de ceros o unos, en ocasiones lo que se realiza es correr un modelo de regresión lineal y coger los coeficientes

obtenidos como iniciales, o para el caso de la regresión binomial negativa, se pueden emplear los coeficientes de una regresión Poisson.

La gradiente o el score se define como:

$$\mathbf{U} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \mathbf{y}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\left(\frac{y_i - \mu_i}{a_i(\phi)} \right) \left(\frac{1}{V(\mu_i)} \right) \left(\frac{1}{g(\mu_i)} \right) (x_{ij}) \right] \sim N(\mathbf{0}, \mathbf{I})$$

a. Algoritmo de Newton Raphson

El método de Newton Raphson resuelve directamente la fórmula iterativa $\boldsymbol{\beta}_r = \boldsymbol{\beta}_{r-1} - \mathbf{H}_{r-1}^{-1}\mathbf{U}_{r-1}$ hasta que la diferencia en la log-verosimilitud o el vector de parámetros $\boldsymbol{\beta}$ estimados entre dos iteraciones consecutivas sea menor al nivel de tolerancia fijado. En este caso, la matriz Hessiana es igual a la matriz de información de Fisher observada:

$$\mathbf{H} = -\sum_{i=1}^{n} \left[\frac{x_{ij} x_{ik}}{a_i(\phi)} \left\{ \left(\frac{1}{V(\mu_i)} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 - (\mu_i - y_i) \left\{ \left(\frac{1}{(V(\mu_i))^2} \right) \left(\frac{\partial V(\mu_i)}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 - \left(\frac{1}{V(\mu_i)} \right) \left(\frac{\partial^2 \mu_i}{\partial \eta_i^2} \right) \right\} \right\} \right]$$

Cuando se emplea la función de enlace canónica, los métodos de Newton Raphson y el scoring de Fisher resultan equivalentes ya que el término que se encuentra luego de $(\mu_i - y_i)$ se cancela y queda como resultado el valor esperado de la matriz de información.

Si se emplea el método de Newton Raphson, el vector $\boldsymbol{\beta}$ puede incluir al parámetro de dispersión en los modelos de tipo binomial negativo, aunque también puede estimarse mediante otras rutinas o ser ingresado como un valor constante, como se vio en las páginas 20 y 21.

b. Scoring de Fisher (Mínimos Cuadrados Iterativamente Ponderados)

Se trata de caso particular y simplificado del algoritmo de Newton Raphson pero que es computacionalmente más simple, por ello fue implementado por *Numerical Algorithms Group in the UK* en el *software* GLIM, uno de los primeros programas que contenía el procedimiento para trabajar con MLG durante la década de los 70. En la actualidad sigue siendo es el método más empleado en el contexto de los Modelos Lineales Generalizados

La metodología que hace uso del scoring de Fisher también es conocida como Mínimos Cuadrados Iterativamente Ponderados (MCIP) ya que la expresión final que se obtiene tiene una estructura similar a los mínimos cuadrados ordinarios (como en una regresión lineal) con la diferencia que su resolución no es directa sino mediante iteraciones y añadiendo pesos que dependen de la variancia.

El valor esperado de la matriz de información de Fisher está dado por:

$$\mathbf{I} = \sum_{i=1}^{n} \left[\left(\frac{x_{ij} x_{ik}}{Var\left(y_i\right)} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right]$$

El procedimiento detallado para obtener la gradiente y Hessiana tanto por el método de Newton Raphson como el de scoring de Fisher puede ser consultado en el libro *Categorical Data Analysis* de Agresti (2002) o en *Negative Binomial Regression* de Hilbe (2011).

Es importante mencionar que mediante el algoritmo estándar de MCIP sólo se estima μ, por ello el parámetro de dispersión en los modelos de regresión de tipo binomial negativo debe ser introducido como constante. Como se vio en las páginas 21 y 22 existen algoritmos que permiten estimar este parámetro pero el valor de su estimación a veces difiere del método de Máxima Verosimilitud.

2.5.2 MÉTODO DE LOS MOMENTOS

La estimación mediante el método de los momentos, propuesta por Pearson (1895) genera estimaciones aproximadas a las obtenidas mediante el algoritmo de Máxima Verosimilitud. Su principal desventaja es no estar disponible para algunas distribuciones y sumado a ello no se pueden asegurar las propiedades de optimalidad como en el caso de la estimación máximo verosímil (Cameron y Trivedi, 1998), por ello a veces sólo es empleado para obtener un valor inicial para el método de máxima verosimilitud.

Este método consiste en comparar los momentos poblacionales con las muestrales, tantas veces como parámetros se desee estimar. El primer momento muestral es la media. Se define el momento poblacional como $\boldsymbol{\mu}_1' = \boldsymbol{\mu} = E[\mathbf{Y}] = \exp(\mathbf{x}'\boldsymbol{\beta})$ mientras que su contraparte muestral viene a ser $\mathbf{M}_1' = \bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i$. Además se sabe que $E[\mathbf{e}_i \mid \mathbf{x}_i] = 0$ entonces $E[\mathbf{e}_i\mathbf{x}_i] = 0$. Por lo tanto, la condición poblacional de primer momento queda determinada por $E[(\mathbf{y}_i - \boldsymbol{\mu}_i)\mathbf{x}_i] = E[(\mathbf{y}_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i] = 0$, cuyo respectivo momento muestral es $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \exp(\mathbf{x}_i'\boldsymbol{\beta}))\mathbf{x}_i = 0$ (Cameron y Trivedi, 1998)

Luego, se tiene el segundo momento poblacional $\boldsymbol{\mu}_2' = V[\mathbf{Y}] + (E[\mathbf{Y}])^2$ y el muestral: $\mathbf{M}_2' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i^2$. En general, la expresión a desarrollar mediante el método de los momentos, propuesta por Moore (1986) es: $\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\phi_i V(\mu_i)} = n.$ Sin embargo, el término que se encuentra a la izquierda es el estadístico Chi cuadrado, cuya distribución (Chi cuadrado) presenta $n - p$ grados de libertad, por ello Breslow (1984) propone el siguiente estadístico: $\sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\phi_i V(\mu_i)} = n - p$

2.5.3 ALGORITMO EM

El algoritmo EM (Esperanza – Maximización, o Expectation – Maximization) es una rutina iterativa empleada para la estimación de parámetros en modelos con datos faltantes y/o mixturas de distribuciones. La distribución de los valores observados de la variable respuesta es obtenida tomando la distribución conjunta de todas las variables (con data completa e incompleta) y luego obtener la distribución marginal sobre los datos faltantes.

Garay, Hashimoto, Ortega y Lachos (2010) presentan de manera detallada el procedimiento que sigue el algoritmo EM en el modelo de regresión binomial negativo inflado en cero, pudiendo obtenerse de manera similar el algoritmo para el modelo de regresión Poisson inflado en cero.

a. Ventajas del Algoritmo EM

- Es numéricamente estable en cada iteración. Se garantiza que log verosimilitud se incremente (o al menos no disminuya) en cada iteración, sin embargo no se garantiza que ésta llegue a un máximo global. No obstante, en la mayoría de los casos sí se alcanza este máximo sin importar el valor inicial indicado para los parámetros. Los problemas pueden surgir si la función de verosimilitud es multimodal, en este caso el algoritmo convergerá a un máximo local dependiendo del valor inicial elegido para los parámetros (Moon, 1996).

- Si el valor esperado es obtenido analíticamente, no es complicado de programar; Ramaswamy, Anderson y DeSarbo (1994) presentan el algoritmo de manera genérica para los modelos inflados en cero. No obstante, la rutina EM viene incluida en los principales *softwares* estadísticos especializados como R CRAN (Zeileis, Kleiber y Jackman, 2008)

b. Desventajas del algoritmo EM

Se presentan algunas desventajas del algoritmo EM al momento de estimar parámetros:

- No estima una matriz de covarianzas para los parámetros, por lo que debe estimarse mediante otras vías. (McLahlan and Krishnan, 1997)

- La convergencia puede darse de manera muy lenta en algunos casos, y/o la expresiones para el paso M es compleja, lo cual torna tediosa la maximización. (McLahlan and Krishnan, 1997)

- Si bien en la mayoría de los casos se llega a un máximo global, éste no está garantizado. (Moon, 1996)

2.6 ANÁLISIS DE RESIDUALES

Los residuales cuantifican la desviación existente entre los valores observados y los ajustados. Su uso será, en mayor parte, mediante gráficas. Entre sus principales funciones están las de identificar un ajuste pobre, detectar valores influyentes, *outliers*, *leverages*, así como la incorrecta especificación de un modelo. Se tienen varios tipos de residuales:

2.6.1 RESIDUAL BRUTO

Cameron y Trivedi (1998) definen los residuales brutos como aquellos que resultan de la diferencia entre el valor observado y el valor ajustado o estimado, entonces para el i-ésimo elemento de la muestra se tiene: $r_i = y_i - \hat{\mu}_i$. A diferencia de la regresión lineal en la que $\mathbf{r} \sim N\left(0, \sigma^2\right)$ para muestras grandes, en los modelos para datos de conteo, el residual bruto no es homoscedástico ni simétrico alrededor de cero, incluso en muestras grandes, lo cual conlleva a plantear otros residuales de mayor utilidad como el residual de Pearson. El residual bruto es ortogonal al vector de regresores $\mathbf{x}_i$

2.6.2 RESIDUAL DE PEARSON

Cameron y Trivedi (1998) obtienen el residual de Pearson al dividir el residual bruto entre la varianza estimada de y_i . Este residual, a diferencia del bruto, tiene media cero y es homoscedástico, para muestras grandes, sin embargo su distribución no es simétrica. No obstante, Garay et al. (2010) lo denominan residual estandarizado de Pearson y los plotean en un gráfico de probabilidad normal, por lo que se puede considerar una distribución asintóticamente normal para éstos. El i-ésimo residual de Pearson se calcula como:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},$$

2.6.3 RESIDUAL DEVIANCE

Si la distribución probabilística asumida para la variable respuesta pertenece a la Familia Exponencial puede emplearse los residuales Deviance.

La expresión analítica para su obtención es:

$$r_i^D = signo\left(y_i - \hat{y}_i\right)\sqrt{D_i}$$

Donde D_i es la Deviance calculada para el i-ésimo elemento de la muestra, la cual será presentada y estudiada más adelante.

2.6.4 RESIDUAL ESTANDARIZADO

De acuerdo como Hilbe (2011), la expresión analítica para su obtención consiste en dividir el residual bruto entre $\sqrt{1-h_{ii}}$, así:

$$r_i^{std} = \frac{r_i}{\sqrt{1-h_{ii}}}$$

De esa manera se logra estabilizar la varianza de los residuales. Sin embargo, es necesario definir $\mathbf{h}$ y $\mathbf{W}$ para este caso.

$\mathbf{h}$ es conocido como valor *hat*, es una medida de influencia de cada predictor en el modelo de regresión. h_{ii} **e**s el valor *hat* que se encuentra en la diagonal de la matriz *hat*: $\mathbf{H}$, específicamente en la i-ésima fila.

La matriz $\mathbf{H}$ se define como $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$, donde la matriz $\mathbf{W}$ es definida como: $\mathbf{W} = diag\left\{\frac{1}{V(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right\}$

La matriz *hat* es una matriz idempotente de orden n, cuya traza es igual al número de regresores, de modo que el valor promedio de h_{ii} (valor de la diagonal principal) es $\frac{p}{n}$, siendo p el número de variables y n el tamaño de muestra. Debido a ello, valores de h_{ii} que superen la cantidad $\frac{2p}{n}$ son conocidos como *leverages* (Cameron y Trivedi, 1998)

Pueden darse problemas computacionales para el cálculo de la matriz *hat* ya que se deben lidiar con n^2 elementos, sin embargo solo son necesarios los valores de la diagonal. En todo caso pueden realizarse programas mediante los cuales sólo se obtenga la diagonal y de ese modo no desperdiciar memoria calculando los n^2 elementos de la matriz *hat*.

La cantidad h_{ii} también sirve para estandarizar otros residuales (Hilbe, 2011)

El residual estandarizado de Pearson es: $r_i^{p*} = \frac{r_i^p}{\sqrt{1-h_{ii}}}$ mientras que el residual Deviance estandarizado es: $r_i^{D*} = \frac{r_i^D}{\sqrt{1-h_{ii}}}$ Sin embargo si se consulta a Cameron y Trivedi (1998), denomina a éstos últimos residuales como residuales estudiantizados en vez de estandarizados.

La presencia de elementos de la muestra cuyo *leverage* es alto (según las reglas ya mencionadas) y residuales de Pearson estandarizados (o estudiantizados según Cameron y Trivedi) que exceden el rango de ± 2 indica un pobre ajuste del modelo. (Hilbe, 2011)

2.7 BONDAD DE AJUSTE

Los siguientes indicadores permitirán evaluar el ajuste de cada uno de los modelos que se propongan. Entre los principales se tiene (a) el estadístico Chi Cuadrado de Pearson, (b) la Deviance, (c) Pseudo-R^2 y (d) la prueba de bondad de ajuste Chi Cuadrado.

2.7.1 ESTADÍSTICO CHI CUADRADO DE PEARSON

A inicios del siglo XX, Karl Pearson propuso el uso de este estadístico, el cual se obtiene mediante la suma de los residuales de Pearson al cuadrado:

$$\chi^2 = \sum_{i=1}^{n} \left(r_i^p \right)^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\mu}_i \right)^2}{V(\hat{\mu}_i)}$$

Donde y_i es la i-ésima observación, $\hat{\mu}_i$ es el i-ésimo valor ajustado (evaluado en el estimador máximo verosímil $\hat{\boldsymbol{\beta}}\big)$ y $V(\hat{\mu}_i)$ es su función de varianza.

Bajo hipótesis nula cierta, la distribución asintótica del estadístico de Pearson es Chi cuadrado con $n-p$ grados de libertad, donde n es el tamaño de muestra y p es el número de coeficientes de regresión (parámetros) estimados. Sin embargo, según Cameron y Trivedi (1998), esta aproximación funciona mejor en datos agrupados, y que si los datos no tienen esta característica debería emplearse una corrección propuesta por McCullagh (1989), pero que en la práctica no es muy empleada.

Para un modelo de regresión correctamente especificado, el estadístico de Pearson debería tomar un valor igual al tamaño de muestra, sin embargo debe ser corregido debido a la pérdida de grados de libertad para estimar $\hat{y}_i$, por ello χ^2 debe ser igual a $n-p$, de lo contrario, la distribución probabilística asumida no es adecuada (Winkelmann, 2008). Para el

caso de la distribución Poisson significaría que $E[y_i] \neq V[y_i]$, por ello sería posible plantear la siguiente relación entre el valor esperado y la varianza $V[y_i] = \varphi E[y_i]$, siendo $\varphi = \frac{\chi_p^2}{n-p}$. De lo contrario, tal como se propone en esta investigación, se puede establecer otro modelo de regresión que recoja y atenúe el efecto de sobredispersión (NB2, inflado en cero, *hurdle*).

2.7.2 DEVIANCE

Para los modelos de regresión cuya variable respuesta presenta una distribución que pertenece a la familia exponencial se tiene disponible el estadístico Deviance (Cameron y Trivedi, 1998), el cual mide la diferencia entre la máxima log-verosimilitud posible $(l(y_i))$ y la log verosimilitud del modelo en estudio $(l(\hat{\mu}_i))$. Es decir, puede ser considerada como una medida de discrepancia o falta de ajuste del modelo a los datos. Es expresada como:

$$D = 2\sum_{i=1}^{n}\left(l(y_i) - l(\hat{\mu}_i)\right)$$

Para un modelo adecuado, el estadístico Deviance tiene distribución asintótica Chi cuadrado con $n-p$ grados de libertad, donde n es el tamaño de muestra y p es el número de coeficientes de regresión (parámetros) estimados (Agresti, 2002). Tanto el estadístico de Pearson como la Deviance presentan esta distribución asintótica, entonces su valor esperado es $n-p$; por lo tanto como regla práctica, si el estadístico de Pearson así como la Deviance son similares a los grados de libertad $(n-p)$, entonces el modelo de regresión tenderá a ser el adecuado. (Hinde y Demétrio, 2008).

2.7.3 PSEUDO-R^2

Es una aproximación del coeficiente de determinación para modelos no lineales (en los que se incluyen los MLG), no debe interpretarse de la misma manera que un R^2de un Modelo Lineal General (Cameron y Trivedi, 1998). Existe una diversidad de indicadores Pseudo-R^2.

a. Basados en Sumas de Cuadrados

Se puede definir un Pseudo-R^2 basado en la suma de cuadrados de los residuales:

$$R^2_{RES} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

También puede plantearse un Pseudo-R^2 en función a la suma de cuadrados que explica la regresión:

$$R^2_{EXP} = \frac{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2_{EXP} y R^2_{RES} son equivalentes para un modelo lineal general, no siendo así en los MLG. Ambos pueden tomar valores negativos ya que $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ puede ser mayor que $\sum_{i=1}^{n}(y_i - \bar{y})^2$ en muestras pequeñas, además debido a que el objetivo no es minimizar la suma de cuadrados de los residuales: $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$. Luego, R^2_{RES} tiene como límite superior la

unidad, sin embargo no siempre se cumple esto para R^2_{EXP}. Otra característica importante de estos indicadores Pseudo-R^2 es que su valor puede disminuir a medida que se agregan variables predictoras al modelo, por ello no son los mejores Pseudo-R^2 a elegir. (Cameron y Windmeijer, 1996).

b. Basados en la Deviance

La generalización de las sumas de cuadrados para los MLG es la Deviance. El uso de la Deviance para construir una medida de bondad de ajuste Pseudo-R^2 fue propuesto por Cameron y Windmeijer (1996).

Se puede definir la siguiente relación: $D(\mathbf{y},\bar{\mathbf{y}}) = D(\mathbf{y},\hat{\boldsymbol{\mu}}) + D(\hat{\boldsymbol{\mu}},\bar{\mathbf{y}})$, donde $D(\mathbf{y},\bar{\mathbf{y}})$ es la Deviance de un modelo nulo (sólo intercepto), $D(\mathbf{y},\hat{\boldsymbol{\mu}})$ es la Deviance del modelo ajustado y $D(\hat{\boldsymbol{\mu}},\bar{\mathbf{y}})$ es la variabilidad explicada mediante el modelo.

Utilizando estas definiciones es posible plantear la reducción en la Deviance (la discrepancia o falta de ajuste del modelo a los datos) al incluir los regresores, en otros términos, compara un modelo nulo con el modelo en estudio, de la siguiente manera:

$$R^2_{DEV} = 1 - \frac{D(\mathbf{y},\hat{\boldsymbol{\mu}})}{D(\mathbf{y},\bar{\mathbf{y}})} = \frac{D(\hat{\boldsymbol{\mu}},\bar{\mathbf{y}})}{D(\mathbf{y},\bar{\mathbf{y}})}$$

Este Pseudo-R^2 fluctúa entre cero y uno, y se incrementa a medida que se agregan variables predictoras. (Cameron y Windmeijer, 1996; Cameron y Trivedi, 1998)

c. Basados en la Verosimilitud

Cameron y Windmeijer (1996) proponen el siguiente Pseudo-R^2 en base a la log verosimilitud: $R_l^2 = 1 - \frac{l(\hat{\mu})}{l(\bar{y})}$, donde $l(\hat{\mu})$ es la log verosimilitud del modelo estimado y $l(\bar{y})$ la del modelo de sólo intercepto. Es proporcional a R_{DEV}^2 definido líneas arriba, pero puede resultar negativo en el modelo Poisson.

2.7.4 PRUEBA DE BONDAD DE AJUSTE CHI-CUADRADO

Consiste en comparar la probabilidad promedio de obtener un resultado igual a j $(\hat{p}_j)$ con la frecuencia relativa de observaciones con valores igual a j $(\bar{P}_j)$. Por ejemplo, para el caso de un modelo de regresión Poisson: $P(Y=j) = \hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-\hat{\mu}_i)\hat{\mu}_i^j}{j!}$. Se recomienda realizar esta comparación para $j = 1, 2, ..., m$ siendo m el máximo valor observado en la variable respuesta (o la mayor categoría si lo valores esperados son menores a cinco). Luego, comparar las discrepancias entre $\hat{p}_j$ y $\bar{p}_j$ mediante un ploteo y analizarlas mediante una prueba Chi cuadrado (Morel y Neerchal, 2012):

$$\sum_{j=1}^{m} \frac{(o_j - e_j)^2}{e_j} = \sum_{j=1}^{m} \frac{(n\bar{p}_j - n\hat{p}_j)^2}{n\hat{p}_j} \sim \chi^2_{m-k-1}$$

Donde m es el número de categorías y k es el número de parámetros estimados en el modelo. Para el modelo Poisson se tiene $k=1$ (media), mientras que para el modelo NB2, $k=2$ (media y parámetro de dispersión). Luego, en los modelos Poisson inflado en cero y *hurdle* Poisson, el valor de $k=2$ (media y proporción), y finalmente en los modelos NB2 inflado en cero y *hurdle* NB2 se tiene $k=3$ (media, parámetro de dispersión y proporción).

2.7.5 OTROS CRITERIOS DE BONDAD DE AJUSTE

Flynn y Francis (2009) emplean la log verosimilitud como criterio de bondad de ajuste para comparar modelos para datos de conteo, sin embargo Famoye y Rothe (2001) e Ismail y Jemain (2009) sugieren el uso del AIC ya que penaliza la cantidad de variables predictoras empleadas en el modelo a evaluar. Esta última medida se discutirá en la siguiente sección de Comparación de Modelos. Finalmente, SAS Institute (2008) propone comparar gráficamente las probabilidades predichas y observadas para cada valor de $y = 0, 1, 2, \ldots$

2.8 COMPARACIÓN DE MODELOS

Las siguientes pruebas permitirán la comparación de dos o más modelos, en casos específicos como modelos anidados, no anidados y debido a la presencia de sobredispersión.

2.8.1 COMPARACIÓN DE DOS MODELOS

Las pruebas estadísticas para comparar modelos no son las mismas si es que éstos son anidados o no lo son. Dos modelos son anidados si uno puede ser reducido a otro imponiendo alguna restricción en el vector de parámetros, es decir, uno representa una generalización del otro (Khoshgoftaar y Szabo, 2001). Por el contrario, se trata de modelos no anidados cuando un modelo no puede ser representado como un caso especial del otro (Agresti, 2002).

Para modelos anidados se utiliza técnicas como la razón de verosimilitudes, el test de Wald y el de Lagrange (Khoshgoftaar y Szabo, 2001). Éstas asumen que la estimación de parámetros se realizó por el método de máxima verosimilitud, asimismo son pruebas asintóticas, es decir, sus resultados serán tomados como válidos para muestras grandes, caso para el cual los resultados obtenidos con las tres pruebas propuestas son equivalentes (Cameron y Trivedi, 1998).

Ejemplos de algunos modelos anidados que pueden probarse con estos tests son:

1. El modelo Poisson y el modelo binomial negativo, en este caso se plantea $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$ (Miller J. y Miller M., 2008)
2. El modelo Poisson inflado en cero y el modelo binomial negativo inflado en cero, con la misma restricción que el caso anterior (Miller J. y Miller M., 2008).
3. El modelo *hurdle logit* (o probit) - Poisson y el modelo *hurdle logit* (o probit) - binomial negativo: También se debe plantear la restricción sobre el parámetro de dispersión (Miller J. y Miller M., 2008).

4. El modelo *hurdle* Poisson – Poisson y el modelo *hurdle* Poisson – binomial negativo: Similar al caso anterior, plantear la restricción en el parámetro de dispersión.
5. El modelo *hurdle* Poisson – Poisson y el modelo Poisson, bajo la restricción que $\boldsymbol{\gamma} = \boldsymbol{\beta}$ siempre y cuando $\mathbf{z} = \mathbf{x}$, como se vio en (43) (Winkelmann, 2008)
6. Restricciones lineales para los coeficientes de regresión, por ejemplo $H_0 : \beta_5 = 0$ versus $H_1 : \beta_5 \neq 0$, o $H_0 : \beta_3 - \beta_1 = 0$ versus $H_0 : \beta_3 - \beta_1 \neq 0$. Se pueden probar ambas restricciones por separado así como en forma conjunta.

Para comparar los distintos modelos Poisson y binomial negativo en los ejemplo 1 – 5, se asume que las covariables, factores y/o interacciones son las mismas en ambos modelos y que el único parámetro adicional es el de dispersión.

Para modelos no anidados el método a emplear es el test de Vuong, por ejemplo:

1. El modelo Poisson y el modelo Poisson inflado en cero: Este último modelo se reduciría a un modelo Poisson si y solo si la proporción de ceros estructurales (π) es igual a cero, sin embargo, como $\pi = \dfrac{\exp(\mathbf{z}'\boldsymbol{\gamma})}{1+\exp(\mathbf{z}'\boldsymbol{\gamma})}$, la única manera de que sea cero es cuando $\boldsymbol{\gamma} \approx -\infty$ (Winkelmann, 2008).
2. El modelo NB2 y el modelo NB2 inflado en cero: Ocurre una situación similar que el caso anterior.
3. El modelo *hurdle logit* (o probit) – Poisson y el modelo Poisson, ocurre una situación similar, para que $\pi = 0$ entonces $\boldsymbol{\gamma} \approx -\infty$, además, en caso se diese que $\pi = 0$ no se obtendría un modelo Poisson, sino un modelo Poisson truncado en cero.
4. El modelo *hurdle logit* (o probit) – NB2 y el modelo NB2: Ocurre una situación similar que su análogo para el caso Poisson.

A continuación se presentan las principales técnicas de comparación de modelos tanto anidados como no anidados.

a. Razón de Verosimilitudes

Cameron y Trivedi (2008) definen el estadístico de Razón de Verosimilitudes como:

$$\Lambda = -2\left(l_R - l_{NR}\right) \sim \chi^2_{1-\alpha,k}$$

Donde l_R y l_{NR} son las log-verosimilitudes de los respectivos modelos bajo una hipótesis nula (restringido) y alterna (no restringido) respectivamente, y $\chi^2_{1-\alpha,k}$ es el valor crítico para decidir si se rechaza o no la hipótesis nula (percentil $1-\alpha$ de la distribución Chi cuadrado con k grados de libertad), cuando la hipótesis nula es cierta y cuando el tamaño de muestra es grande (Wilks, 1935, 1938). Los grados de libertad pueden definirse de dos maneras:

1. El número de parámetros desconocidos bajo la hipótesis alterna menos el número de parámetros desconocidos de la hipótesis nula.
2. El número de restricciones impuestas bajo una hipótesis nula correcta.

Para el caso específico de la comparación entre los modelos Poisson y NB2, se está probando la correcta especificación del modelo de regresión Poisson, es decir si se cumple el supuesto distribucional asumido que la media es igual a la varianza (Zorn, 1996; Winkelmann, 1998). Sin embargo, se debe tener un especial cuidado en este caso, dado que al probar la hipótesis $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$, el parámetro se encuentra limitado inferiormente por la hipótesis nula, entonces su estimador sólo puede ser igual o mayor a cero, por lo tanto, Chernoff (1954) y Lawless (1987) demuestran que la distribución del estimador en estudio presenta una probabilidad de 0.5 de ser igual a cero y el 0.5 restante asociado a una distribución Chi cuadrado con k grados de libertad. Es por esta razón, que para probar esta hipótesis se debe realizar un ajuste en el valor crítico, el cual sería $\chi^2_{1-2\alpha,k}$ en vez de $\chi^2_{1-\alpha,k}$ (Cameron y Trivedi, 1998).

Anteriormente se consideraba una desventaja práctica de la prueba de razón de verosimilitudes el hecho de estimar el vector de parámetros para dos modelos (bajo hipótesis nula y bajo hipótesis alterna), sin embargo en la actualidad, la potencia de las computadoras minimizan este hecho. Una gráfica (Fox, 1997) que permite observar el comportamiento de esta prueba es:

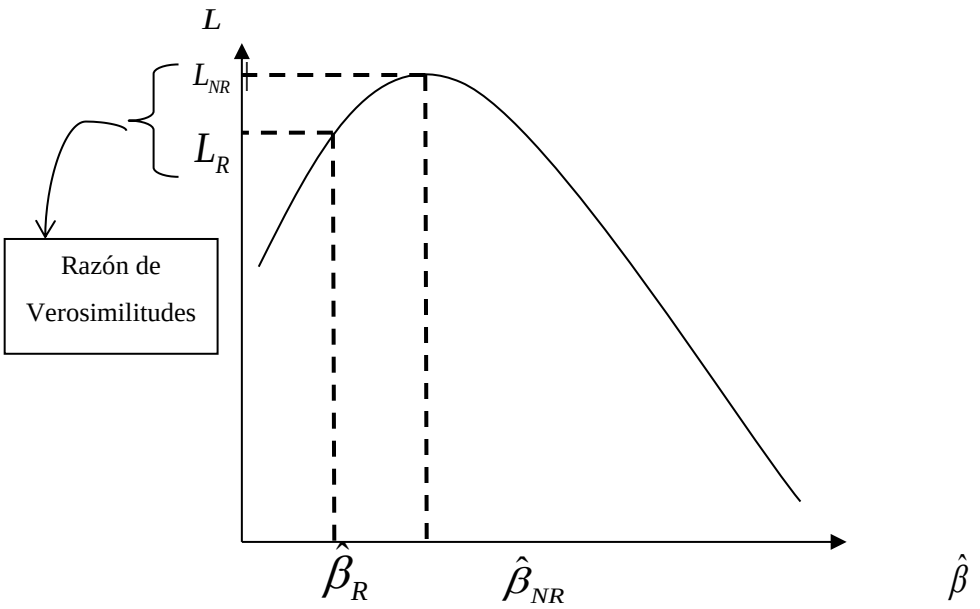

Figura 1: Representación gráfica de la prueba de Razón de Verosimilitudes

b. Prueba de Wald

El test de Wald, cuyo nombre recae en honor a Abraham Wald (1902-1950), es una prueba estadística que a diferencia del test de Razón de Verosimilitudes, no necesita estimar dos modelos (uno bajo cada hipótesis), sino sólo un modelo bajo la hipótesis alterna.

Cameron y Trivedi definen una simbología similar a la siguiente:

- $\boldsymbol{\theta}$ es el vector de parámetros de interés. El test de Wald también puede usarse cuando el interés se encuentra en: (a) una combinación lineal de parámetros: $\mathbf{c}\boldsymbol{\beta}$, donde $\mathbf{c}$ es un vector, (b) varias combinaciones lineales de parámetros: $\mathbf{C}\boldsymbol{\beta}$, donde $\mathbf{C}$ es una matriz, (c) una relación no lineal: $c(\boldsymbol{\beta})$ o (d) varias relaciones no lineales: $C(\boldsymbol{\beta})$ (Cameron y Trivedi, 1998).

- $\boldsymbol{\beta}_0$ es el valor propuesto en la hipótesis. Se tienen las mismas cuatro variantes del caso anterior, dependiendo del tipo y la cantidad de relaciones.
- $\hat{\boldsymbol{\beta}}_{NR}$ es el estimador máximo verosímil para el modelo no restringido (bajo hipótesis alterna), cuya distribución asintótica es $\hat{\boldsymbol{\beta}}_{NR} \overset{a}{\sim} N\left(\boldsymbol{\beta}_0, V\left(\hat{\boldsymbol{\beta}}_{NR}\right)\right)$

En el caso más sencillo que se tiene una única restricción para sólo un parámetro, es decir una hipótesis de la forma $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$ se utiliza el estadístico de prueba: $Z = \dfrac{\hat{\beta}_{NR} - \beta_0}{\sqrt{V\left(\hat{\beta}_{NR}\right)}} \overset{a}{\sim} N(0,1)$, lo cual es equivalente a: $\dfrac{\left(\hat{\beta}_{NR} - \beta_0\right)^2}{V\left(\hat{\beta}_{NR}\right)} \overset{a}{\sim} \chi^2_{(1)}$

También es posible determinar la distribución para una combinación lineal de parámetros, de la forma $\mathbf{c}\hat{\boldsymbol{\beta}}_{NR} - q$, de modo que $\mathbf{c}\hat{\boldsymbol{\beta}}_{NR} - q \overset{a}{\sim} N\left(\mathbf{c}\hat{\boldsymbol{\beta}}_0 - q, \mathbf{c}V\left(\hat{\boldsymbol{\beta}}_{NR}\right)\mathbf{c}'\right)$.

Entonces, para la hipótesis $\begin{matrix} H_0 : \mathbf{c}\boldsymbol{\beta}_0 - q = 0 \\ H_1 : \mathbf{c}\boldsymbol{\beta}_0 - q \neq 0 \end{matrix}$, se define: $\dfrac{\mathbf{c}\hat{\boldsymbol{\beta}}_{NR} - q}{\sqrt{\mathbf{c}V\left(\hat{\boldsymbol{\beta}}_{NR}\right)\mathbf{c}'}} \overset{a}{\sim} N(0,1)$

Aunque también es posible definir: $\left(\mathbf{c}\hat{\boldsymbol{\beta}}_{NR} - q\right)' \left(\mathbf{c}V\left(\hat{\boldsymbol{\beta}}_{NR}\right)\mathbf{c}'\right)^{-1} \left(\mathbf{c}\hat{\boldsymbol{\beta}}_{NR} - q\right) \sim \chi^2_{(k)}$

Donde k es el número de restricciones impuestas bajo la hipótesis nula verdadera, en este caso $k = 1$ porque se trata sólo de una combinación lineal. Se puede generalizar el procedimiento presentado para casos en los que hay más de una combinación lineal (restricción) así como relaciones no lineales.

Gráficamente, Fox (1997) propone:

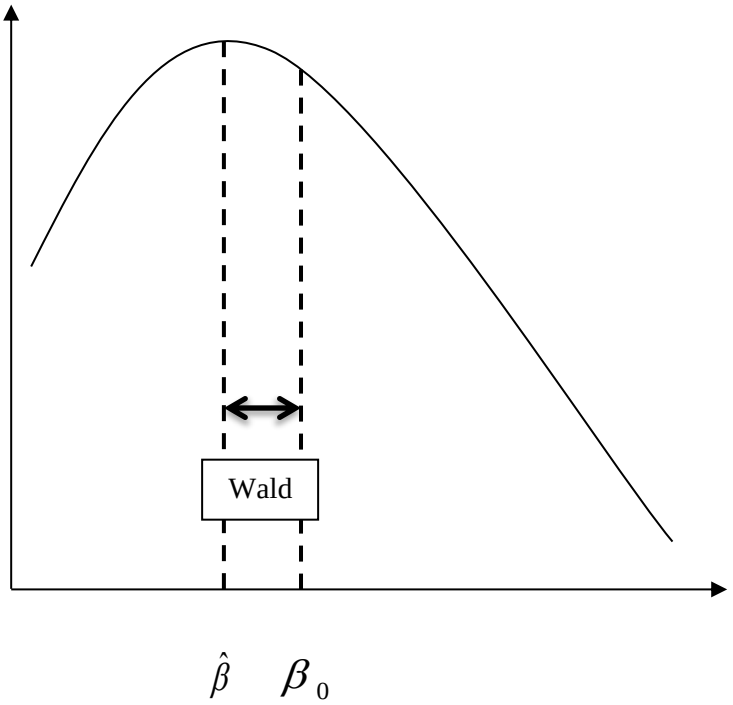

Figura 2: Representación gráfica de la prueba Wald

c. Prueba de los Multiplicadores de Lagrange

Cameron y Trivedi (1998) señalan que es una prueba basada en el score de la función de log-verosimilitud en aquel punto donde es maximizada. El score evaluado en el parámetro fijado en la hipótesis nula debe ser cercano (o igual) a cero, lo cual significa que en el valor de dicho parámetro β_0 se maximiza la función de log verosimilitud, por lo tanto no se rechaza la hipótesis nula.

Luego, dado que la distribución del vector gradiente (o score) bajo hipótesis nula cierta es asintóticamente Normal con media $\mathbf{0}$ y matriz de varianza covarianza $I(\boldsymbol{\beta})$ se tiene que el

estadístico de prueba es: $Lg = \left(\frac{\partial l(\beta_{NR})}{\partial \beta} \right)' \Big|_{\hat{\theta}_R} (\mathbf{I})^{-1} \left(\frac{\partial l(\beta_{NR})}{\partial \beta} \right) \Big|_{\hat{\theta}_R} \overset{a}{\sim} \chi^2_{(k)}$, donde k es el número de restricciones bajo la hipótesis nula verdadera, y además $\mathbf{I} = -E\left(\frac{\partial^2 l(\hat{\beta}_0)}{\partial \beta \partial \beta'} \right)$.

En la siguiente gráfica, para el Test de Lagrange, el eje X está conformado por distintos valores de $\hat{\beta}$ y el eje Y por valores de la función de log-verosimilitud para diferentes $\hat{\beta}$. Se busca encontrar la pendiente igual a cero, lo cual es cierto sólo para el estimador de máxima verosimilitud. A medida, que los valores de $\hat{\beta}$ difieren del estimador máximo verosímil, la pendiente toma valores distintos de cero.

Gráficamente, Fox (1997) propone:

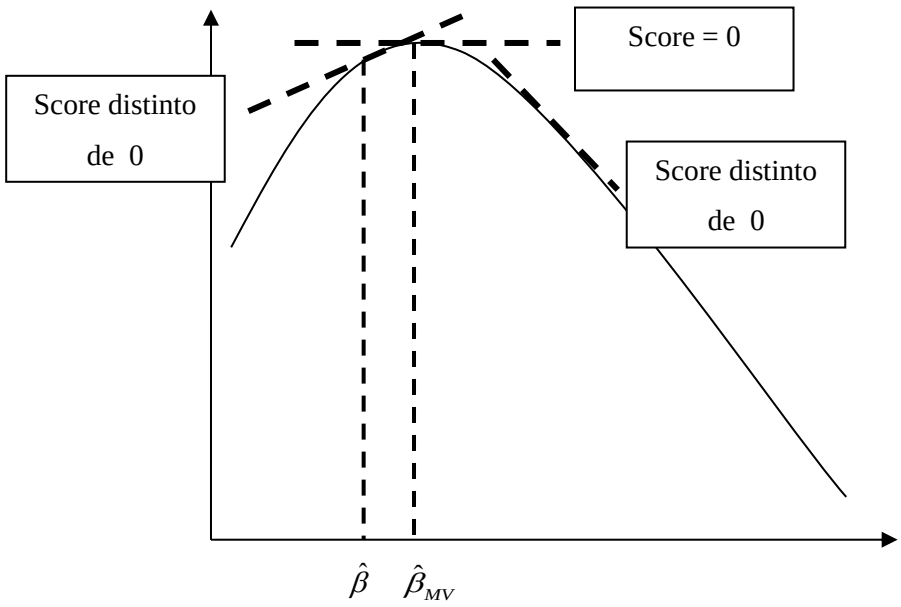

Figura 3: Representación Gráfica de la prueba de los Multiplicadores de Lagrange

d. Test de Voung

Zorn (1996), Cameron y Trivedi (1998), Khoshgoftaar y Szabo (2001), Gao y Khoshgoftaar (2007) citan a Voung (1989), quien en su artículo publicado en la revista *Econometrica*, presenta este test que lleva su apellido, el cual sirve para comparar dos modelos no anidados. Se parte de la siguiente expresión:

$$m_i = \ln\left(\frac{f_1\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)}{f_2\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)}\right) = \ln\left(f_1\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)\right) - \ln\left(f_2\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)\right), \text{ para } i = 1, 2, \ldots, n$$

Donde m_i será el valor a calcular para cada observación, $f_1\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)$ y $f_2\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)$ son las funciones de densidad de los modelos a comparar, evaluadas en la i-ésima observación, es decir, la probabilidad de $y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i$ bajo el modelo cuya función de densidad es $f_k\left(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}_i\right)$. Entonces se plantea la siguiente hipótesis $H_0 : E\left(m_i\right) = 0$ versus $H_1 : E\left(m_i\right) \neq 0$, de modo que H_0 será rechazada siempre y cuando las probabilidades estimadas mediante los modelos 1 y 2, no sean las mismas.

Luego, si se define: $\bar{m} = \dfrac{\sum_{i=1}^{n} m_i}{n}$ y $s_m = \sqrt{\dfrac{\sum_{i=1}^{n}\left(m_i - \bar{m}\right)^2}{n}}$, Voung (1989) señala que el estadístico de prueba $v = \dfrac{\bar{m}}{s_m / \sqrt{n}} \sim N\left(0,1\right)$, de modo que si el valor calculado de v es mayor a $Z_{1-\alpha/2}$, entonces debe optarse por el modelo uno; si es menor a $-Z_{1-\alpha/2}$, se debe elegir el modelo dos, mientras que si $|v| \leq Z_{1-\alpha/2}$ significa que los modelos son similares, y debe emplearse otro criterio para decidir entre ellos. El *software* R brinda directamente el pvalor obtenido para esta prueba de comparación.

2.8.2 COMPARACIÓN DE DOS O MÁS MODELOS

a. Deviance

La Deviance, definida anteriormente, también es útil para la comparación de modelos. Sin embargo, la Deviance no sólo sirve para probar la adecuación de un modelo, sino que también es empleada para la comparación de éstos. Según la definición trivial de Deviance (medida de falta de ajuste del modelo a los datos), se debe elegir el modelo que cuente con menor Deviance (Hilbe, 2011).

Para el caso de dos modelos anidados se puede plantear lo siguiente: Sea M_0 el modelo bajo la hipótesis nula y M_1 el modelo bajo la hipótesis alterna, es decir el modelo para el cual los efectos a probar son distintos de cero. Entonces M_0 está anidado en M_1. Luego, sean D_0 y gl_0 la Deviance y sus respectivos grados de libertad para M_0. De manera análoga, se tiene D_1 y gl_1, la Deviance y grados de libertad para M_1. Finalmente la diferencia de Deviances se distribuye asintóticamente como una Chi cuadrado con $gl_0 - gl_1$ grados de libertad $\left(D_0 - D_1 \overset{a}{\sim} \chi^2_{gl_0 - gl_1} \right)$. Este procedimiento es equivalente al de Razón de Verosimilitudes (Agresti, 2002).

b. Criterios de Información

Permiten comparar dos o más modelos (Hilbe, 2011), sean éstos anidados o no. A medida que se añaden parámetros a dicho modelo, la función de log verosimilitud de éste también se incrementa, debido a ello estos indicadores de criterio de información penalizan el número de parámetros de los modelos a comparar. En general, se opta por elegir el modelo con menor criterio de información (AIC, CAIC, AICc o BIC).

- **AIC**

AIC (Criterio de Información de Aikake), fue propuesto por Hirotsugu Aikake en el año 1947 como un criterio general (no enfocado en los modelos de conteo). Sus resultados son válidos asintóticamente, es decir para muestras de tamaño grande. Se obtiene mediante:

$$AIC = -2(l + p) = -2\ln L + 2p$$

Donde l es a función de log-verosimilitud, L es la función de verosimilitud y p el número de parámetros estimados.

Hilbe (2011) propone una tabla basada en simulaciones que pretende ser sólo una guía, ya que no todas las diferencias posibles entre AICs aparecen en dicha tabla sugerida por Hilbe, la cual se muestra a continuación:

Cuadro N° 2: Diferencias entre AICs

$AIC_B - AIC_A$	Decisión
$(0, 2.5]$	No existe diferencia alguna entre los modelos
$(2.5, 6]$	Preferir el modelo A para tamaños de muestra mayores a 256
$(6, 9]$	Preferir el modelo A para tamaños de muestra mayores a 64
$(10, \infty)$	Preferir el modelo A

FUENTE: Negative Binomial Regression, Hilbe (2011)

Si bien, la regla general induce a elegir el mejor modelo, existe también la alternativa de realizar inferencia multimodelo. Burnham & Anderson (2008) brindan mayores detalles sobre ello.

- **AICc**

AICc es una corrección al AIC para tamaños de muestra pequeños propuesta por Sugiura (1978). Burnham & Anderson (2002) recomiendan su uso para tamaños de muestra pequeños o un número grande de parámetros estimados, específicamente cuando el ratio $n/p < 40$.

Una consecuencia de utilizar AIC en vez del AICc cuando la muestra es pequeña es el incremento en la probabilidad de seleccionar modelos con una mayor cantidad de parámetros, es decir, con sobreajuste.

El criterio AICc se obtiene mediante la siguiente expresión:

$$AIC_c = AIC + \frac{2p(p+1)}{n-k-1} = -2\ln L + 2p + \frac{2p(p+1)}{n-p-1} = -2\ln L + 2p\left(\frac{n}{n-p-1}\right)$$

Para el AICc se pueden emplear los mismos indicadores del AIC $(w_j \text{ y } RE)$.

- **CAIC**

Se trata de otra variante para el AIC. Bozdogan (1987) propuso el uso del criterio de información de Aikake consistente (CAIC), el cual también penaliza el número de parámetros y el tamaño de muestra de la siguiente manera: $CAIC = -2\ln L + p(\ln(n)+1)$

De manera similar al AICc, su uso es de mayor provecho para conjuntos de datos con gran cantidad de parámetros estimados (Nylund, 2007).

- **BIC**

Schwarz (1978) propone una modificación al AIC, la cual incluye una penalización adicional mediante el número de observaciones y la denominó el Criterio de Información Bayesiano (BIC): $BIC = -2l + p\ln(n) = -2\ln L + p\ln(n)$

Adrian Raftery (1986) también propone una expresión para BIC, basada en la Deviance: $BIC_R = D - (n-p)\ln(n)$, donde D es la Deviance del modelo y $(n-p)$ son los grados de libertad del error.

Para esta propuesta, Raftery elaboró una tabla de preferencias, como se muestra a continuación:

Cuadro N° 3: Diferencias entre BIC's

$\lvert\Delta BIC\rvert$	Preferencia por el modelo con menor BIC
0-2	Débil
2-6	Positiva
6-10	Fuerte
10+	Muy fuerte

FUENTE: Negative Binomial Regression, Hilbe (2011)

2.8.3 COMPARACIÓN DE MODELOS USANDO EL CRITERIO DE LA DISPERSIÓN

Estas pruebas se encuentran enfocadas específicamente a la comparación de los modelos Poisson (equidispersión) y NB2 (sobredispersión).

a. Evaluar el ratio media – varianza

Anteriormente se mencionó que si la varianza muestral es más del doble que la media muestral, está presente el problema de sobredispersión (Cameron y Trivedi, 1998), sin embargo de acuerdo con Hilbe (2011), este criterio no permite tomar una decisión ya que dependerá también del tamaño de muestra, por lo que no lo recomienda.

b. Estadístico Chi Cuadrado de Pearson

El estimado muestral para la sobredispersión es la relación proporcional que existe entre la media y la varianza, es decir se tiene que estimar ϕ donde $V[Y]=\phi E[Y]$. Esta estimación también puede realizarse a través de $\varphi=\frac{\chi_p^2}{n-p}$, donde χ_p^2 es el estadístico chi cuadrado de Pearson, n es el tamaño de muestra y p es el número de parámetros estimados en el modelo. Si $\phi>1$ implica Sobredispersión (Ismail & Aziz, 2007).

c. Prueba de Hipótesis para el parámetro de dispersión

Como ya se vio anteriormente al desarrollar la prueba de razón de verosimilitudes (2.8.1.a), se debe probar la hipótesis $H_0:\alpha=0$ versus $H_1:\alpha\neq 0$ donde α es el parámetro de dispersión en un modelo NB2. El rechazo de la hipótesis nula es indicador de sobredispersión.

d. Prueba de sobredispersión de Böhning

Böhning (1994), propone un test de sobredispersión que prueba la siguiente hipótesis $H_0 : \mathrm{E}(\bar{Y}) = E(S^2)$ versus $H_1 : \mathrm{E}(\bar{Y}) \neq E(S^2)$. El estadístico de prueba, construido en base a la media, la desviación estándar y el tamaño muestral es: $O = \sqrt{\frac{n-1}{2}}\left(\frac{s_y^2 - \bar{y}}{\bar{y}}\right) \sim N(0,1)$. También se puede utilizar la aproximación: $O' = n\frac{s_y^2}{\bar{y}} \sim \chi^2_{(n-1)}$. El rechazo de la hipótesis nula es indicador de sobredispersión para la variable respuesta en un modelo Poisson.

e. Pruebas de Cameron y Trivedi

Cameron y Trivedi propusieron un indicador de sobredispersión. En su artículo *Regression Based Tests for Overdispersion* (1985) mencionan la siguiente hipótesis nula $H_0 : E[Y_i] = V[Y_i] = \mu_i$, con su respecta estadística de prueba implementada en el *software* R (función `dispersiontest` en el paquete `AER`).

Cameron y Trivedi también presentaron otras pruebas para verificar la presencia de sobredispersión o infradispersión una vez mediante los residuales de la regresión Poisson.

La primera de esas pruebas (1986) consiste en efectuar una regresión de la varianza estimada sobre la media estimada: $(y_i - \hat{\mu}_i)^2 = \beta\hat{\mu}_i + \varepsilon_i$. Bajo el supuesto de que la media y la varianza son idénticas en un modelo de regresión Poisson, se espera obtener un coeficiente de regresión $\beta = 1$. Se aplica el test de Wald para probar $H_0 : \beta = 1$. Rechazar esta hipótesis es indicador de sobredispersión para la variable respuesta en un modelo Poisson.

Una variante de esta prueba, propuesta también por Cameron y Trivedi (1990) consiste en obtener una ecuación de regresión de la forma $(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i = \beta\hat{\mu}_i^2 + \varepsilon_i$. Bajo el supuesto de

igualdad del primer y segundo momento (media y varianza) de la variable respuesta, se cumple que $E\left[V\left[Y\right]-E\left[Y\right]\right]=0$. Entonces se plantea la hipótesis nula $\beta=0$ utilizando las observaciones (y_i) y predicciones $(\hat{y}_i)$ provenientes del modelo Poisson. Rechazar la hipótesis nula es indicador de sobredispersión para la variable respuesta en un modelo Poisson

f. Prueba de Dean y Lawless

Dean y Lawless (1989) quienes presentan otro test de sobredispersión, el cual utiliza el estadístico de prueba $T=\dfrac{\sum_{i=1}^{n}\left\{\left(Y_i-\hat{\mu}_i\right)^2-Y_i\right\}}{\left(2\sum_{i=1}^{n}\hat{\mu}_i^2\right)^{1/2}}$ cuya distribución asintótica es normal estándar. Valores altos para T indican sobredispersión para la variable respuesta en un modelo Poisson.

2.9 CONSIDERACIONES PARA EL TAMAÑO DE MUESTRA

Signorini (1991) planteó un modo de obtener el tamaño de muestra para el modelo de regresión *Poisson*, haciendo uso de la potencia de prueba, aplicándola para decidir si una covariable tiene influencia en la tasa de eventos o no. La hipótesis nula que se prueba es $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Entonces, la expresión para obtener el tamaño de muestra óptimo es:

$$n\mu_T \exp(\beta_0) \geq \frac{\left(z_\alpha \sqrt{V(\beta_N)} + z_\gamma \sqrt{V(\beta_A)} \right)^2}{\tilde{\beta}^2}$$

Donde:

n : Tamaño de muestra óptimo

μ_T : Intervalo en el cual sucede el evento.

$\exp(\beta_0)$: Tasa media del evento bajo H_0 cierta

$\exp(\beta_1)$: Efecto de la covariable X_1 (*Rate Ratio*)

$V(\beta_N)$: Varianza del vector de covariables bajo la $H_0 : \beta = \beta_N = \left(0 \quad \beta_2 \quad \ldots \quad \beta_p\right)'$

$V(\beta_A)$: Varianza del vector de covariables bajo la $H_1 : \beta = \beta_A = \left(\tilde{\beta} \quad \beta_2 \quad \ldots \quad \beta_p\right)'$

z_α : Cuantil de la distribución normal para el nivel de significación α

z_γ : Cuantil de la distribución normal para la potencia de prueba 1-γ

$\tilde{\beta}$: Valor fijado para β en H_1

Signorini (1991) detalla el procedimiento para obtener $V(\beta_\bullet)$ para el caso de uno y dos o más variables predictoras. Hsieh et al (1998) también proponen una corrección para n, la cual consiste en dividir n entre $\left(1 - R^2\right)$, donde R^2 es el coeficiente de correlación múltiple al efectuar una regresión entre X_1 (la variable con mayor contribución) y las demás covariables.

Cabe mencionar que esta metodología propuesta para la estimación del tamaño de muestra es válida para modelos de regresión Poisson, no siendo necesariamente el mismo para modelos de tipo NB, inflados en cero o *hurdle*.

Algunos estudios en los que se hizo uso de los modelos de regresión para datos de conteo (sin hacer énfasis en la determinación del tamaño de muestra), y que pueden servir como referencia en cuanto al tamaño de muestra, se listan a continuación.

Cuadro N° 4: Estudios previos sobre modelos de regresión para datos de conteo

Área y/o detalles del estudio	Variable en Estudio	Características de la muestra
Investigación sobre accidentes de tránsito. Kuan et al. (1991)	Número de accidentes por conductor.	Tamaño de la muestra: **5422** registros obtenidos del archivo del Departamento de Vehículos Motorizados de California. Estadísticas descriptivas de la variable Respuesta: $\bar{y} = 0.203$, $s_y^2 = 0.2365$ .Se obtuvo 4499 registros con valor cero (82.97 por ciento)
Estudio prospectivo de una escuela en un área urbana de Belo Horizonte. Mendonça and Böhning (1994)	Índice DMF (*Number of Decay, Missing and Filled teeth*)	Tamaño de muestra: **797** niños de colegios del área. Est. Descriptiva de la variable respuesta: $\bar{y} = 3.23$, $s_y^2 = 6.639$. Alrededor del 40 por ciento de los registros son ceros.

Continuación…

Evaluating Zero-Inflated and hurdle Poisson Specifications. Midwest Political Science Association. Abril 1996.	Número de respuestas (acciones) del Congreso de EEUU ante las decisiones de la Corte	Tamaño de muestra: **4052** registros (decisiones tomadas por la Corte entre 1979 y 1988) Est. Descriptiva de la variable Respuesta: $\bar{y} = 0.11,\ s^2 = 0.64$, $Min = 0, Max = 11.$ 95.8 por ciento de los registros son ceros.
The Application of Poisson Random-Effects Regression Models to the Analyses of Adolescents' Current Level of Smoking.	Número de cigarrillos fumados en los últimos siete días.	Tamaño de muestra: 913 estudiantes de 35 escuelas de Los Ángeles y 12 de San Diego. El 68.57 por ciento de la muestra no fumó cigarrillos la última semana.
Estudio automovilístico. Yip &Yau (2001)	Número de reclamos de usuarios de automóviles	Tamaño de muestra: 10000 registros obtenidos del historial de los clientes en los últimos cinco años. Aproximadamente el 60 por ciento de los registros no presentan reclamos (cero).

Continuación…

Estudio llevado a cabo por Department of Conservative Dentistry and Endodontics, A. B. Shetty Memorial Institute of Dental Sciences of coastalen Karnataka, India.	Número de dientes perdidos.	Tamaño de muestra: 2000 individuos mayores de 15 años. El 68.1 por ciento (1362 individuos) no presentó pérdida alguna de dientes.
Determinants Number of Cigarette Smoked with Iranian Adolescents: A Multilevel Zero Inflated Poisson Regression Model. Publicado en Iranian J Publ Health, Vol. 38, No.4, 2009, pp.91-96	Número de cigarillos fumados al día	Tamaño de la muestra: 1745 adolescentes de entre 15 y 20 años residentes en ocho provincias del noroeste de Irán.

… Continuación

Application of Negative binomial Regression for Assessing Public Awareness of the Health Effects of Nicotine and Cigarettes. Investigación realizada por School of Statistics and Actuarial Science,University of KwaZulu-Natal(2010)	Número de mensajes anti-smoking en la publicidad de cigarrillos	Tamaño de muestra: 343 individuos Variable Respuesta: $\bar{y} = 3.09$, $s^2 = 5.99$
On Estimation and Influence Diagnostics for Zero-Inflated Negative binomial Regression Models. (2010)	Número de raíces producidas.	Tamaño de muestra: **270** brotes micropropagados de manzana en su variedad Trajan. El porcentaje observado de ceros es de 23.7 por ciento
Negative binomial Model with an Application to Special Treatment Count Data. *School of Economics of Shanghai University* (2011)	ST (Número de tratamientos especiales, una regulación importante en China).	Tamaño de la muestra: **540** eventos en total, entre Febrero del 2003 y Agosto del 2010.

FUENTE: Elaboración propia

III. MATERIALES Y MÉTODOS

3.1 HIPÓTESIS DE INVESTIGACIÓN

1. Las variables “Número de cigarros consumidos semanalmente” y “Número de peces capturados por pescador” presentan sobredispersión.

2. El modelo Poisson no se ajusta a los datos cuando está presente el problema de sobredispersión.

3. Ante la ausencia de equidispersión, la inferencia en los modelos de regresión Poisson es inválida ya se subestiman los errores estándar de los coeficientes estimados.

4. Ante la ausencia de equidispersión, no existe un único modelo que presente una adecuada bondad de ajuste.

3.2 PROCEDIMIENTO DE ANÁLISIS

El procedimiento de análisis consta de los siguientes pasos: (a) Análisis Exploratorio, (b) Construcción de modelos, (c) Ajuste del Modelo, (d) Comparación de modelos. También se consideran pruebas de sobredispersión en caso se modele una regresión Poisson. Para el desarrollo del análisis se utilizará el *software* R en su versión 2.14.2. La lista de comandos empleados aparece en el ANEXO 17.

3.2.1 ANÁLISIS EXPLORATORIO DE DATOS

Se realiza con la finalidad de conocer los principales indicadores de las variables en estudio, así como para eliminar factores con alguna categoría preponderante (por encima del 80 u 85 por ciento de valores en una sola categoría) y variables con exceso de valores perdidos, también para verificar la independencia entre las variables predictoras.

3.2.2 CONSTRUCCIÓN DE MODELOS

Al momento de construir los modelos, R estima los parámetros mediante el método de máxima verosimilitud con el algoritmo de scoring de Fisher. Para el caso de los modelos inflados en cero, también se vale del algoritmo EM y el algoritmo de optimización BFGS (Broyden – Fletcher – Goldfarb – Shano), este último también utilizado en la estimación de los modelos *hurdle*. El procedimiento a seguir para seleccionar variables en cada modelo es el siguiente:

(a) Obtener un modelo de regresión de la variable respuesta con la regresora cuya contribución sea superior a las demás. Para ello, probar hipótesis de la forma $\beta_j = 0$ mediante el test de Wald y considerando $\alpha = 0.01$ para el ingreso de dicha variable predictora, siempre y cuando ésta sea cuantitativa o categórica binaria, en caso se trate de una variable con más de dos categorías utilizar la estadística Deviance para probar la contribución de esta variable mediante el método del modelo reducido.

(b) Luego de haber ingresado la primera variable predictora, probar cuál es la segunda variable en ingresar, nuevamente aplicando el test de Wald o el modelo reducido como en el paso anterior. De no haber ninguna variable que ingresa al modelo, termina el proceso de selección de variables, de lo contrario, la segunda variable ingresa al modelo y se continúa con el siguiente paso.

(c) Verificar si existe interacción entre estas dos primeras variables; de existir, agregarla al modelo, de lo contrario continuar con el siguiente paso.

(d) Continuar con la búsqueda de la tercera variable que ingresaría al modelo. Cada vez que se agrega una nueva variable, verificar la existencia de interacción de a dos. No se consideran interacciones de a 3 por la complejidad de su interpretación.

(e) Obtener los valores de log verosimilitud y AIC en cada paso, siendo el más importante del modelo final (con las variables ya seleccionadas) ya que permitirá la comparación entre modelos.

(f) Finalmente, una vez seleccionadas las variables pertinentes para cada modelo, y ninguna otra contribuye significativamente, plantear el modelo estimado e interpretar sus coeficientes estimados.

3.2.3 AJUSTE DEL MODELO

Utilizar las siguientes pruebas y/o indicadores para verificar el ajuste del modelo:

- Para la Deviance se prueba la hipótesis nula de que el modelo se ajusta a los datos.
- Para el estadístico Chi cuadrado de Pearson, el cociente de este entre sus grados de libertad debe ser cercano a uno para indicar que el modelo está correctamente especificado, es decir, que se asumió la correcta distribución para los datos.
- Para los residuales de Pearson, tomar nota cuántos de estos sobrepasan el rango de ± 2 (lo ideal sería cero). Para comparar los modelos Poisson y NB2, tomar nota de los *leverages* que sobrepasan el límite $2p/n$ siendo p el número de parámetros estimados y n el tamaño de muestra (del mismo modo, lo ideal sería cero). Las observaciones con altos residuales y/o *leverages* deben ser eliminadas si se verifica que existió algún error en la recolección de datos.
- Para el PseudoR^2 basado en la verosimilitud: A mayor Pseudo R^2, mejor ajuste.
- Para el AIC: Un menor AIC indica un mejor ajuste en el modelo, es un indicador de bondad de ajuste netamente comparativo.

Adicionalmente a lo propuesto, en los modelos inflados en cero es posible predecir: (a) la probabilidad $P(Y = y)$, donde se muestran valores para la variable respuesta (Y) desde cero hasta el máximo que se encuentre en la muestra y (b) la probabilidad de que la observación provenga de un cero estructural. Dado que se conoce si un cero es estructural (nunca fumó) o aleatorio (alguna vez fumó pero no lo hace ahora) pero este dato no es empleado en la estimación del modelo, puede servir para validar si la tasa de clasificación correcta de ceros es adecuada.

3.2.4 COMPARACIÓN DE MODELOS

Adicionalmente a las pruebas y/o indicadores de bondad de ajuste planteados en la sección anterior, utilizar la prueba de razón de verosimilitudes para comparar modelos anidados y el test de Vuong para las comparaciones de modelos no anidados. Además, como ya se mencionó, se empleará el AIC para cualquier comparación.

3.2.5 SOBREDISPERSIÓN

Para el caso específico del modelo Poisson, plantear hipótesis para probar la existencia de sobredispersión, se utilizará el test de Böhning y las pruebas de Cameron y Trivedi, así como la de Dean y Lawless. Resumiendo la metodología en el siguiente diagrama de flujo:

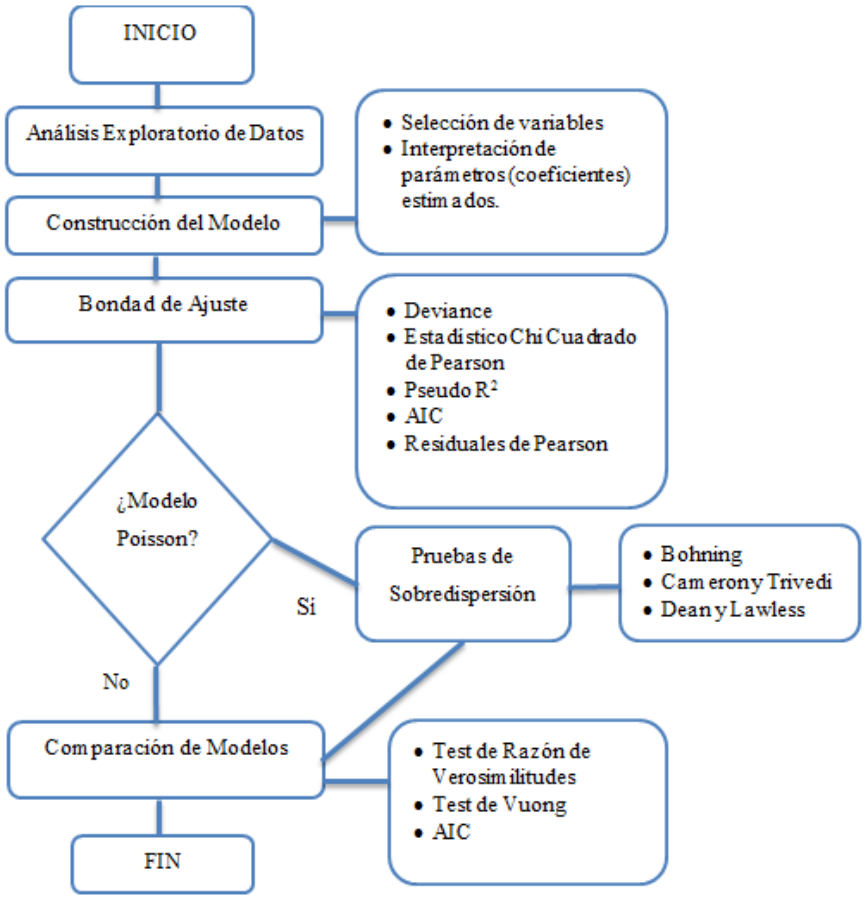

Figura 4: Diagrama de flujo de la metodología propuesta

3.3 APLICACIONES

3.3.1 APLICACIÓN UNO: CONSUMO DE CIGARROS

a. Descripción de la aplicación

El cigarro es el producto más común mediante el cual se consume tabaco. Cada uno de estos contiene más de 4700 sustancias químicas, entre las cuales hay varias que son cancerígenas; una de las más conocidas es la nicotina, la cual es de seis a ocho veces más adictiva que el alcohol y puede generar una dependencia más fuerte que la heroína. En ese contexto, los datos a analizar corresponden al número de cigarros consumidos semanalmente, la cual es una variable que, través de los modelos de regresión para datos de conteo ya expuestos, se relacionará con diversos factores y covariables con la finalidad de obtener estimaciones.

Los datos fueron recolectados en una muestra de 266 ingresantes a la Universidad Nacional Agraria La Molina en el curso de propedéutico (herramientas informáticas) al inicio del semestre 2012-1, sin hacer distinción entre modalidades, a través de una encuesta anónima. La elección del estudio de sólo ingresantes se debe a que la etapa de inicio en los hábitos de fumador se sitúa en la adolescencia, sin embargo la tendencia es que su inicio sea cada vez más precoz. Según Cedro, para este año (2013) el consumo de tabaco en Perú se inicia a los 13 años en promedio. Además, el 22 por ciento de los adolescentes entre los 12 y 18 años ha fumado alguna vez en su vida y el 6.5 por ciento lo ha hecho en el último mes.

b. Variables en estudio

Cuadro N° 5: Lista de variables en estudio para la aplicación uno

Nombre de la variable	Codificación de la variable en R	Tipo de Variable	Descripción
Cigarros semanales	CigarrosSemanales	Cuantitativa Discreta	Número de cigarros que un alumno fuma a la semana.
Sexo	Sexo_	Cualitativa	Condición biológica que distingue hombres de mujeres
Edad	Edad	Cuantitativa continua	Tiempo transcurrido (en años) desde el nacimiento de la persona hasta el momento de la encuesta.
Situación Sentimental	Sít_Sent_	Cualitativa	Estado en el que se encuentra respecto a tenencia de pareja
Prevalencia de vida	Fuma	Cualitativa	Indicador binario (Sí / No) del consumo de cigarros en el periodo que corresponde a toda la vida del alumno. Responde a la pregunta ¿Ha fumado cigarros alguna vez en su vida?
Edad a la que fumó por primera vez	EdadFumo	Cualitativa	Edad a la que fumó por primera vez, categorizada. Categorías: Etapa escolar / Etapa post escolar / Nunca.
Prevalencia de año	CigarroAño	Cualitativa	Indicador binario (Sí / No) del consumo de cigarros en el periodo que corresponde al último año de vida del alumno. Responde a la pregunta: ¿Fumó cigarros durante el último año?

Prevalencia de mes	CigarroMes	Cualitativa	Indicador binario (Sí / No) del consumo de cigarros en el periodo que corresponde al último mes de vida del alumno. Responde a la pregunta ¿Ha fumado cigarros durante el último mes?
Prevalencia puntual	CigarroAhora	Cualitativa	Medición binaria (Sí / No) del consumo de cigarros en el momento de la aplicación de la encuesta. Fumador regular: ¿Fuma actualmente cigarros?
Intento de dejar de fumar	DejarFumar	Cualitativa	Indica si el alumno trató de dejar de fumar cigarros: Categorías: No (Nunca ha fumado) / No (Alguna vez ha fumado) / Sí
Fumadores en casa	FumaCasa	Cualitativa	Indicador binario (Sí / No) que informa si existe algún familiar que fuma regularmente en casa
Fumadores en el entorno de amigos	FumaCompañeros	Cualitativa	Indicador binario (Sí / No) que informa si existe alguien que fuma regularmente en el entorno de sus amigos.
Recepción de charlas	Charlas	Cualitativa	Indicador binario (Sí / No) sobre si el alumno recibió charlas acerca del consumo de cigarros y sus consecuencias.
Consumo de bebidas alcohólicas	Bebidas	Cualitativa	Indicador binario (Sí / No) sobre si el alumno consume bebidas alcohólicas de modo regular, definido como al menos tres fines de semana al mes
Consumo de puros	Puro	Cualitativa	Indicador binario (Sí / No) sobre si el alumno alguna vez ha consumido puros

Consumo de tabaco de mascar	TabacoMascar	Cualitativa	Indicador binario (Sí / No) sobre si el alumno alguna vez ha consumido tabaco de mascar
Consumo de tabaco de pipa	TabacoPipa	Cualitativa	Indicador binario (Sí / No) sobre si el alumno alguna vez ha consumido tabaco de pipa
Consumo de tabaco en polvo	TabacoPolvo	Cualitativa	Indicador binario (Sí / No) sobre si el alumno alguna vez ha consumido tabaco de pipa
Consumo de cigarros electrónicos	CigarroElectronico	Cualitativa	Indicador binario (Sí / No) sobre si el alumno alguna vez ha consumido cigarros electrónicos

FUENTE: Elaboración propia

3.3.2 APLICACIÓN DOS: TASA DE PESCA

a. Descripción de la aplicación

UCLA Academic Technology Services, en su página web, presenta una aplicación para datos de conteo en la que un grupo de biólogos desea modelar el número de peces capturados por pescadores en un lago estatal. Los pescadores son encuestados para recolectar datos acerca del número de peces capturados, número de personas acompañantes, número de niños y si acudieron en casa rodante al parque, sin embargo la encuesta fue aplicada a todos los pescadores, pudiendo darse el caso de que alguno no haya logrado capturar pescar nada porque no pescó (cero estructural) o porque tuvo una mala jornada de pesca (cero aleatorio), sin embargo, a diferencia de la aplicación uno, esta información no se encuentra disponible. Se encuestó a un total de 250 pescadores; la descripción de las variables se muestra en el Cuadro N° 6 que se muestra a continuación.

b. Variables en estudio

Cuadro N° 6: Lista de variables en estudio para la aplicación dos

Nombre de la variable	Codificación de la variable en R	Tipo de Variable	Descripción
Número de peces capturados	count	Cuantitativa discreta	Número de peces capturados por pescador
Número de personas	persons	Cuantitativa discreta	Número de personas que acompaña al pescador
Número de niños	child	Cuantitativa discreta	Número de niños que acompañan al pescador
Casa rodante	camper	Cualitativa	Indica si el pescador fue en casa rodante al parque (1) o no (0)

FUENTE: Elaboración propia

IV. RESULTADOS Y DISCUSIÓN

4.1 APLICACIÓN UNO: CONSUMO DE CIGARROS

4.1.1 ANÁLISIS EXPLORATORIO

4.1.1.1 ANÁLISIS EXPLORATORIO DE LA VARIABLES RESPUESTA

La variable respuesta de conteo, *número de cigarros semanales*, presenta exceso de ceros (83.45 por ciento) y posible sobredispersión pues presenta un promedio muestral de 0.7030 cigarros y varianza de 6.9266 cigarros^2; además de estos indicadores, si se simula una variable con distribución Poisson y parámetro (media) igual a 0.7030, se esperaría obtener 50 por ciento de ceros, sin embargo este porcentaje es excedido largamente en este conjunto de datos. Un gráfico de varas para esta variable se muestra a continuación:

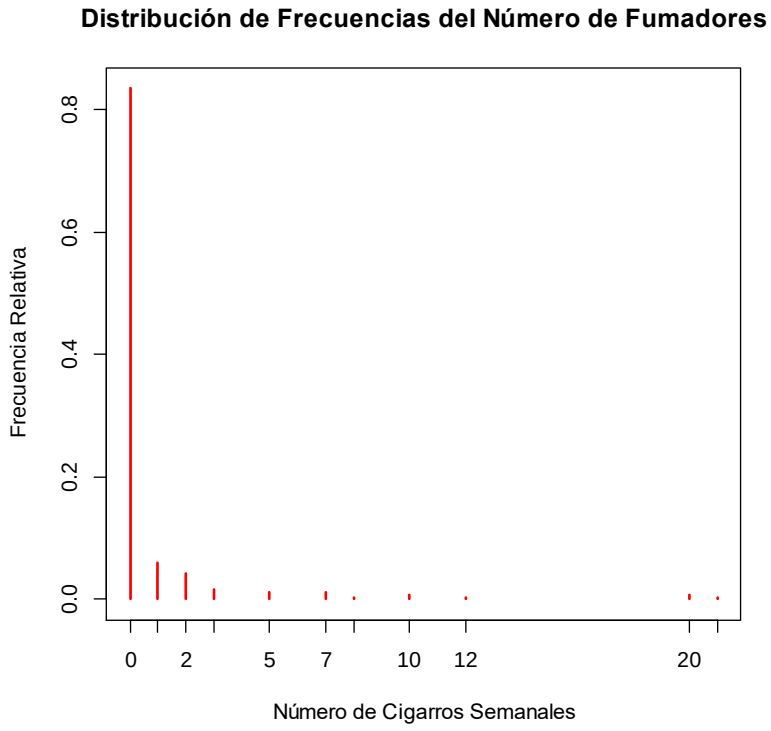

Figura 5. Ap1: Distribución de frecuencias del Número de Cigarros Semanales consumidos

Los modelos inflados en cero y *hurdle* no sólo estiman la tasa media de cigarros consumidos semanalmente, sino también el estatus del fumador. Los modelos inflados en cero diferencian los ceros estructurales de los ceros (más los conteos) aleatorios. Para esta aplicación, los ceros estructurales corresponderían a aquellos alumnos que nunca han fumado, mientras que los ceros aleatorios se encontrarían representados por aquellos que alguna vez fumaron pero que actualmente deciden no fumar y finalmente, los conteos aleatorios provendrían de aquellos que actualmente fuman. Sin embargo para este tipo de modelos se asume que la procedencia del cero (estructural o aleatoria) es desconocida por lo que debe ser estimada, entonces esta variable no será incluida en el modelamiento sino que será utilizada para contrastar la proporción observada de ceros estructurales versus la estimada. Por otro lado, los modelos *hurdle* sólo diferencian los ceros de los conteos positivos, entonces para esta aplicación, los ceros corresponden a aquellos alumnos que actualmente no fuman, y los conteos positivos provienen de quienes sí fuman.

La estimación de la prevalencia de vida de alumnos ingresantes fumadores indica que el 43.23 por ciento declara haber fumado alguna vez en su vida, mientras que la prevalencia actual de alumnos ingresantes fumadores es de 16.54 por ciento. Ambas prevalencias pueden observarse en el siguiente cuadro:

Cuadro N° 7: Ap1: Prevalencias de consumo de cigarros

		Fuma Cigarros Ahora		Total
		Sí	No	
Ha fumado alguna vez	Sí	44 (0.1654)	71 (0.2669)	115 (0.4323)
	No	0 (0.0000)	151 (0.5677)	151 (0.5677)
Total		44 (0.1654)	222 (0.8346)	266 (1.000)

FUENTE: Elaboración propia

4.1.1.2 ANÁLISIS EXPLORATORIO DE LAS VARIABLES PREDICTORAS

El 54.88 por ciento de los alumnos son de sexo masculino frente a una 45.12 por ciento de sexo femenino, por otro lado, la edad promedio es de 18.27 años y la edad mediana, así como la modal es 18. Respecto al departamento de nacimiento, el 74.06 por ciento de los alumnos es limeño mientras que el porcentaje restante nació en algún otro departamento. Todos los alumnos encuestados son solteros, sin embargo el 18.05 por ciento se encuentra en una relación. Luego, se tiene que las variables Distrito de Residencia y Carrera tienen demasiadas categorías (más de 10) con pocos casos por categoría (en algunos menos de 15), razón por la cual no se considerarán en el modelamiento.

Para la variable Edad a la que fumó por primera vez se tiene que el 34.21 por ciento comenzó a fumar entre los 10 y 16 años (denominada etapa escolar) y el 9.2 por ciento luego de los 16 años. El 56.77 por ciento restante nunca ha fumado. Respecto a la variable Dejar de fumar (si ha intentado dejar de fumar o no), se tiene que el 11.66 por ciento sí ha intentado dejar de fumar, el 31.57 por ciento no ha intentado dejar de fumar y el 56.77 por ciento restante nunca ha fumado. Ambas variables no pueden ser usadas en los modelos de regresión clásicos para datos de conteo (Poisson, NB2) ya que si el alumno **Nunca empezó a fumar** o **No ha intentado dejar de fumar porque nunca lo ha hecho**, entonces el número de cigarros semanales (variable respuesta) toma automáticamente el valor de cero, es decir no existiría aleatoriedad. Ocurre un escenario similar para los modelos inflados en cero ya que si el alumno **Nunca empezó a fumar** o **No ha intentado dejar de fumar porque nunca lo ha hecho,** entonces automáticamente se tratará de un cero estructural, de lo contrario, será un cero aleatorio, perdiendo de esa manera la aleatoriedad del componente binario.

Además, se estimó que la edad promedio, así como la mediana, de inicio en los hábitos de fumador es de 15.07 años. Luego, en el hogar de 25.67 por ciento de los alumnos, fuman regularmente, sin embargo ocurre lo contrario con su entorno de compañeros: el 52.63 por ciento contestó que su en su entorno compañeros fuman regularmente. Asimismo, el 14.66 por ciento de los alumnos consume bebidas alcohólicas regularmente; por otro lado el 57.77 por

ciento ha recibido alguna charla acerca de las consecuencias que conlleva fumar. Finalmente, se tienen otras variables como el consumo de puros, tabaco de mascar, tabaco en polvo y cigarro electrónico cuyas proporciones de consumo son de sólo 8.27, 1.13, 0.38 y 1.5 por ciento respectivamente, por lo que no serán consideradas en el modelamiento.

Uno de los supuestos para el desarrollo de modelos lineales generalizados es la independencia entre variables explicativas, por ello se muestra a continuación la correlación entre los factores y/o variables predictoras a utilizar en los modelos de regresión para datos de conteo.

Cuadro N° 8: Ap1: Matriz de correlaciones entre las variables predictoras

	Sexo	Depa Nac	Sit Sent	Edad Fumó	Dejar de Fumar	Fuman en casa	Fuman Comp.	Bebidas	Charlas	Edad
Sexo	1									
Depa Nac	0.028	1								
Sit Sent	0.081	0.088	1							
Edad Fumó	0.145	0.037	0.185	1						
Dejar de Fumar	0.135	0.028	0.18	0.442	1					
Fuman en Casa	0.007	0.062	0.022	0.028	0.106	1				
Fuman Compañeros	0.085	0.048	0.015	0.286	0.286	0.102	1			
Bebidas	0.17	0.034	0.124	0.233	0.233	0.126	0.227	1		
Charlas	0.04	0.098	0.015	0.063	0.043	0.068	0.121	0.029	1	
Edad	-0.11	-0.10	-0.012	-0.123	-0.09	0.01	-0.07	-0.04	0.124	1

FUENTE: Elaboración propia

El tipo de correlaciones empleadas es la siguiente:

	Coeficiente de contingencia (cuando está presente una variable nominal).
	Coeficiente biserial (Entre una variable nominal dicotómica y una cuantitativa).
	Coeficiente poliserial (Entre una variable ordinal y una cuantitativa).
	Coeficiente de Goodman y Kruskal (Entre dos variables ordinales).

Se observa que, a excepción de la correlación (coeficiente de Goodman y Kruskal) entre Dejar de Fumar y Edad a la que empezó a fumar, las correlaciones son menores a 0.30, pudiéndose asumir que la correlación es baja entre los factores y variable explicativas.

4.1.2 MODELO DE REGRESIÓN POISSON

a. Modelo estimado

El modelo de regresión Poisson se construye según la metodología planteada en la página 73, usando la función de enlace logarítmica. En el ANEXO 2 se presenta en detalle los modelos estimados paso a paso; el modelo estimado resultante es el siguiente:

$$\hat{\mu}_i = \exp(-4.57 + 2.331X_{1i} - 0.495X_{2i} + 1.98X_{3i} + 0.116X_{4i} - 11.768X_{5i} - 1.681X_{1i}X_{2i} + 0.6418X_{4i}X_{5i})$$

Para $i = 1, \ldots, 266$

Donde: $X_1 : Bebidas$ $X_2 : Charlas$ $X_3 : FumanCompañeros$

$X_4 : Edad$ $X_5 : FumanEnCasa$

Como se observa en el ANEXO 2, la prueba de Wald resulta significativa para todos los factores e interacciones, excepto para la covariable Edad, sin embargo ésta última se incluye en el modelo ya que el efecto de una interacción en la que está presente esta variable resulta significativo. La interpretación de los coeficientes estimados es como sigue (la obtención de las razones de tasas en el caso de interacciones se muestra en el ANEXO 2):

- La única variable que no interactúa es $X_3 : FumanCompañeros$, para ella se tiene que cuando fuman en el entorno de compañeros del ingresante, la tasa de consumo semanal de cigarros es 7.24 veces que cuando en su entorno no fuman, manteniendo las demás variables en valores constantes.
- Para las variables cuya interacción resulta significativa, la interpretación se realiza para cada combinación de niveles. Primero, para las variables $X_1 : Bebidas$ y $X_2 : Charlas$ se tiene que para los ingresantes que *no consumen bebidas alcohólicas regularmente*, la asistencia a charlas disminuye aproximadamente en un 40 por ciento la tasa de consumo semanal de cigarros, manteniendo las demás variables en valores constantes; luego para aquellos que *sí consumen bebidas alcohólicas de manera regular*, la asistencia a charlas disminuye aproximadamente en un 89 por ciento la tasa de consumo semanal de cigarros,

manteniendo las demás variables en valores constantes. Se observa que la asistencia a charlas tiene una mayor incidencia en aquellos ingresantes que consumen bebidas alcohólicas regularmente. Se puede realizar de la misma manera para comparar los estudiantes que asisten a charlas versus los que no asisten.

- Por otro lado, para las variables $X_4 : Edad$ y $X_5 : FumanEnCasa$ también se tiene que analizar según la interacción; primero cuando la edad del ingresante (cuyos familiares no fuman) se incrementa en un año, la tasa de consumo semanal de cigarros se incrementa en un 12.3 por ciento $\left(\exp(0.116) = 1.123\right)$, mientras que cuando la edad del ingresante (cuyos familiares fuman) se incrementa en un año, la tasa de consumo semanal de cigarros aproximadamente se duplica $\left(\exp(0.116 + 0.6148) = 2.134\right)$, en ambos casos, manteniendo las demás variables en valores constantes. Se observa entonces que la edad tiene una mayor incidencia en el consumo semanal de cigarros cuando los familiares del ingresante fuman.

b. Ajuste del modelo Poisson

Cuadro N° 9: Ap1: Indicadores de ajuste del modelo Poisson

Criterio	Interpretación
Deviance	Al poner a prueba la bondad de ajuste mediante el test de Deviance, se rechaza la hipótesis nula de que el modelo se ajusta a los datos, al obtener $Deviance = 422.64 \sim \chi^2_{(258)}$ con pvalor de 4.24×10^{-10}.
Estadístico Chi Cuadrado de Pearson	Respecto al estadístico Chi cuadrado de Pearson, el cociente de éste entre sus grados de libertad es 3.22, indicando que el modelo no está correctamente especificado (la función de varianza no es adecuada: la distribución probabilística elegida no es la correcta).
Pseudo R^2 basado en la verosimilitud	El Pseudo R^2 basado en la verosimilitud arroja un valor de 0.5678.

… continuación

AIC	Toma el valor de 564.3847. Como referencia, se tiene que el AIC del modelo de sólo intercepto es de 998.96
Residuales de Pearson	Analizando los residuales bajo este modelo Poisson se tienen 26 residuales de Pearson que sobrepasan el rango de las 2 desviaciones estándar, y 30 *leverages* que sobrepasan el límite $2p/n = 2\times 8/266 = 0.06$. Se presentan a continuación las gráficas de residuales y *leverages*.

FUENTE: Elaboración propia

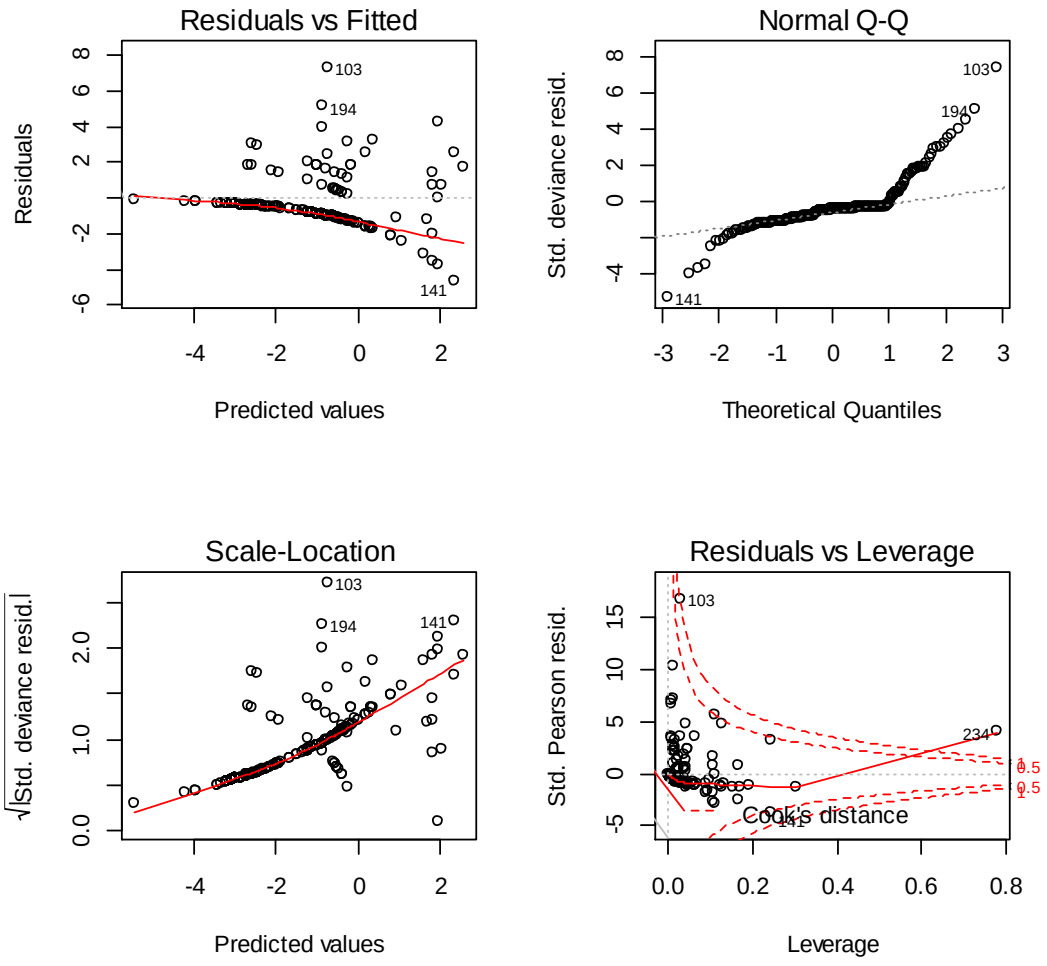

Figura 6. Ap1: Gráfico de residuales y *leverages* del modelo Poisson

Se observa en la Figura 6 de la página anterior que en el gráfico *Residuals vs Fitted*, los *outliers* detectados para el modelo ocupan las posiciones 103, 141 y 194. Se observa un patrón indicando que no hay homogeneidad de varianzas, lo cual es aceptable ya que por definición la varianza no es constante sino que depende de la media. Luego, en el segundo gráfico (*Normal Q-Q*), no se espera obtener un gráfico que muestre normalidad de los residuales, y efectivamente sucede tal cual. Después, en el gráfico *Scale Location* también se presenta un patrón que no indica homogeneidad de varianzas. Finalmente, en el gráfico de *Residuales estandarizados de Pearson versus leverages* (*hat*), los valores extremos horizontalmente indican altos *leverages*. Residuales estandarizados de Pearson mayores a |2| y con *leverages* altos, en este caso el punto 234, indica un mal ajuste en el modelo. Si se amplía esta última gráfica, se tiene la Figura 7 como se muestra a continuación:

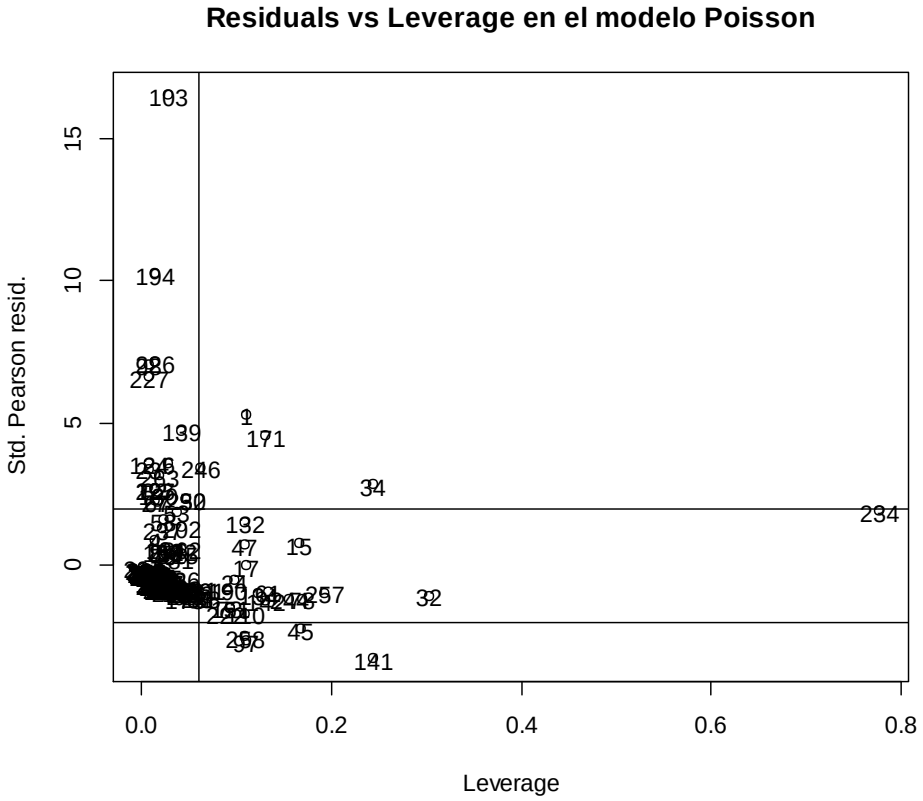

Figura 7. Ap1: Gráfico de *leverages* versus residuales del modelo Poisson

Los *leverages*, aquellos cuyos valores en la diagonal de la matriz *hat* se encuentran por encima de 2p/n, son 30; gráficamente, los valores a la derecha de la línea vertical indican los *leverages* de este modelo, sin embargo el punto 234 resalta ante los demás. Los puntos por encima de la línea horizontal son *outliers*. Las observaciones 34, 1 y 171 son *leverages* y *outliers* moderados.

Según los resultados de mal ajuste obtenidos para el modelo Poisson, es probable que exista sobredispersión para la variable respuesta en este modelo, lo cual debe ser confirmado mediante los tests propuestos de Bohning y Dean & Lawless.

c. Pruebas de sobredispersión

- **Test de Bohning**

Se prueba la hipótesis nula $H_0 : \mu_Y = \sigma_Y^2$, la cual se rechaza ya que se obtiene un estadístico de prueba de 26.52687 y pvalor equivalente a cero, entonces se puede afirmar que la media de cigarros semanales consumidos difiere de su varianza.

- **Test de Cameron y Trivedi (1985)**

La hipótesis a probar es $H_0 : E[Y_i] = V[Y_i] = \mu_i$. El p-valor obtenido para el estadístico de prueba $Z_{calc} = 2.1783$ es de 0.01469, con un estimado de dispersión de 3.2, similar al del estadístico chi cuadrado. Por lo tanto se rechaza la hipótesis nula; existe sobredispersión

- **Regresión de la varianza estimada sobre la media estimada**

Se contrastan las siguientes hipótesis según el test propuesto por Cameron y Trivedi (1986): $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 \neq 1$, obteniendo $t_{calc} = 10.46147 \sim t_{265}$ cuyo $pvalor \approx 0$, entonces existe suficiente evidencia estadística para rechazar la hipótesis nula, dicho de otro modo, no existe equidispersión.

Luego, según el test propuesto por Cameron y Trivedi (1990), las hipótesis a contrastar son $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, obteniendo $t_{calc} = 9.96 \sim t_{252}$ cuyo $pvalor \approx 0$, entonces existe suficiente evidencia estadística para rechazar la hipótesis nula, es decir, no existe equidispersión

- **Prueba de Dean y Lawless**

El estadístico T calculado es igual a 18.8849, dado que su distribución asintótica es normal, el $pvalor = P(Z > 18.8849) \approx 0$, entonces se rechaza la hipótesis nula de equidispersión de la variable respuesta en el modelo Poisson.

Por lo visto según el estadístico Chi Cuadrado de Pearson, el test de Bohning, los *tests* de Cameron y Trivedi y el de Dean y Lawless, está presente el problema de sobredispersión en la variable respuesta del modelo Poisson, por ello la metodología indica que se deben modelar los datos utilizando la regresión NB2, inflada en cero o hurdle.

4.1.3 MODELO DE REGRESIÓN BINOMIAL NEGATIVO (NB2)

a. Modelo estimado

Se mostró que el modelo anterior (Poisson) no era adecuado debido al problema de sobredispersión. Debido a ello, los errores estándar de los coeficientes de regresión se encontraban subestimados y entonces, la lista de factores, covariables e interacciones cuya prueba de Wald resultó significativa era extensa (cuatro factores, una covariable, dos interacciones). En contraparte a ello, mediante el modelo de regresión NB2, se corrige la función de varianza, la cual pasa de ser μ a $\mu+\alpha\mu^2$. Luego, para estimar el número de cigarros consumidos semanalmente, la ecuación de regresión es:

$$\hat{\mu}_i = \exp\left(-8.2060 + 1.6443X_{1i} - 1.0708X_{2i} + 2.1794X_{3i} + 0.3278X_{4i}\right); \quad \hat{\alpha} = 0.1903^{-1}$$

Para $i = 1,\ldots,266$, donde:

$X_1 : Bebidas$ $X_2 : Charlas$ $X_3 : FumanCompañeros$ $X_4 : Edad$

El procedimiento de selección de variables y algunas salidas de R se encuentran en el ANEXO 3. Ahora, interpretando los coeficientes estimados:

- La tasa semanal de consumo de cigarros de un alumno que consume bebidas alcohólicas regularmente es cinco veces $\left(\exp(1.6443) = 5.18\right)$ la de un alumno que no las consume de manera regular, manteniendo las demás variables en valores constantes.
- La tasa de consumo semanal de cigarros disminuye aproximadamente en 66 por ciento $\left(\exp(-1.0708) = 0.34\right)$ cuando el alumno asiste a charlas, manteniendo las demás variables en valores constantes.
- Cuando fuman en el entorno de compañeros del ingresante, la tasa de consumo semanal de cigarros es aproximadamente nueve veces $\left(\exp(2.1794) = 8.84\right)$ la de una alumno cuyo entorno no es fumador, manteniendo las demás variables en valores constantes.

- Cuando la edad del ingresante se incrementa en un año, la tasa de consumo semanal de cigarros se incrementa en un 38.8 por ciento $\left(\exp(0.3278)=1.388\right)$, manteniendo las demás variables en valores constantes.
- El estimado $\hat{\alpha}=0.1903^{-1}$ sirve para cuantificar la relación entre la media y varianza en el modelo NB2, entonces se tiene que la función de variancia estimada para este modelo es $\sigma^2=\mu+6.255\mu^2$.

b. Ajuste del modelo NB2

Cuadro N° 10: Ap1: Indicadores de ajuste del modelo NB2

Criterio	Interpretación
Deviance	Se obtiene $Deviance=111.45\sim\chi^2_{(261)}$, para el cual el pvalor resultante es 1. Entonces no se puede afirmar que el modelo no se ajuste a los datos.
Estadístico Chi Cuadrado de Pearson	Toma el valor de 0.988, podría indicarse que el modelo de regresión binomial negativo está especificado correctamente (la distribución empleada es adecuada para el modelamiento), ya que el cociente obtenido es cercano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 calculado para este caso es de 0.1194
AIC	El valor AIC: del modelo es de 393.35, mientras que para el modelo nulo o minimal fue de 437.23.
Residuales de Pearson	Analizando los residuales bajo este modelo Poisson se tienen 13 residuales de Pearson que sobrepasan el rango de las 2 desviaciones estándar, y 25 *leverages* que sobrepasan el límite $2p/n=2\times5/266=0.0375$.

FUENTE: Elaboración propia

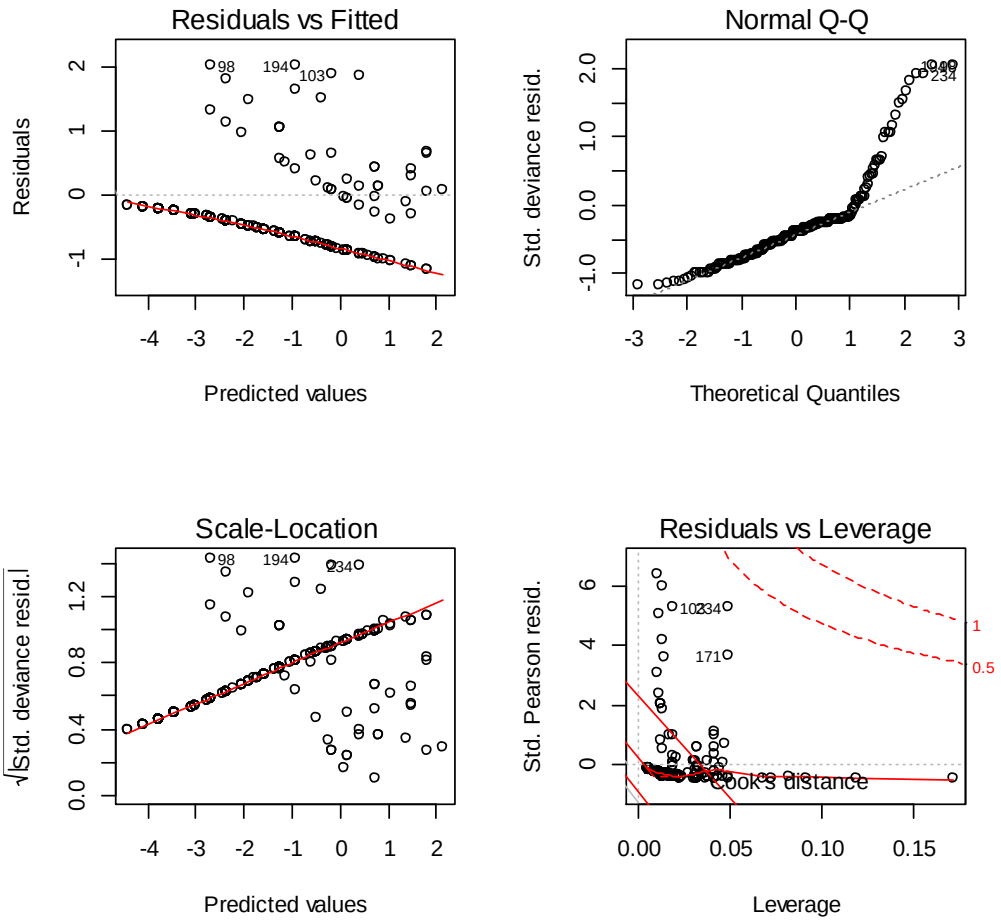

Figura 8. Ap1: Gráfico de residuales y *leverages* del modelo NB2

En el gráfico *Residuals vs Fitted*, los *outliers* detectados para el modelo ocupan las posiciones 98, 103 y 194, aunque están en el límite de ± 2. Asimismo, se observa un patrón indicando que no hay homogeneidad de varianzas, lo cual es aceptable ya que por definición la varianza no es constante. Luego, en el gráfico *Normal Q-Q* no se espera obtener un gráfico que muestre normalidad de los residuales, y efectivamente sucede tal cual. Se observan varios *outliers*, en el límite de las dos desviaciones estándar. En el tercer gráfico, *Scale Location*, también se presenta un patrón que no indica homogeneidad de varianzas. Finalmente, en la gráfica de *residuales estandarizados de Pearson versus leverages* (*hat*), se observan los

puntos 103, 171 y 234 con residuales altos pero no son *leverages*, sin embargo en esta última gráfica no se aprecian muy bien los puntos por lo que se mostrará ampliada a continuación:

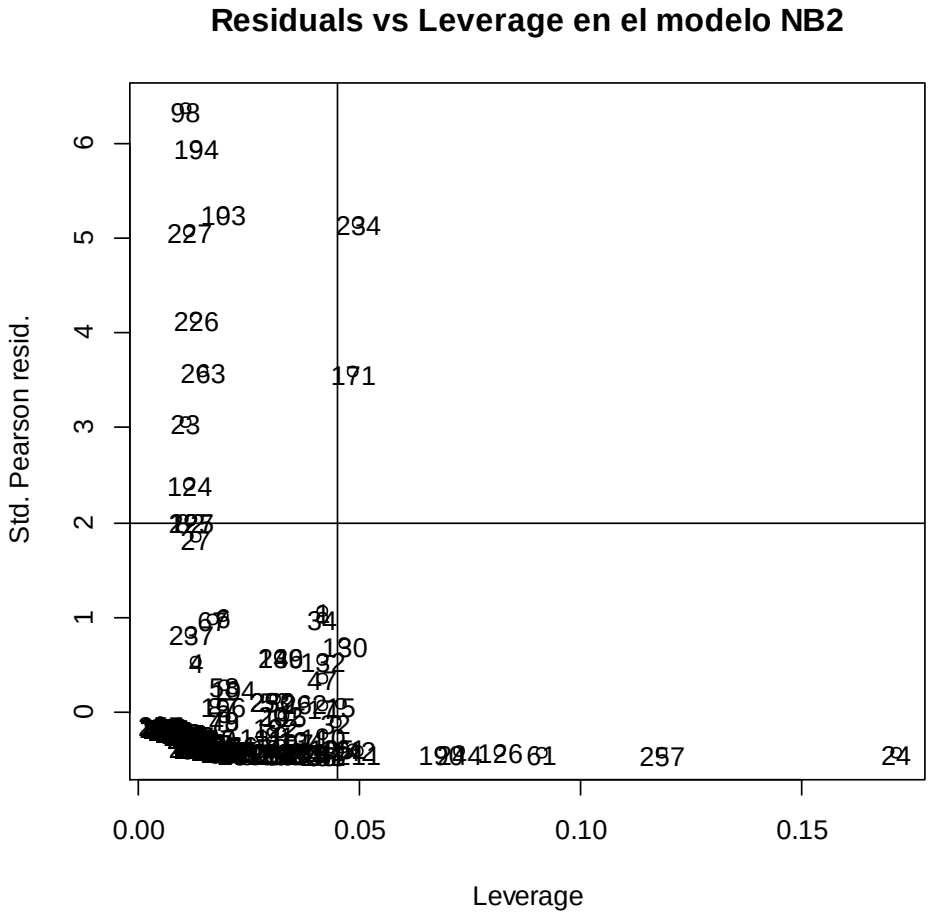

Figura 9. Ap1: Gráfico de *leverages* versus residuales del modelo NB2

Los puntos por encima de la línea horizontal indican residuales de Pearson altos (*outliers*), mientras que los puntos a la derecha de la línea vertical son *leverages*. El modelo NB2 presenta menos *outliers* y *leverages* que el modelo Poisson.

4.1.4 MODELO DE REGRESIÓN POISSON INFLADO EN CERO

a. Modelo estimado

Ecuación de regresión estimada para la media en los conteos aleatorios

$$\hat{\mu}_i = \exp\left(-6.07 + 1.48X_{1i} + 0.059X_{2i} + 1.478X_{3i} + 0.286X_{4i} + 0.815X_{5i} - 2.24X_{1i}X_{2i}\right)$$

Para $i = 1,\ldots,266$, donde: $X_1 : Bebidas$ $X_2 : Charlas$

$X_3 : FumanCompañeros$ $X_4 : Edad$ $X_5 : FumanEnCasa$

La selección de variables, algunas salidas de R y el cálculo de las razones de tasa para la interacción se detallan en el ANEXO 4. Se presenta a continuación la interpretación de los coeficientes de regresión estimados:

- La tasa promedio de consumo semanal de cigarros para un ingresante cuyos compañeros fuman en su entorno es aproximadamente el cuádruple $\left(\exp(1.478) = 4.384\right)$ que para aquellos cuyos compañeros no fuman, manteniendo en valores fijos las demás variables.
- A medida que la edad de un ingresante se incrementa en un año, su tasa de consumo semanal de cigarros se incrementa en 33 por ciento $\left(\exp(0.286) = 1.331\right)$, manteniendo las demás variables en valores fijos.
- La tasa promedio de consumo semanal de cigarros para un ingresante cuyos familiares fuman en casa es aproximadamente el doble $\left(\exp(0.85) = 2.25\right)$ que para aquellos cuyos familiares no fuman en casa, manteniendo las demás variables en valores fijos.
- Como existe interacción entre los factores $X_1 : Bebidas$ y $X_2 : Charlas$, la interpretación se realiza para cada combinación de niveles. Primero, se tiene que para aquellos que *sí consumen bebidas alcohólicas regularmente*, la tasa semanal de consumo de cigarros cuando reciben charlas cerca de los daños del cigarro a la salud es aproximadamente 89 por ciento menor que cuando no las reciben, mientras que para los ingresantes que *no consumen*

bebidas alcohólicas regularmente, esta misma tasa para cuando reciben charlas acerca de los daños del cigarro a la salud es 6.5 por ciento mayor que la tasa de aquellos que no han recibido dichas charlas. Se observa que en aquellos ingresantes que consumen bebidas alcohólicas regularmente, la no asistencia a charlas sobre el consumo dañino del cigarro incrementa la tasa de consumo semanal de cigarros a diferencia de lo que sucede en aquellos que no consumen bebidas de manera regular.

- Para aquellos que *han asistido a charlas sobre el consumo de cigarros*, la tasa semanal de consumo de cigarros cuando consumen bebidas alcohólicas regularmente es 90 por ciento menor cuando no lo hacen, manteniendo las demás variables en valores fijos. Por otro lado, para aquellos que *no han asistido a charlas sobre el consumo de cigarros*, la tasa de consumo semanal de cigarros cuando consumen bebidas alcohólicas de manera regular es aproximadamente el cuádruple que cuando no lo hacen, manteniendo en valores fijos las demás variables. Se aprecia que cuando el ingresante no asiste a charlas sobre consumo de cigarros, el consumo regular de bebidas alcohólicas es un factor de riesgo en el consumo de cigarros, no siendo así en el grupo de ingresantes que asiste a charlas.

Ecuación de regresión estimada para la proporción de ceros estructurales

$$\text{logit}\left(\hat{\pi}_i\right) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.53 - 1.95X_{1i}$$

Para $i = 1,\ldots,266$, donde $X_1 : \textit{Bebidas}$

La chance de que un ingresante no haya fumado nunca disminuye en 86 por ciento $\left(\exp(-1.95) = 0.14\right)$ cuando éste consume bebidas alcohólicas regularmente.

b. Ajuste del modelo Poisson inflado en cero

Cuadro N° 11: Ap1: Indicadores de ajuste del modelo Poisson Inflado en Cero

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 321.1303, de modo que $P\left(\chi^2_{(257)} > 321.1303\right) = 0.004$. Además el cociente de este estadístico entre sus grados de libertad es 1.2495, un valor no tan cercano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.3364
AIC	El AIC del modelo en estudio es 390.5283, mientras que el del modelo de sólo intercepto es 565.3325.
Residuales de Pearson	Se observan 10 residuales de Pearson por encima de las dos desviaciones estándar. Ningún residual por debajo de las dos desviaciones estándar.

FUENTE: Elaboración propia

En la Figura 10 de la siguiente página se tiene que las observaciones 98, 103, 23, 194, 227, 139, 226 y 263 destacan por tener altos residuales de Pearson. Las observaciones están en el límite 1 y 6.

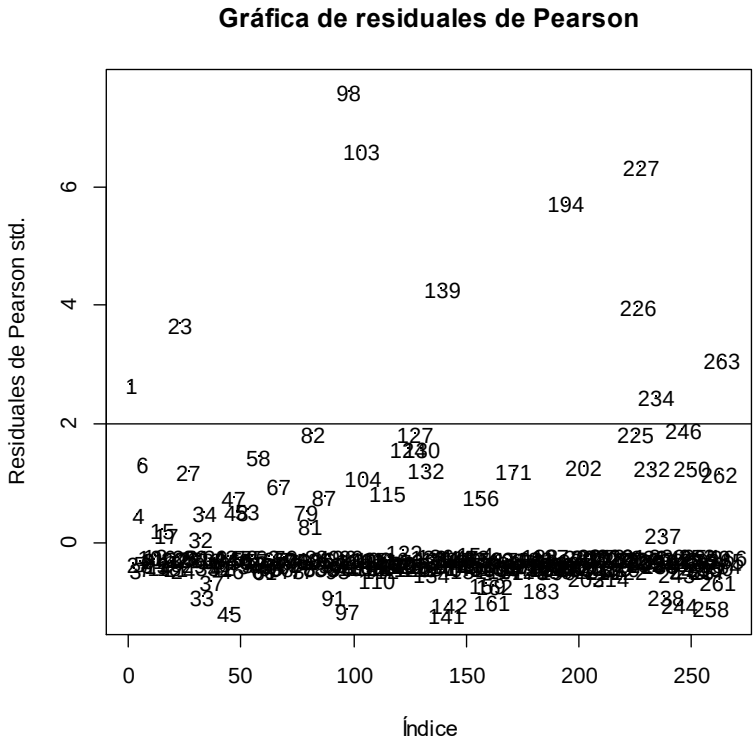

Figura 10. Ap1: Gráfico de residuales de Pearson del modelo Poisson inflado en cero

Analizando las probabilidades de que una observación provenga de un cero estructural, considerando que si ésta es mayor a 0.5, se estima como cero estructural se tiene:

Cuadro N° 12: Ap1: Comparación de valores observados versus predichos (Modelo Poisson Inflado en Cero)

		Observación		
	Status	No ha Fumado nunca	Alguna vez ha Fumado	TOTAL
Predicción	Ceros estructurales	**140**	87	227
	Ceros + Conteos Aleatorios	11	**28**	39
TOTAL		151	115	**266**

FUENTE: Elaboración propia

Entonces, respecto a su origen estructural o aleatorio, se han clasificado correctamente el 63.15 por ciento de los casos. El modelo Poisson inflado en cero predice que el 76.5 por ciento de las observaciones son ceros estructurales, y que el 72.18 por ciento son ceros (tanto estructurales como aleatorios). De los datos observados se tiene que este porcentaje es de 83.46 por ciento, es decir en este caso la estimación es algo cercana.

En la siguiente página se muestran los resultados para las observaciones 15 – 30: en la primera columna aparece el número de cigarros semanales observado, en la siguiente columna la probabilidad que la observación provenga de un cero estructural y a partir de la tercera columna la probabilidad de que la variable respuesta tome el valor cero, uno, dos, etc. Si bien el rango de la distribución Poisson se extiende hasta infinito, R sólo muestra los valores hasta 21, ya que es el máximo obtenido en la muestra (sin embargo por cuestión de espacio sólo se muestra en esta hoja hasta $y = 17$). Por ejemplo, para la observación 19, su respuesta fue que consume cero cigarros semanalmente. Según el modelo, se obtiene una predicción de cero cigarros semanales con 0.895 de probabilidad, sin embargo, de esta probabilidad se puede decir que se tiene 0.82 de probabilidad de ser cero estructural (nunca ha fumado) versus 0.075 de que sea aleatorio (alguna vez ha fumado pero no lo hace actualmente). Luego, la probabilidad de que consuma un cigarro semanalmente es de 0.065, de que consuma dos cigarros semanales es de 0.029 y así sucesivamente.

Cuadro N° 13: Ap1: Predicción en el modelo Poisson inflado en cero

Id	Y	Cero estructural	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
15	10	0.40	0.396	0.000	0.000	0.000	0.001	0.003	0.007	0.013	0.023	0.034	0.046	0.056	0.063	0.066	0.063	0.057	0.048	0.038
16	0	0.82	0.952	0.041	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	7	0.40	0.396	0.000	0.001	0.004	0.010	0.021	0.036	0.052	0.066	0.074	0.075	0.070	0.059	0.046	0.033	0.023	0.014	0.009
18	0	0.82	0.851	0.052	0.048	0.029	0.013	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0	0.82	0.895	0.065	0.029	0.009	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0	0.82	0.939	0.049	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
21	0	0.82	0.829	0.023	0.037	0.039	0.032	0.021	0.011	0.005	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
22	0	0.82	0.939	0.049	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
23	1	0.82	0.954	0.039	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
24	0	0.82	0.822	0.001	0.004	0.010	0.017	0.024	0.027	0.026	0.023	0.017	0.012	0.007	0.004	0.002	0.001	0.000	0.000	0.000
25	0	0.82	0.952	0.041	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
26	0	0.82	0.939	0.049	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
27	1	0.82	0.877	0.064	0.038	0.015	0.004	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
28	0	0.82	0.939	0.049	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
29	0	0.82	0.871	0.063	0.041	0.018	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
30	0	0.82	0.952	0.041	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

FUENTE: Elaboración propia

4.1.5 MODELOS DE REGRESIÓN BINOMIAL NEGATIVO INFLADO EN CERO

a. Modelo estimado

Ecuación de regresión estimada para la media en conteos aleatorios

$$\hat{\mu}_i = \exp\left(-8.2605 + 1.6422X_{1i} - 1.0708X_{2i} + 2.1794X_{3i} + 0.3278X_{4i}\right); \quad \hat{\alpha} = \left(\log\left(-1.659\right)\right)^{-1}$$

Para $i = 1,\ldots,266$, donde:

$X_1 : Bebidas$ $X_2 : Charlas$ $X_3 : FumanCompañeros$ $X_4 : Edad$

Al ser los coeficientes estimados los mismos que en el modelo NB2, las interpretaciones son las mismas.

Ecuación de regresión estimada para la proporción de ceros estructurales

$$\text{logit}\left(\hat{\pi}_i\right) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -8.291$$

Ninguna variable predictora contribuye a explicar la proporción de ceros estructurales y dicha proporción es estimada como $\hat{\pi}_i = 0.00025$, es decir este modelo llega a la conclusión de que casi no existen ceros estructurales, pero según los resultados descriptivos (un cero estructural proviene de un ingresante que nunca fumó) esto no es así, ya que el 56.77 por ciento de los ingresantes nunca ha fumado.

b. Ajuste del modelo NB2 inflado en cero

Cuadro N° 14: Ap1: Indicadores de ajuste del modelo NB2 inflado en cero

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 257.8707, de modo que $P\left(\chi^2_{(259)} > 257.707\right) = 0.5082$. Además el cociente de este estadístico entre sus grados de libertad es 0.9956, un valor bastante cercano a uno, , aunque no tanto como el del modelo NB2.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 es igual a 0.1195753
AIC	El AIC del modelo ajustado es igual a 395.3509 y el del modelo nulo (de sólo intercepto) es 439.2339, la diferencia de 44 según Hilbe (ver Cuadro N° 2) puede ser considerado un indicador de que el modelo ajustado es mejor que el minimal.
Residuales de Pearson	Se tienen 13 residuales por encima de las 2 desviaciones estándar y ninguno por debajo de las 2 desviaciones estándar.

FUENTE: Elaboración propia

Cuadro N° 15: Ap1: Comparación de valores observados versus predichos (Modelo binomial negativo inflado en cero)

		Observación		
	Status	No ha Fumado nunca	Alguna vez ha Fumado	TOTAL
Predicción	Ceros estructurales	**0**	0	0
	Ceros + Conteos Aleatorios	151	**115**	266
TOTAL		151	115	**266**

FUENTE: Elaboración propia

La tasa de clasificación correcta es de 43.23 por ciento. Luego, el modelo NB2 inflado en cero predice que el 70.5 por ciento de las observaciones son ceros estructurales, y que el 83.1 por ciento son ceros, tanto estructurales como aleatorios. De los datos observados se tiene que este porcentaje es de 83.46 por ciento, es decir esta estimación es bastante cercana.

Como ninguna variable predictora contribuye a explicar la proporción de ceros estructurales, entonces la generación de ceros no es explicada por dos mecanismos distintos: uno para ceros estructurales y los conteos aleatorios por el modelo binomial negativo. A primera vista se preferiría el modelo binomial negativo sin inflación en cero, sin embargo el tema de comparación de modelos se verá más adelante.

4.1.6 MODELOS DE REGRESIÓN HURDLE POISSON

a. Modelo estimado

Se muestra a continuación el modelo estimado, se empleó la función de enlace logarítmica para estimar la media de los datos positivos (Poisson truncado en cero) y se muestran los enlaces *logit* y logarítmico para estimar la proporción de ceros.

Ecuación de regresión estimada para los datos positivos

$$\hat{\mu}_i = \exp\left(-5.5899 + 1.55X_{1i} + 0.256X_{2i} + 1.22X_{3i} + 0.27X_{4i} + 0.808X_{5i} - 2.5X_{1i}\,\mathrm{X}_{2i}\right)$$

Para $i = 1,\ldots,266$, donde:

$X_1 : Bebidas$ $X_2 : Charlas$ $X_3 : FumanCompañeros$ $X_4 : Edad$

$X_5 : FumanEnCasa$ $X_6 : Depa.Nacimiento$

El proceso de selección de variables, algunas salidas en R y el cálculo de la razón de tasa para la interacción se muestran en el ANEXO 6. La interpretación de los coeficientes estimados es como sigue:

- La tasa de consumo semanal de cigarros de un ingresante se incrementa en 31 por ciento $\left(\exp(0.27) = 1.31\right)$ al cumplir éste un año más de vida, manteniendo constantes los valores para las demás variables.
- La tasa de consumo semanal de cigarros de un ingresante cuyo entorno de compañeros suele fumar es aproximadamente el triple $\left(\exp(1.22) = 3.39\right)$ que la de un ingresante cuyos compañeros no suelen realizarlo (fumar).
- La tasa de consumo semanal de cigarros de un ingresante cuyos familiares suelen fumar es aproximadamente el doble $\left(\exp(0.808) = 2.24\right)$ que la de un ingresante cuyos familiares no acostumbran fumar.

- Las variables Bebidas y Charlas interactúan de manera significativa, por lo que debe interpretarse para cada combinación de niveles de estos factores en estudio, para los ingresantes que *consumen bebidas alcohólicas regularmente*, la tasa semanal de consumo de cigarros de aquellos que han recibido charlas acerca de los daños del cigarro a la salud es 90 por ciento menor que la tasa de aquellos que no han recibido dichas charlas, mientras que para los ingresantes que *no consumen bebidas alcohólicas regularmente* la tasa semanal de consumo de cigarros de aquellos que han recibido charlas acerca de los daños del cigarro a la salud es 67 por ciento mayor que la tasa de aquellos que no han recibido dichas charlas. Se observa que en aquellos ingresantes que consumen bebidas alcohólicas regularmente, la no asistencia a charlas sobre el consumo dañino del cigarro incrementa la tasa de consumo semanal de cigarros (factor de riesgo), mientras que sucede lo contrario para los que no consumen bebidas alcohólicas de manera regular: aquellos que han asistido a charlas tienen una tasa de consumo semanal de cigarros mayor (la asistencia a charlas se convierte en un factor de riesgo).

Ecuación de regresión estimada para la proporción de ceros

Usando enlace *logit*: $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -3.03 + 1.7048X_{1i} + 1.5089X_{2i}$

La interpretación de los coeficientes estimados se presenta a continuación:

- La chance de que un ingresante fume cuando consume bebidas alcohólicas regularmente es de 5.5 veces $\left(\exp(1.7048) = 5.5\right)$ la de un ingresante que no las consume regularmente.
- La chance de que un ingresante fume cuando su entorno de compañeros suele fumar es 4.5 veces $\left(\exp(1.5089) = 4.52\right)$ la de un ingresante cuyos compañeros no fuman de manera regular.

b. Bondad de ajuste

Cuadro N° 16: Ap1: Indicadores de ajuste del modelo *hurdle logit* Poisson

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 306.3178, de modo que $P\left(\chi^2_{(256)} > 306.3178\right) = 0.017$. Luego, el cociente de este estadístico entre sus grados de libertad es 1.196554, un valor algo cercano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.3427
AIC	El AIC del modelo en estudio es 388.9188, mientras que el del modelo de sólo intercepto es 565.3325.
Residuales de Pearson	Se observan 13 residuales de Pearson por encima de las dos desviaciones estándar. Ningún residual por debajo de las 2 desviaciones estándar.

FUENTE: Elaboración propia

En la Figura 11 se observa las observaciones con residuales de Pearson altos: 98, 103, 227, 194, 139, 23, 226, 1.

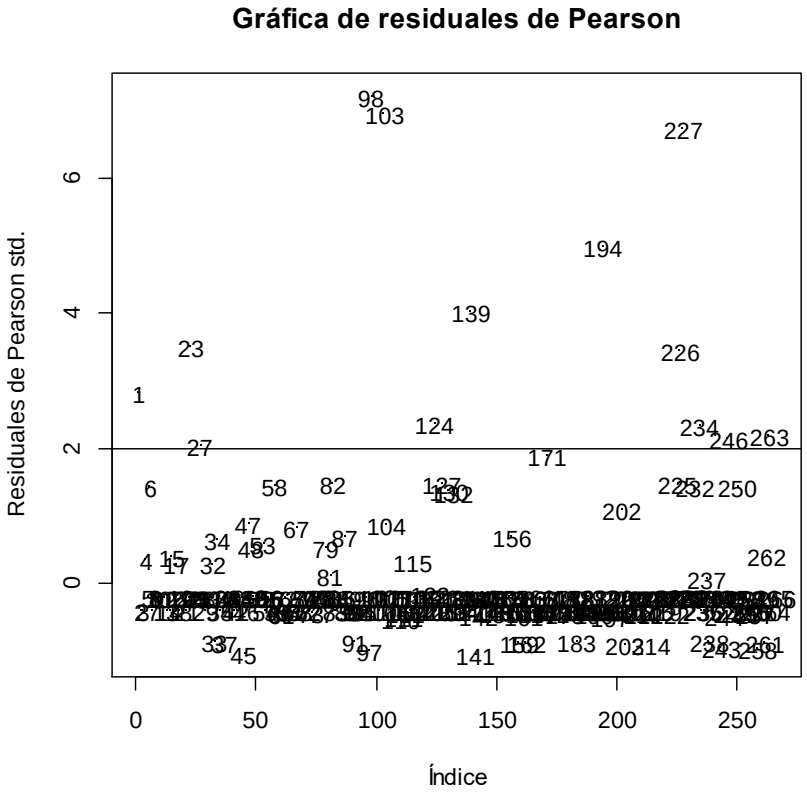

Figura 11. Ap1: Gráfico de residuales de Pearson del modelo *Hurdle logit* Poisson

Adicionalmente, se tiene que para el modelo hurdle *logit* Poisson, la tasa de clasificación correcta es de 70.3 por ciento.

Cuadro N° 17: Ap1: Comparación de valores observados versus predichos (modelo *hurdle logit* Poisson)

		Observación		
	Status	No Fuma	Fuma	TOTAL
Predicción	Ceros	**159**	16	175
	Positivos	63	**28**	91
TOTAL		222	44	**266**

FUENTE: Elaboración propia

Cuadro N° 18: Ap1: Predicción en el modelo *hurdle logit* Poisson

Id	Y	Ratio - 0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
15	10	0.55	0.452	0	0	0	0.001	0.004	0.008	0.015	0.024	0.035	0.046	0.055	0.06	0.06	0.056	0.049	0.04	0.031
16	0	0.12	0.954	0.036	0.009	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	7	0.55	0.452	0	0.001	0.004	0.01	0.021	0.035	0.049	0.062	0.069	0.069	0.062	0.052	0.04	0.028	0.019	0.012	0.007
18	0	0.21	0.82	0.053	0.056	0.039	0.02	0.008	0.003	0.001	0	0	0	0	0	0	0	0	0	0
19	0	0.07	0.954	0.026	0.014	0.005	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0.1	0.954	0.034	0.01	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0.19	0.82	0.019	0.033	0.04	0.035	0.025	0.015	0.008	0.003	0.001	0	0	0	0	0	0	0	0
22	0	0.1	0.954	0.034	0.01	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	1	0.15	0.954	0.038	0.007	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0.05	0.954	0	0.001	0.002	0.004	0.006	0.007	0.007	0.006	0.005	0.003	0.002	0.001	0.001	0	0	0	0
25	0	0.12	0.954	0.036	0.009	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0.1	0.954	0.034	0.01	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	1	0.06	0.954	0.021	0.015	0.007	0.002	0.001	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0.1	0.954	0.034	0.01	0.002	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0.25	0.82	0.091	0.056	0.023	0.007	0.002	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0.12	0.954	0.036	0.009	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FUENTE: Elaboración propia

La primera columna muestra el valor observado del número de cigarros semanales, en la siguiente columna se muestra el ratio $\frac{1-f_{cero}(0,\mathbf{z},\boldsymbol{\gamma})}{1-f_{conteo}(0,\mathbf{x},\boldsymbol{\beta})}$, finalmente las columnas restantes muestran los valores para $P(Y_i = y_i)$, donde $\pi_i = P(Y_i > 0)$ entonces $1-\pi_i = P(Y_i = 0)$, y los valores de $P(Y_i = y_i)$ para $y_i > 0$ se obtienen haciendo uso de la distribución truncada en cero.

4.1.7 MODELOS DE REGRESIÓN *HURDLE* BINOMIAL NEGATIVO

a. Modelo estimado

El proceso de selección de variables, algunas salidas en R y el cálculo de la razón de tasa para la interacción se muestran en el ANEXO 7. Se muestra a continuación el modelo estimado, se empleó la función de enlace logarítmica para estimar la media de los datos positivos (binomial negativo truncado en cero) y se muestran el enlace *logit* para estimar la proporción de ceros.

Ecuación de regresión estimada para los datos positivos

$$\hat{\mu}_i = \exp(-6.294), \text{ con } \hat{\alpha} = \frac{1}{\exp(-8.626)} = 5574.735, \text{ para } i = 1,\ldots,266.$$

La tasa semanal de consumo de cigarros es de 0.001 cigarros para todos los ingresantes. Ninguna variable predictora ingresó en esta ecuación estimada.

Ecuación de regresión estimada para la proporción de ceros

$$\log it(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.0229 + 1.705X_{1i} + 1.509X_{3i}$$

Para $i = 1,\ldots,266$, donde:

$X_1 : Bebidas$ $X_3 : FumanCompañeros$

- La chance de que un ingresante fume cuando consume bebidas alcohólicas regularmente es de 5.5 veces $(\exp(1.7048) = 5.5)$ la de un ingresante que no las consume regularmente.

- La chance de que un ingresante fume cuando su entorno de compañeros suele fumar es 4.5 veces $\left(\exp(1.5089) = 4.52\right)$ la de un ingresante cuyos compañeros no fuman de manera regular.

b. Bondad de ajuste

Cuadro N° 19: Ap1: Indicadores de ajuste del modelo hurdle NB2

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 133.0347, de modo que $P\left(\chi^2_{(261)} > 133.0347\right) \approx 1$, sin embargo, el cociente de este estadístico entre sus grados de libertad es 0.5097, un valor lejano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.0924.
AIC	El AIC del modelo en estudio es 402.2991, mientras que el del modelo de sólo intercepto es 439.0045.
Residuales de Pearson	Se observan sólo 5 residuales de Pearson por encima de las dos desviaciones estándar. Ningún residual por debajo de las 2 desviaciones estándar.

FUENTE: Elaboración propia

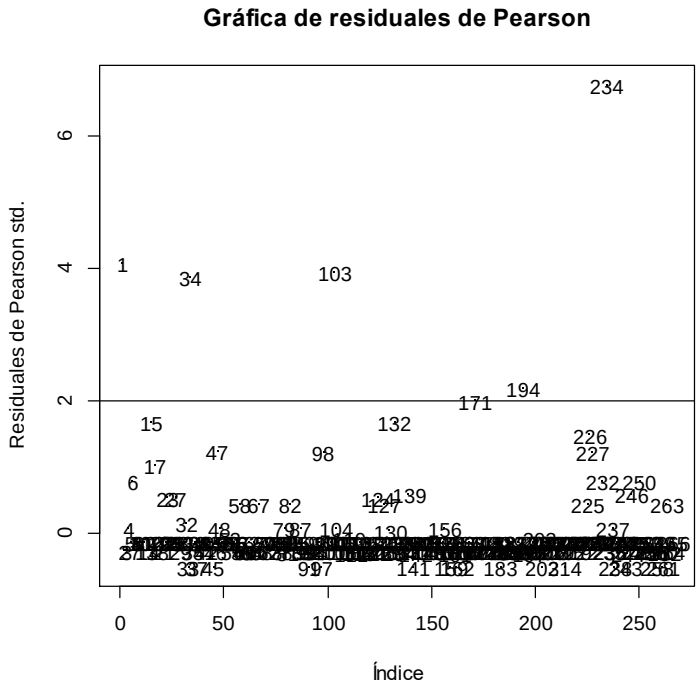

Figura 12 Ap1: Gráfica de residuales de Pearson del modelo *Hurdle logit* NB2

Para este modelo también se tiene la tasa de clasificación correcta para el estatus de fumador, la cual es de 57.52 por ciento, menor a la obtenida mediante el modelo *hurdle logit* Poisson.

Cuadro N° 20: Ap1: Comparación de valores observados versus predichos (modelo *hurdle logit* NB2)

		Observación		
	Status	No Fuma	Fuma	TOTAL
Predicción	Ceros	**159**	16	175
	Positivos	63	**28**	91
TOTAL		222	44	**266**

FUENTE: Elaboración propia

Este modelo no es adecuado por el hecho de no poder explicar los conteos mediante ningún factor, covariable o interacción de éstos. De todos modos, se utilizará en la comparación de modelos, sin embargo no se puede elegir éste como el mejor.

4.1.8 COMPARACIÓN DE MODELOS

Cuadro N° 21: Ap1: Indicadores de ajuste de todos los modelos

	Poisson	NB2	Poisson Inflado en cero	NB2 Inflado en cero	Hurdle (*logit*) Poisson	Hurdle (*logit*) NB2
Deviance	*422.64*	**111.45**	---	---	---	---
$\frac{\chi^2_{Pearson}}{gl}$	*3.235*	**0.988**	1.2495	**0.9956**	1.1966	*0.5097*
Pseudo R^2 basado en la verosimilitud	**0.5678**	0.1190	0.3364	0.1196	0.3427	*0.0924*
AIC	*564.3847*	393.3500	**390.5283**	395.3500	**388.919**	402.2991
$\lvert r_{Pearson} \rvert > 2$	*26*	13	**10**	13	13	**5**
Número de parámetros estimados	8	6	9	7	10	5
Número de variables predictoras	5	4	5	4	5	2

FUENTE: Elaboración propia

De acuerdo al Cuadro N° 21, el modelo NB2 es mejor al modelo Poisson ya que su Deviance es menor y el ratio $\frac{\chi^2_{Pearson}}{gl}$ es cercano a uno, además de ello, el AIC del modelo NB2 es menor que en el modelo Poisson, tiene menos residuales de Pearson que sobrepasan las dos desviaciones estándar, asimismo emplea un menor número de variables para construir el modelo y estima menos parámetros. Respecto a los otros modelos se tienen indicadores similares; en la página 105 se indicó que el modelo NB2 inflado en cero era incapaz de

explicar la proporción de ceros estructurales, debido a ello si se comparan los indicadores entre el modelo NB2 y el modelo NB2 inflado en cero, éstos resultan muy similares. Ante un ajuste similar (Pseudo-R^2 y $\Delta AIC < 2$) se prefiere el modelo NB2 el cual es más simple que el modelo NB2 inflado en cero.

Se presentan en negrita y mayor tamaño los *mejores* indicadores obtenidos para los modelos y en cursiva y menor tamaño los peores indicadores, de esta manera se observa que los *peores* modelos que ajustan al conjunto de datos de consumo de cigarros semanales son **Poisson** y ***hurdle logit* NB2**, mientras que los que comparten más indicadores de *mejor* ajuste son el modelo **NB2** y el modelo **Poisson inflado en cero**.

No obstante, estas comparaciones deben ser contrastadas mediante pruebas de hipótesis utilizando alguno de los tests ya propuestos (razón de verosimilitudes o el test de Vuong). En la siguiente página se muestra un diagrama que resume las comparaciones entre todos los modelos propuestos, los cuales resultaron en su totalidad no anidados; las líneas azules indican que los modelos son similares mientras que las líneas rojas señalan que es preferible elegir el modelo con menor AIC. Para cada comparación se muestra la diferencia de AICs y el pvalor obtenido en la prueba de Vuong, donde la hipótesis nula indica similitud de los modelos en comparación. Luego, se consolida la información de la Figura 13 en el Cuadro N° 23 de la página 117 donde se muestran los modelos estimados ordenados de mayor a menor AIC.

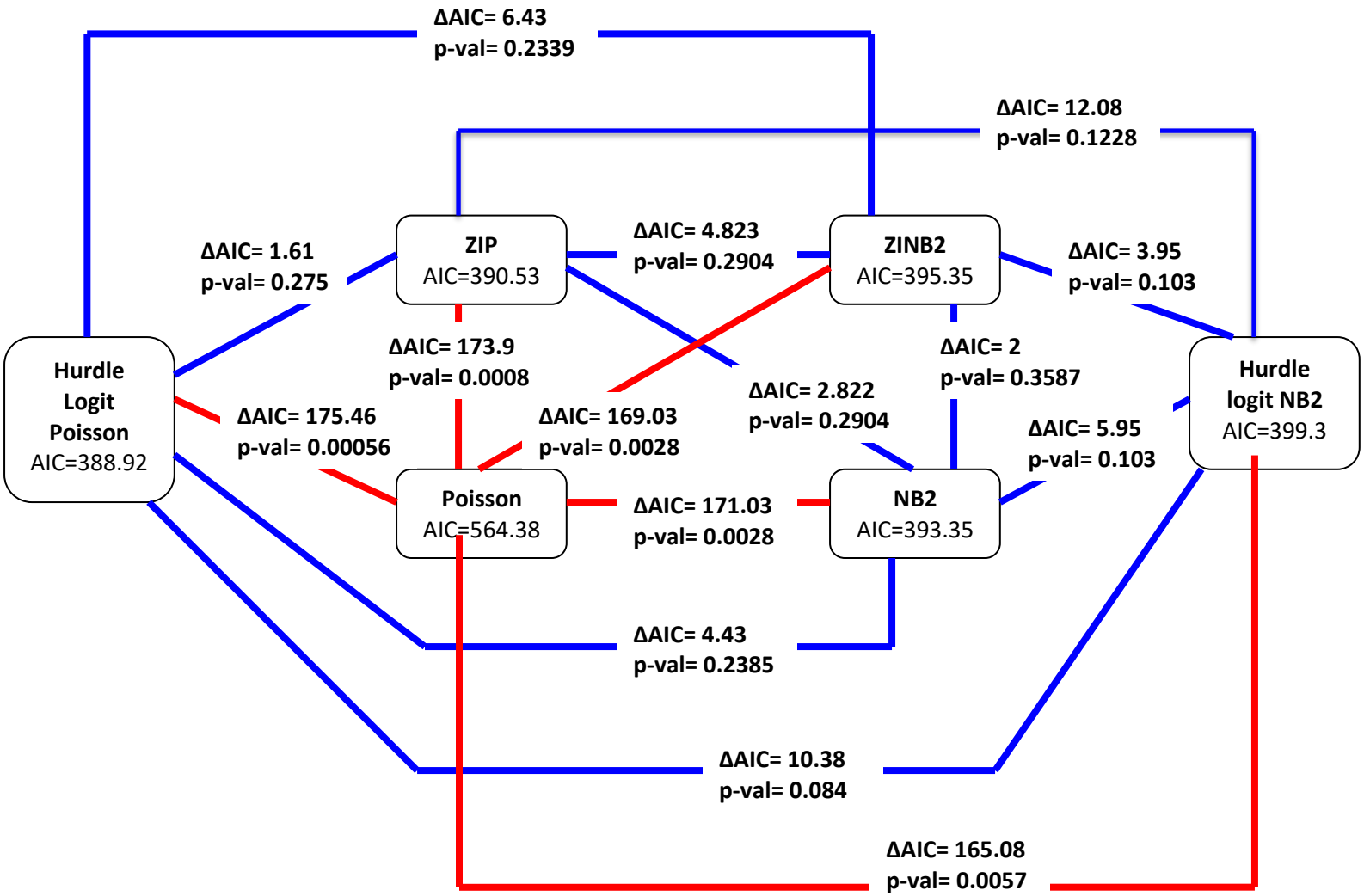

Figura 13 Ap1: Comparación de los modelos propuestos: Clásicos, inflados en cero y *hurdle*

Cuadro N° 22: Ap1: Resumen de modelos obtenidos

	POISSON	**NB2**	**ZIP**		**ZINB2**		***Hurdle logit* Poisson**		***Hurdle logit* NB2**	
			C.C.	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-4.5704	-8.2606	-6.07	1.53	-8.2605	-8.291	-5.59	-3.02	-6.294	-2.6986
	***	***	***	***	***	n.s.	***	***	n.s.	***
Bebidas	2.3313	2.1794	1.48	-1.95	2.1794		1.553	1.705		1.7787
	***	***	***	***	***		***	***		***
FumaComp	1.9804	1.6443	1.478		1.6422		1.217	1.509		1.6576
	***	***	***		***		***	***		***
Charlas	-0.495	-1.0708	0.059		-1.0708		0.256			-0.8373
	*	**	n.s.		*		n.s.			*
Edad	0.1159	0.3278	0.286		0.3278		0.27			
	.	**	***		**		***			
FumaCasa	-11.77		0.815							
	***		***							
Beb*Charla	-1.681		-2.24				-2.501			
	***		***				***			
Edad*Fcasa	0.6418									
	***									
Theta	---	0.1903	---	---	0.1903	---	---	---	0.0001794	---
Deviance	422.64	111.45	---	---	---	---	---	---	---	---
AIC	**564.38**	**393.35**	**390.5283**		**395.3509**		**388.9188**		**399.2992**	
LogVero	**-274.19**	**-190.675**	**-186.3**		**-190.7**		**-184.4594**		**-193.6**	

FUENTE: Elaboración propia

Cuadro N° 23: Ap1: Resumen para la comparación de modelos

Poisson	Hurdle *Logit* NB2	NB2 inflado en cero	NB2	Poisson inflado en cero	Hurdle Poisson
	a	a			
	b		b		
	c			c	
	d				d
		e	e		
		f		f	
		g			g
			h	h	
			i		i
				j	j

FUENTE: Elaboración propia

En este cuadro, dos letras iguales indican que los modelos son indistinguibles (uno no es mejor que otro). Según la disposición de los modelos por AIC se observa que cualquier modelo es mejor que el modelo Poisson. Luego, según lo discutido en los modelos NB2 inflado en cero y *hurdle logit* NB2, éstos no son adecuados porque no son capaces de modelar la proporción de ceros estructurales y conteos positivos, respectivamente; dicho de otra manera, se podría emplear el modelo NB2 y se llegan a las mismas estimaciones y conclusiones. Entonces, restan comparar tres modelos: binomial negativo, Poisson inflado en cero y *hurdle* Poisson. De estos se puede elegir el modelo **NB2** por ser más parsimonioso, ya que estima sólo seis parámetros respecto a los nueve parámetros estimados en el modelo Poisson inflado en cero y los diez del modelo *hurdle* Poisson, además de ello, el ajuste es similar en los tres modelos, tanto en AIC como en cantidad de residuales de Pearson que sobrepasan las dos desviaciones estándar.

Una consecuencia de trabajar con un modelo Poisson cuando este no se ajusta correctamente es la subestimación de los errores estándar de los coeficientes estimados, como se muestra a continuación:

Cuadro N° 24: Ap1: Errores estándar de los modelos estimados

Variable Predictora	Poisson	Hurdle *Logit* NB2	NB2 inflado en cero	NB2	Poisson Inflado en Cero	Hurdle Poisson
Intercepto	1.163	83.187	2.181	1.978	1.136	1.160
Bebidas	0.205	---	0.460	0.459	0.215	0.231
Charlas	0.244	---	0.423	0.403	0.261	0.278
Fuma Comp.	0.276	---	0.472	0.468	0.348	0.361
Edad	0.059	---	0.117	0.101	0.060	0.060
Fuma Casa	1.832	---	---	---	0.174	0.181
Bebidas*Charlas	0.333	---	---	---	0.363	0.393
Edad*FumaCasa	0.094	---	---	---	---	---

FUENTE: Elaboración propia

Para el caso de los modelos inflados en cero y *hurdle* sólo se muestran los coeficientes estimados del submodelo Poisson o binomial negativo según sea el caso. Se aprecia que los modelos de tipo Poisson estiman una mayor cantidad de coeficientes y los errores estándar asociados a éstos son menores que en los modelos NB2.

4.2 APLICACIÓN DOS: TASA DE PESCA

4.2.2 ANÁLISIS EXPLORATORIO

4.2.2.1 ANÁLISIS EXPLORATORIO DE LA VARIABLE RESPUESTA

La variable respuesta de conteo, *número de peces capturados por pescador*, presenta exceso de ceros (56.8 por ciento) y posible sobredispersión ya que se obtiene un promedio muestral de 3.296 peces y varianza de 135.38 peces^2; además de ello si se simula una variable aleatoria con distribución Poisson y media 3.296, se esperaría obtener sólo tres o cuatro por ciento de ceros, sin embargo, este porcentaje es vastamente excedido en este conjunto de datos. Un gráfico de varas para esta variable se muestra a continuación:

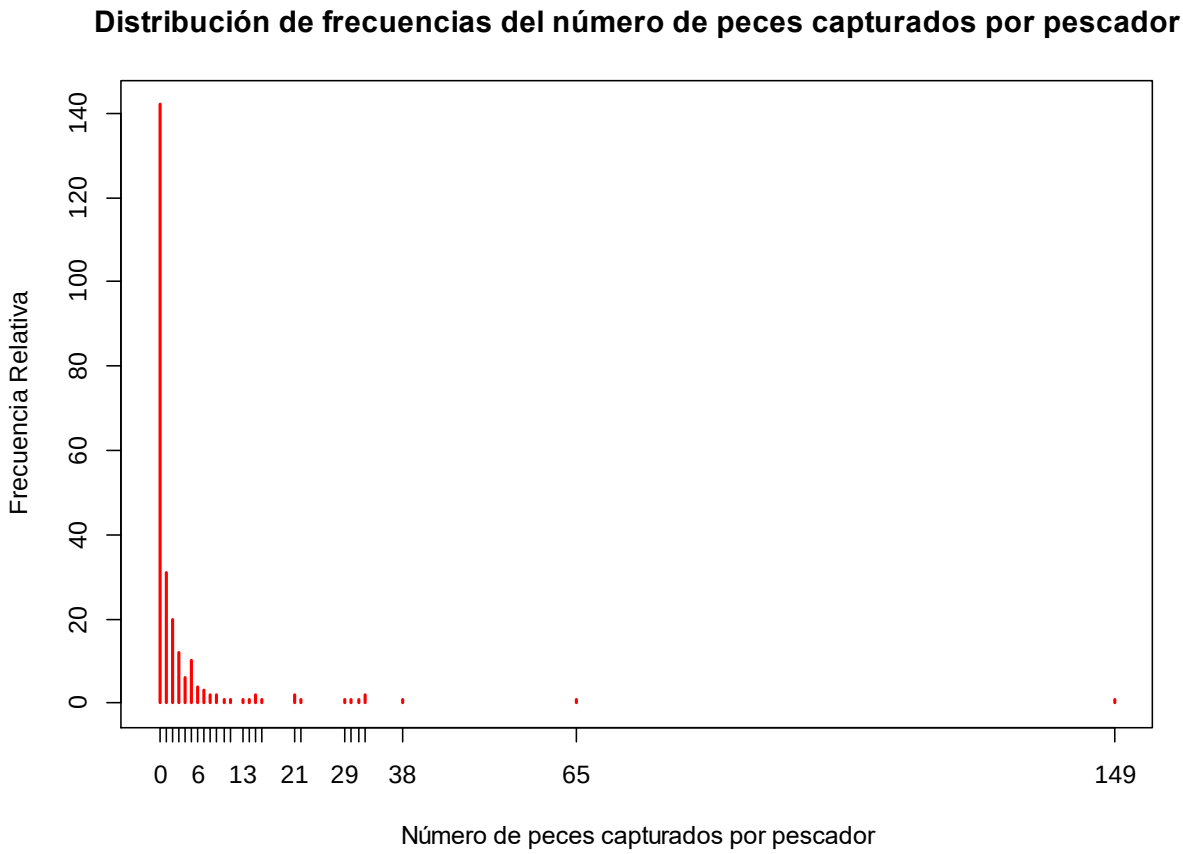

Figura 14. Distribución de frecuencias del Número de peces capturados por pescador

4.2.2.2 ANÁLISIS EXPLORATORIO DE LAS VARIABLES PREDICTORAS

El 50.8 por ciento de los pescadores acude acompañado de una o dos personas, siendo la media de 2.528 personas, además el 52.8 por ciento de los pescadores no acude a pescar con niños, siendo la media en este caso de 0.684 niños por pescador. Por otro lado, el 41.2 por ciento de los pescadores encuestados no va a pescar en casa rodante.

Uno de los supuestos para el desarrollo de modelos lineales generalizados es la independencia entre variables explicativas, por ello se muestra a continuación la correlación entre los factores y/o variables predictoras a utilizar en los modelos de regresión para datos de conteo.

Cuadro N° 25: Matriz de correlaciones entre las variables predictoras

	Persons	Child	Camper
Persons	1		
Child	0.546	1	
Camper	-0.052	-0.039	1

FUENTE: Elaboración propia

El tipo de correlaciones empleadas es la siguiente:

	Coeficiente de correlación de Pearson (entre dos variables cuantitativas).
	Coeficiente biserial (Entre una variable nominal dicotómica y una cuantitativa).

Se observa que, a excepción de la correlación entre el número de niños y personas, la cual es moderada, las otras dos correlaciones son bastante cercanas a cero.

4.2.3 MODELO DE REGRESIÓN POISSON

a. Modelo estimado

El modelo de regresión Poisson se construye según la metodología planteada en la página 73, usando la función de enlace logarítmica. En el ANEXO 9 se presenta en detalle los modelos estimados paso a paso; el modelo estimado resultante es el siguiente:

$$\hat{\mu}_i = \exp(-1.161 + 0.832X_{1i} - 1.17X_{2i} - 0.163X_{3i} + 0.338X_{1i}X_{3i})$$

Para $i = 1, ..., 250$

Donde: $X_1 : \text{N}° \text{Personas}$ $X_2 : \text{N}° \text{niños}$ $X_3 : \text{Casa rodante}$

Como se observa en el ANEXO 9, la prueba de Wald resulta significativa para todos los factores e interacciones, excepto para el factor casa rodante, sin embargo se incluye ya que el efecto de una interacción en la que está presente esta variable resulta significativo.

La única variable que no interactúa es $X_2 : \text{N}° \text{niños}$, para ella se tiene que cuando el número de niños acompañando al pescador se incrementa en uno, la tasa de peces capturados disminuye en 69 por ciento, manteniendo las demás variables en valores constantes. Luego, para las variables cuya interacción resulta significativa, la interpretación se realiza para la combinación de niveles. Para las variables $X_1 : \text{N}° \text{Personas}$ y $X_3 : \text{Casa rodante}$ se tiene que analizar según la interacción; primero cuando el pescador no va en casa rodante se tiene que cuando el número de personas aumenta en uno, la tasa de peces capturados aproximadamente se duplica mientras que si el pescador sí va en casa rodante, al añadir un acompañante adicional, la tasa de peces capturados aproximadamente se triplica.

d. Ajuste del modelo Poisson

Cuadro N° 26: Ap2: Indicadores de ajuste en el modelo Poisson

Criterio	Interpretación
Deviance	Al poner a prueba la bondad de ajuste mediante el test de Deviance, se rechaza la hipótesis nula de que el modelo se ajusta a los datos, al obtener $Deviance = 1323.7 \sim \chi^2_{(245)}$ con pvalor aproximado de cero.
Estadístico Chi Cuadrado de Pearson	Respecto al estadístico Chi cuadrado de Pearson, el cociente de éste entre sus grados de libertad da el valor de 11.45627, indicando que el modelo Poisson no está correctamente especificado (la función de varianza no es adecuada, es decir, la distribución probabilística elegida no es la correcta).
Pseudo R^2 basado en la verosimilitud	El Pseudo R^2 basado en la verosimilitud arroja un valor de 0.4961.
AIC	Toma el valor de 1670.728. Como referencia, se tiene que el AIC del modelo de sólo intercepto es de 3297.431
Residuales de Pearson	Analizando los residuales bajo este modelo Poisson se tienen 40 residuales de Pearson que sobrepasan el rango de las 2 desviaciones estándar, y 19 *leverages* que sobrepasan el límite $2p/n = 2\times5/250 = 0.04$. Se presentan en la siguiente página las gráficas de residuales y *leverages*.

FUENTE: Elaboración propia

En el gráfico *Residuals vs Fitted*, los *outliers* detectados para el modelo ocupan las posiciones 89, 138 y 179. Se observa un patrón indicando que no hay homogeneidad de varianzas, lo cual es aceptable ya que por definición la varianza no es constante sino que depende de la media. Luego, en el segundo gráfico (*Normal Q-Q*), no se espera obtener un

gráfico que muestre normalidad de los residuales, y efectivamente sucede tal cual, nuevamente las observaciones 89, 138 y también la 44 se alejan considerablemente de la línea normal.

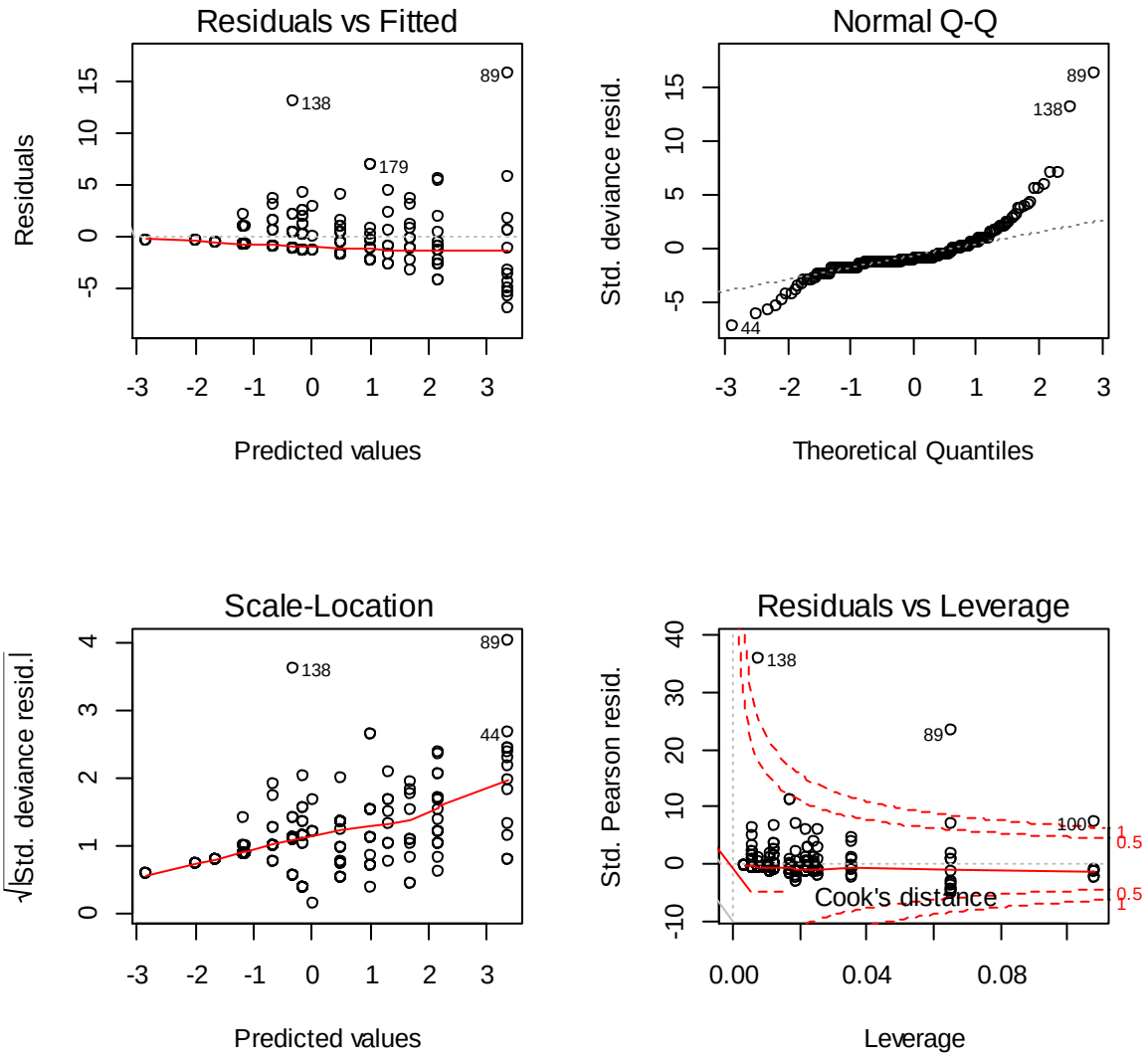

Figura 15. Ap2: Gráfico de residuales y *leverages* del modelo Poisson

Después, en el gráfico *Scale Location* también se presenta un patrón que no indica homogeneidad de varianzas, con *outliers* los puntos 44, 89 y 138. Finalmente, en el gráfico de *Residuales estandarizados de Pearson versus leverages* (*hat*), los valores extremos

horizontalmente indican altos *leverages*. Residuales estandarizados de Pearson mayores a |2| y con *leverages* altos, en este caso el punto 100, indica un mal ajuste en el modelo. Si se amplía esta última gráfica, los *leverages* son aquellos cuyos valores en la diagonal de la matriz *hat* se encuentran por encima de 2p/n, los cuales son 19, gráficamente, los valores a la derecha de la línea vertical indican los *leverages* de este modelo, sin embargo los puntos 100, 89 y 160 resaltan ante los demás. Los puntos por encima de la línea horizontal superior y por debajo de la línea horizontal inferior son *outliers*.

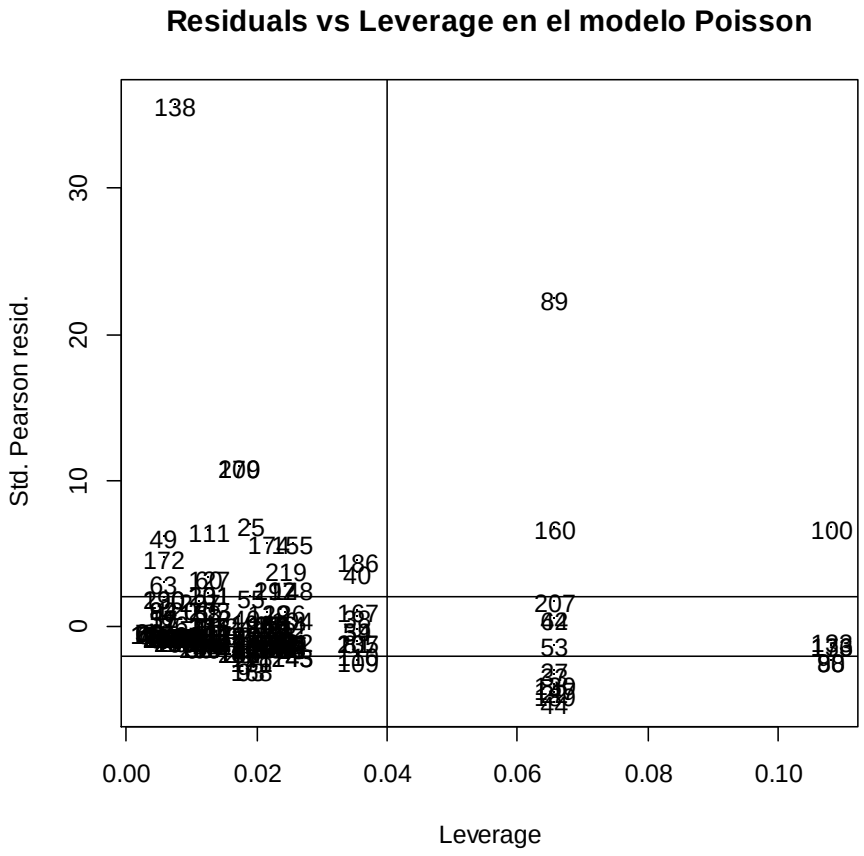

Figura 16. Ap2: Gráfico de *leverages* versus residuales del modelo Poisson

Según los resultados de mal ajuste obtenidos para el modelo Poisson, es probable que exista sobredispersión para la variable respuesta en este modelo, lo cual debe ser confirmado mediante los tests propuestos de Bohning y Dean & Lawless.

e. Pruebas de sobredispersión

- **Test de Bohning**

Se prueba la hipótesis nula $H_0 : \mu_Y = \sigma_Y^2$, la cual se rechaza ya que se obtiene un estadístico de prueba de 116.5504 y pvalor equivalente a cero, entonces se puede afirmar que la media de peces capturados por pescador difiere de su varianza.

- **Test de Cameron y Trivedi (1985)**

Se busca probar $H_0 : E[Y_i] = V[Y_i] = \mu_i$, entonces el p-valor obtenido para el estadístico de prueba $Z_{calc} = 1.9123$ es de 0.0279, con un estimado de dispersión de 11.25, similar al del estadístico chi cuadrado. Por lo tanto se rechaza la hipótesis nula; existe sobredispersión

- **Regresión de la varianza estimada sobre la media estimada**

Se plantea la hipótesis según el test propuesto por Cameron y Trivedi (1986).

$$H_0 : \beta_1 = 1 \qquad t_{calc} = \frac{44.472 - 1}{7.698} = 5.647 \sim t_{249} \qquad pvalor \approx 0$$
$$H_1 : \beta_1 \neq 1$$

Existe suficiente evidencia estadística para rechazar la hipótesis nula, es decir, no existe equidispersión.

Luego, la hipótesis según el test propuesto por Cameron y Trivedi (1990).

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0 \qquad t_{calc} = \frac{1.7732 - 0}{0.2909} = 6.095 \sim t_{249} \qquad pvalor \approx 0$$

Existe suficiente evidencia estadística para rechazar la hipótesis nula, es decir, no existe equidispersión

- **Prueba de Dean y Lawless**

El estadístico T calculado es igual a 139.2557, dado que su distribución asintótica es normal, el $pvalor = P(Z > 139.2557) \approx 0$, entonces se rechaza la hipótesis nula de equidispersión de la variable respuesta en el modelo Poisson.

Por lo visto según el estadístico Chi Cuadrado de Pearson, el test de Bohning, los *tests* de Cameron y Trivedi y el de Dean y Lawless, está presente el problema de sobredispersión en la variable respuesta del modelo Poisson, por ello la metodología indica que se deben modelar los datos utilizando la regresión NB2, inflada en cero o *hurdle*.

4.2.4 MODELO DE REGRESIÓN BINOMIAL NEGATIVO (NB2)

a. Modelo estimado

Se mostró que el modelo anterior (Poisson) no era adecuado debido al problema de sobredispersión. Debido a ello, los errores estándar de los coeficientes de regresión se encontraban subestimados; en contraparte a ello, mediante el modelo de regresión NB2, se corrige la función de varianza, la cual pasa de ser μ a $\mu+\alpha\mu^2$. Luego, para estimar el número de cigarros consumidos semanalmente, la ecuación de regresión es:

$$\hat{\mu}_i = \exp\left(-1.625+1.0608X_{1i}-1.7805X_{2i}+0.6211X_{3i}\right); \quad \hat{\alpha}=0.4635^{-1}$$

Para $i=1,...,250$

Donde: X_1 : N° Personas X_2 : N° niños X_3 : Casa rodante

Interpretando los coeficientes estimados:

- Cuando el número de personas que acompaña al pescador se incrementa en una, la tasa de peces capturados de este pescador aproximadamente se triplica $\left(\exp(1.0608)=2.889\right)$, manteniendo las demás variables en valores constantes.
- Cuando el número de niños que acompaña al pescador se incrementa en uno, la tasa de peces capturados de este pescador disminuye en 83 por ciento $\left(\exp(-1.7805)=0.169\right)$, manteniendo las demás variables en valores constantes.
- Cuando el pescador va al lago en casa rodante, su tasa de peces capturados es 86 por ciento mayor $\left(\exp(0.6211)=1.861\right)$ que cuando no la lleva.
- El estimado $\hat{\alpha}=0.4635^{-1}$ sirve para cuantificar la relación entre la media y varianza en el modelo NB2, entonces se tiene que la función de variancia estimada para este modelo es $\sigma^2=\mu+2.157\mu^2$.

b. Ajuste del modelo NB2

Cuadro N° 27: Ap2: Indicadores de ajuste en el modelo NB2

Criterio	Interpretación
Deviance	Se obtiene $Deviance = 210.65 \sim \chi^2_{(246)}$, para el cual $pvalor = P\left(\chi^2_{(246)} > 210.65\right) = 0.95$. Entonces no se puede afirmar que el modelo no se ajuste a los datos.
Estadístico Chi Cuadrado de Pearson	Toma el valor de 2.4658, si bien es menor al valor obtenido para el modelo Poisson, aún no podría indicarse que el modelo de regresión binomial negativo está especificado correctamente (la distribución empleada es adecuada para el modelamiento), ya que el cociente obtenido no es cercano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 obtenido para el modelo es de 0.1275, menor que en el modelo Poisson.
AIC	AIC: 820.44, mientras que para el modelo nulo AIC: 932.8786
Residuales de Pearson	Analizando los residuales bajo este modelo Poisson se tienen 12 residuales de Pearson que sobrepasan el rango de las 2 desviaciones estándar, y ningún que sobrepasan el límite $2p/n = 2\times 5/250 = 0.04$. Se presentan en la siguiente página las gráficas de residuales y *leverages*

FUENTE: Elaboración propia

En el gráfico *Residuals vs Fitted*, el *outlier* detectado para el modelo ocupa la posición 138. Además, se observa un patrón indicando que no hay homogeneidad de varianzas, lo cual es aceptable ya que por definición la varianza no es constante. Luego, en el gráfico *Normal Q-Q* no se espera obtener un gráfico que muestre normalidad de los residuales, y efectivamente

sucede tal cual. Se observa nuevamente el *outlier* en la observación 138. En el tercer gráfico, *Scale Location*, también se presenta un patrón que no indica homogeneidad de varianzas.

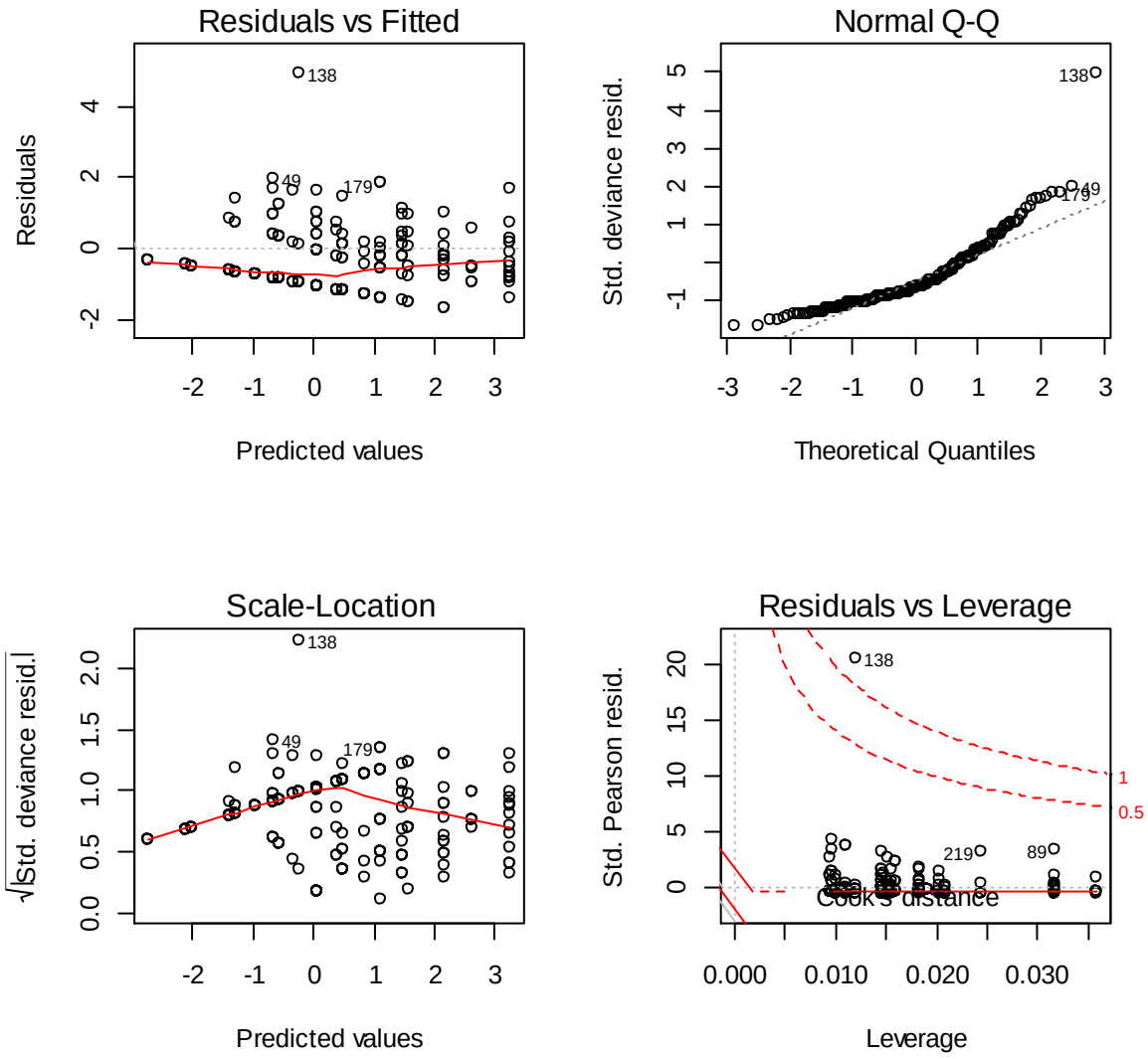

Figura 17. Ap2: Gráfico de residuales y *leverages* del modelo NB2

Finalmente, en la gráfica de *residuales estandarizados de Pearson versus leverages* (*hat*), se observan el punto 138 con residual alto mas no es un *leverage*. Esta última gráfica se muestra ampliada en la siguiente página.

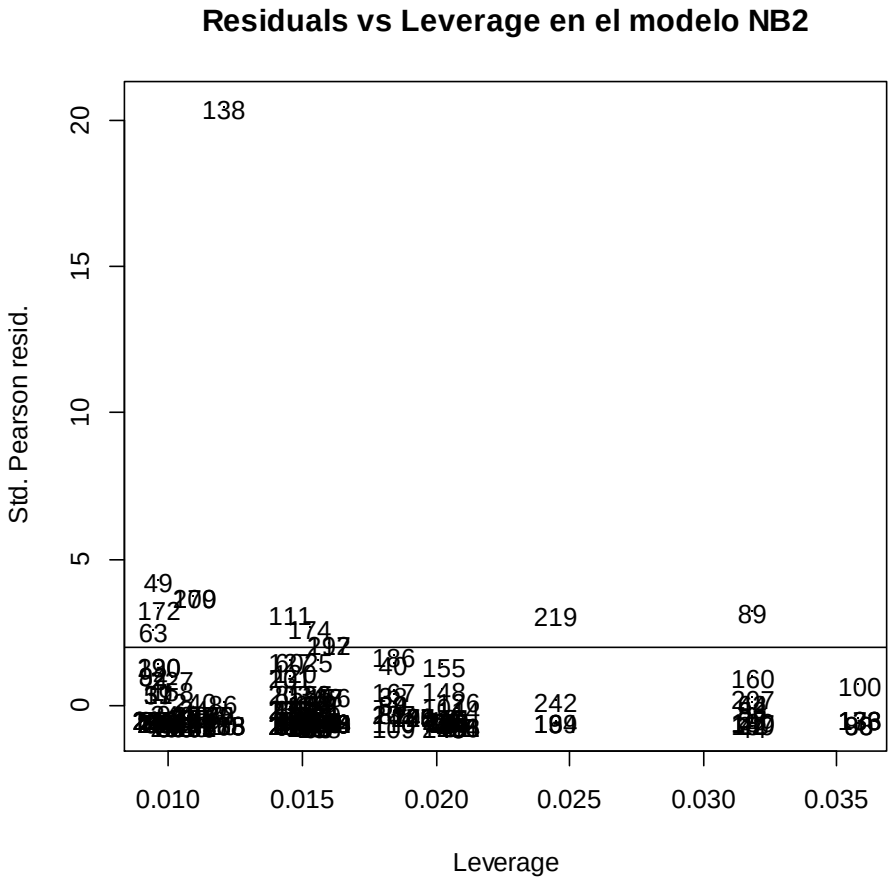

Figura 18. Ap2: Gráfico de *leverages* versus residuales del modelo NB2

Los puntos por encima de la línea horizontal indican residuales de Pearson altos (*outliers*); este modelo presenta menos *outliers* y ningún *leverage*, lo cual lo diferencia bastante del modelo Poisson, en el cual tanto *outliers* como *leverages* eran abundantes.

4.2.5 MODELO DE REGRESIÓN POISSON INFLADO EN CERO

a. Modelo estimado

Ecuación de regresión estimada para la media en los conteos aleatorios

$$\hat{\mu}_i = \exp\left(-0.053 + 0.544X_{1i} + 1.0831X_{2i} - 0.475X_{3i} - 0.3423X_{1i}X_{2i} - 1.1975X_{2i}X_{3i} + 0.4255X_{1i}X_{3i}\right)$$

Para $i = 1,...,250$

Donde: X_1 : N° Personas X_2 : N° niños X_3 : Casa rodante

Las interpretaciones de las tasas se realizan en función a los *Rate Ratios* dependiendo los valores que tomen las variables predictoras, de manera similar a los modelos anteriores y/o a los modelos propuestos en la aplicación uno.

Ecuación de regresión estimada para la proporción de ceros estructurales

$$\text{logit}\left(\hat{\pi}_i\right) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = 1.8726 - 0.956X_{1i} + 2.039X_{2i} - 1.126X_{3i}$$

Para $i = 1,...,250$

Donde: X_1 : N° Personas X_2 : N° niños X_3 : Casa rodante

Los coeficientes estimados se interpretan de la siguiente manera:

- La chance de que un pescador no capture ningún pez debido a que no pescó, disminuye en 62 por ciento $\left(\exp(-0.956) = 0.38\right)$ cuando el número de acompañantes se incrementa en uno, manteniendo las demás variables en valores fijos.
- La chance de que un pescador no capture ningún pez debido a que no pescó, aproximadamente se octuplica $\left(\exp(2.039) = 7.68\right)$ cuando el número de niños se incrementa en uno, manteniendo las demás variables en valores fijos.

- Cuando el pescador va al lago en casa rodante, la chance de que no capture ningún pez debido a que no pescó, disminuye en 68 por ciento $\left(\exp(-1.126) = 0.324\right)$ que cuando no va en casa rodante.

c. Ajuste del modelo Poisson inflado en cero

Cuadro N° 28: Ap2: Indicadores de ajuste en el modelo Poisson inflado en cero

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 874.1753, de modo que $P\left(\chi^2_{(239)} > 874.1753\right) \approx 0$. Además el cociente de este estadístico entre sus grados de libertad es 3.6576, un valor no cercano a uno.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.3570
AIC	El AIC del modelo en estudio es 1471.256, mientras que el del modelo de sólo intercepto es 2258.046
Residuales de Pearson	Se observan 19 residuales de Pearson por encima de las dos desviaciones estándar. Ocho residual por debajo de las dos desviaciones estándar.

FUENTE: Elaboración propia

En la Figura 19 de la siguiente página se tiene que las observaciones 89, 138, 179 y 200 destacan por tener altos residuales de Pearson.

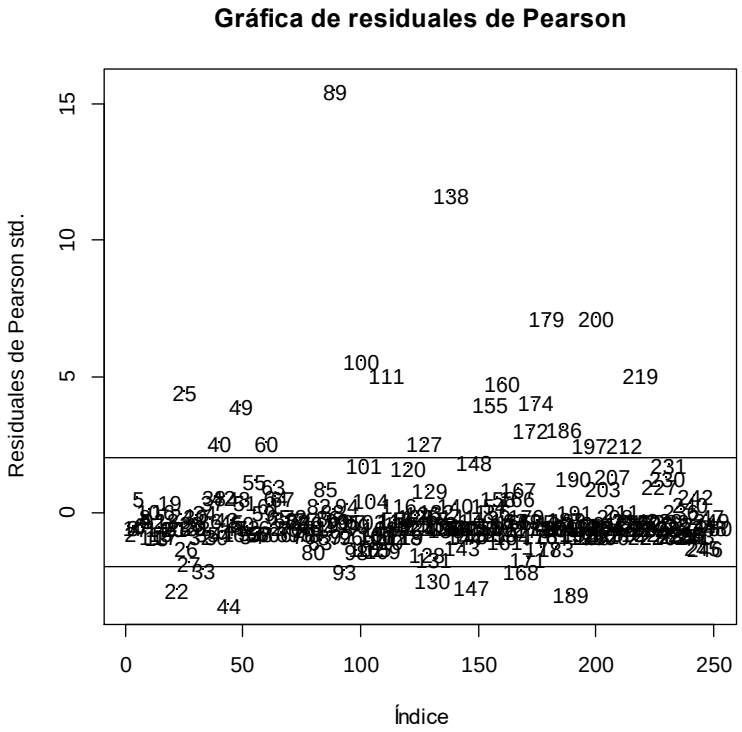

Figura 19. Ap2: Gráfico de residuales de Pearson del modelo Poisson inflado en cero

En el Cuadro N° 29 de la siguiente página se muestran los resultados para las observaciones 15 – 30: en la primera columna aparece el número observado de peces capturados por cada pescador, en la siguiente columna la probabilidad que la observación provenga de un cero estructural y a partir de la tercera columna la probabilidad de que la variable respuesta tome el valor cero, uno, dos, etc. Si bien el rango de la distribución Poisson se extiende hasta infinito, R sólo muestra los valores hasta 149, ya que es el máximo obtenido en la muestra (sin embargo por cuestión de espacio sólo se muestra en esta hoja hasta $y = 17$).

Para la observación 22, su respuesta fue que consume cinco cigarros semanalmente. Según el modelo, se obtiene una predicción de cero peces capturados con 0.75 de probabilidad, sin embargo, de esta probabilidad se puede decir que se tiene 0.71 de probabilidad de ser cero estructural (no pescó) versus 0.04 de que el cero sea aleatorio (pescó pero no cogió ningún pez). Luego, la probabilidad de que capture sólo un pez es de 0.09, de que capture dos es 0.08 y así sucesivamente.

Cuadro N° 29: Ap2: Predicción en el modelo Poisson inflado en cero

Id	Y	Cero estructural	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
15	0	0.45	0.56	0.18	0.14	0.07	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	1	0.45	0.56	0.18	0.14	0.07	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0	0.52	0.52	0.01	0.02	0.04	0.06	0.07	0.08	0.07	0.05	0.04	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
18	0	0.88	0.92	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	1	0.71	0.75	0.09	0.08	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0	0.74	0.74	0.01	0.02	0.03	0.04	0.05	0.04	0.03	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	1	0.52	0.52	0.01	0.02	0.04	0.06	0.07	0.08	0.07	0.05	0.04	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00
22	5	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
23	0	0.71	0.75	0.09	0.08	0.05	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24	3	0.24	0.25	0.05	0.11	0.15	0.15	0.12	0.08	0.05	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	30	0.11	0.11	0.00	0.00	0.00	0.01	0.02	0.04	0.06	0.08	0.10	0.11	0.11	0.10	0.08	0.06	0.04	0.03	0.02
26	0	0.24	0.25	0.05	0.11	0.15	0.15	0.12	0.08	0.05	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27	13	0.04	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
28	0	0.88	0.88	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29	0	0.71	0.77	0.09	0.07	0.04	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	0	0.98	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

FUENTE: Elaboración propia

.

4.2.6 MODELOS DE REGRESIÓN BINOMIAL NEGATIVO INFLADO EN CERO

a. Modelo estimado

Ecuación de regresión estimada para la media en conteos aleatorios

$$\hat{\mu}_i = \exp\left(-1.3119 + 1.0783X_{1i} - 1.2339X_{2i}\right); \quad \hat{\alpha} = \left(\log\left(0.6197\right)\right)^{-1}$$

Para $i = 1,...,250$ Donde: X_1 : N° Personas X_2 : N° niños

Para los coeficientes estimados se tiene las siguientes interpretaciones:

- Cuando el número de acompañantes se incrementa en uno, la tasa de peces capturados aproximadamente se triplica $\left(\exp\left(1.0783\right) = 2.94\right)$, manteniendo constante el número de niños que acompañan al pescador.
- Cuando el número de niños que acompañan al pescador se incrementa en uno, la tasa de peces capturados disminuye en 71 por ciento $\left(\exp\left(-1.2339\right) = 0.29\right)$, manteniendo constante el número de acompañantes del pescador.

Ecuación de regresión estimada para estimar la proporción de ceros estructurales

$$\text{logit}\left(\hat{\pi}_i\right) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -4.2514 + 2.8694X_{2i}$$

Para $i = 1,...,250$ Donde: X_2 : N° niños

El coeficiente de regresión estimado se interpreta como la chance de que un pescador no capture ningún pez porque no estuvo pescando se incrementa en 17.62 veces $\left(\exp\left(2.8694\right) = 17.626\right)$ cuando el número de niños que acompaña al pescador se incrementa en uno.

b. Ajuste del modelo NB2 inflado en cero

Cuadro N° 30: Ap2: Indicadores de ajuste en el modelo NB2 inflado en cero

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 334.3229, de modo que $P\left(\chi^2_{(244)} > 334.3229\right) = 0.0001$. Además el cociente de este estadístico entre sus grados de libertad es 1.3701, un valor tan cercano a uno si se lo compara con el de otros modelos.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 es igual a 0.1328
AIC	El AIC del modelo ajustado es igual a 817.5458 y el del modelo nulo (de sólo intercepto) es 934.8791.
Residuales de Pearson	Se tienen 11 residuales por encima de las dos desviaciones estándar y ninguno por debajo de las dos desviaciones estándar.

FUENTE: Elaboración propia

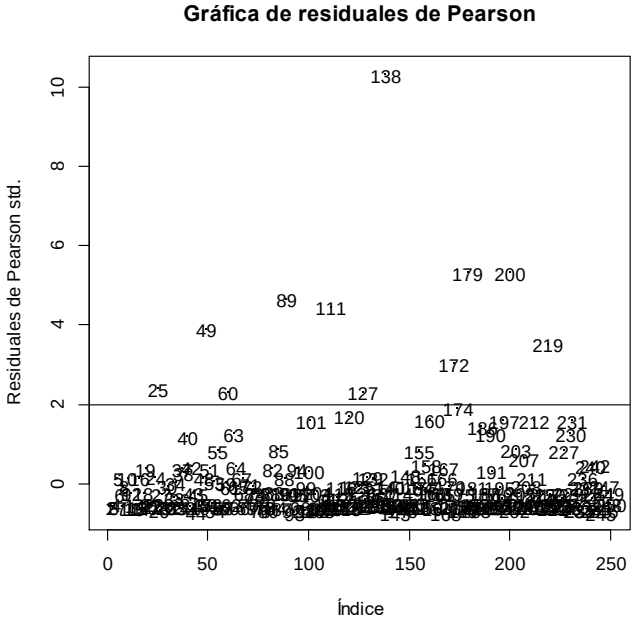

Figura 20. Ap2: Gráfico de residuales de Pearson en el modelo NB2 inflado en cero

4.2.7 MODELOS DE REGRESIÓN HURDLE POISSON

a. Modelo estimado

Se muestra a continuación el modelo estimado, se empleó la función de enlace logarítmica para estimar la media de los datos positivos (Poisson truncado en cero) y se muestran los enlaces *logit* y logarítmico para estimar la proporción de ceros.

Ecuación de regresión estimada para los datos positivos

$$\hat{\mu}_i = \exp\left(0.1439 + 0.4789X_{1i} - 0.0203X_{2i} - 0.538X_{3i} + 0.4535X_{1i}\,\mathrm{X}_{3i} - 1.374X_{2i}\,\mathrm{X}_{3i}\right)$$

Para $i = 1,\ldots,250$

Donde: $X_1 : \mathrm{N}°\,\text{Personas}$ $X_2 : \mathrm{N}°\,\text{niños}$ $X_3 : \text{Caravana}$

Debido a la interacción entre las variables predictoras se interpreta en cada nivel para éstas. Cuando el pescador no va en casa rodante, el incremento unitario de un acompañante, aumenta en 61 por ciento la tasa de peces capturados, manteniendo fijo el número de niños que acude al lago con el pescador. Por otro lado, el incremento unitario de un niño que acompaña al pescador disminuye en dos por ciento la tasa de peces capturados, manteniendo fijo el número de acompañantes al lago.

Luego se tiene la situación en la que el pescador sí lleva casa rodante: el incremento unitario de un acompañante, aumenta en 154 por ciento la tasa de peces capturados, manteniendo fijo el número de niños que acude al lago con el pescador. Por otro lado, el incremento unitario de un niño que acompaña al pescador disminuye en 75 por ciento la tasa de peces capturados, manteniendo fijo el número de acompañantes al lago.

Tomando en cuenta ambos casos se aprecia que el número de acompañantes incrementa la tasa de peces capturados y más aún cuando el pescador va en casa rodante, ocurre lo contrario con el número de niños que acuden al lago con el pescador, éstos disminuyen la tasa de peces capturados y la disminuyen aún más si el pescador va en casa rodante.

Ecuación de regresión estimada para la proporción de ceros

$$\text{logit}(\hat{\pi}_i) = \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = -0.073 - 1.129 X_{2i} + 0.7775 X_{3i}$$

Para $i = 1, ..., 250$ Donde: X_2 : N° niños X_3 : Caravana

La chance de que un pescador "cruce la valla", es decir que empiece a pescar, disminuye en 68 por ciento $\left(\exp(-1.129) = 0.32\right)$ por cada niño adicional que lo acompañe, manteniendo constante el factor casa rodante. Por otro lado, la chance de que un pescador empiece a pescar cuando va en casa rodante es aproximadamente el doble $\left(\exp(0.7775) = 2.18\right)$ que cuando no la lleva, manteniendo fijo el número de niños que lo acompañan.

c. Bondad de ajuste

Cuadro N° 31: Ap2: Indicadores de ajuste en el modelo *hurdle logit* Poisson

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 666.7922, de modo que $P\left(\chi^2_{(241)} > 666.7922\right) \approx 0$. Luego, el cociente de este estadístico entre sus grados de libertad es 2.7667, un valor no cercano a uno.

… continuación

Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.3348
AIC	El AIC del modelo en estudio es 1517.354, mientras que el del modelo de sólo intercepto es 2258.046
Residuales de Pearson	Se observan 16 residuales de Pearson por encima de las dos desviaciones estándar. Ningún residual por debajo de las 2 desviaciones estándar, resaltando los puntos 138 y 89.

FUENTE: Elaboración propia

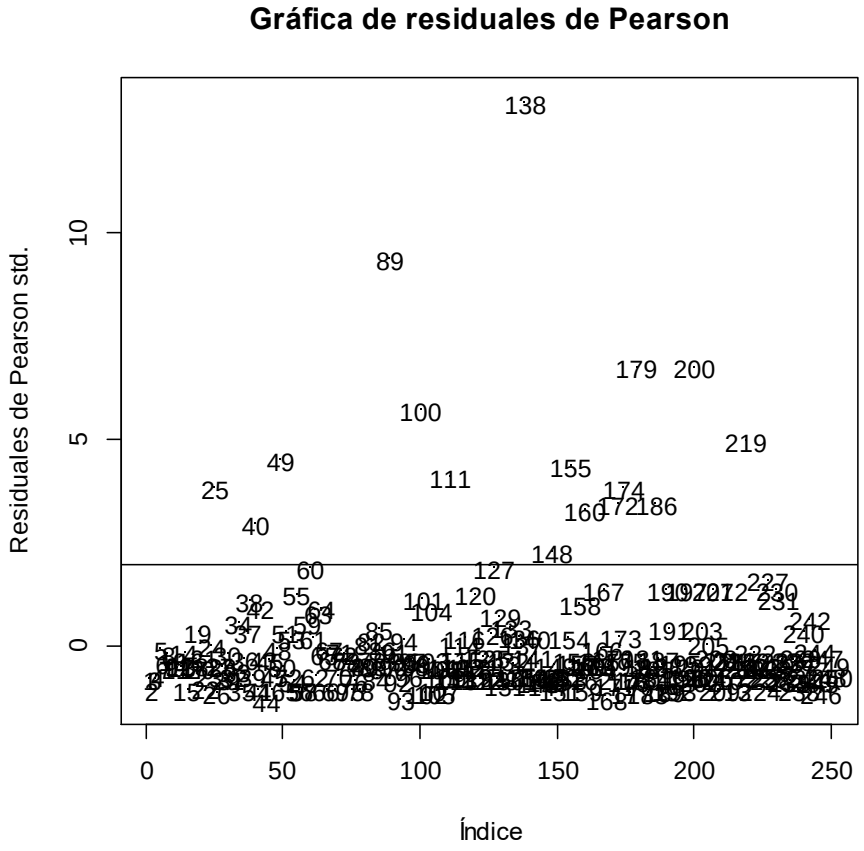

Figura 21. Ap2: Gráfico de residuales de Pearson del modelo *Hurdle logit* Poisson

4.2.8 MODELOS DE REGRESIÓN *HURDLE* BINOMIAL NEGATIVO

a. Modelo estimado

Se muestra a continuación el modelo estimado, se empleó la función de enlace logarítmica para estimar la media de los datos positivos (binomial negativo truncado en cero) y se muestran el enlace *logit* para estimar la proporción de ceros.

Ecuación de regresión estimada para los datos positivos

$$\hat{\mu}_i = \exp\left(-1.4334 + 1.026X_{1i} - 1.1642X_{2i}\right), \text{ con } \hat{\alpha} = \frac{1}{\exp\left(-1.1294\right)} = 3.0938$$

Para $i = 1,...,250$

Donde: $X_1 : \text{N}° \text{Personas}$ $X_2 : \text{N}° \text{niños}$

Se tiene que un incremento unitario en el número de acompañantes aproximadamente triplica la tasa de peces capturados por pescador, manteniendo constante el número de niños que acuden al lago con el pescador. Por otro lado, cuando el número de niños se incrementa en uno, la tasa de peces capturados por pescador disminuye en 69 por ciento $\left(\exp\left(-1.1642\right) = 0.3121\right)$, manteniendo constante el número de acompañantes.

- **Ecuación de regresión estimada para la proporción de ceros**

$$\log it\left(\hat{\pi}\right) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -2.3087 + 1.1104X_{1i} - 2.138X_{2i} + 1.0179X_{3i}$$

Para $i = 1,...,250$

Donde: $X_1 : \text{N}° \text{Personas}$ $X_2 : \text{N}° \text{niños}$ $X_3 : \text{Caravana}$

La chance de que un pescador empiece a pescar se triplica $\left(\exp(1.1104) = 3.035\right)$ cuando el número de acompañantes se incrementa en uno, manteniendo las demás variables en valores fijos. Luego, esta misma chance disminuye en 88 por ciento $\left(\exp(-2.138) = 0.1178\right)$ cuando el número de niños que acuden al lago con el pescador se incrementa en uno, manteniendo las demás variables en valores fijos. Finalmente, cuando el pescador lleva una casa rodante aproximadamente triplica $\left(\exp(1.018) = 2.7673\right)$ esta chance respecto a cuando no la lleva.

b. Bondad de ajuste

Cuadro N° 32: Ap2: Indicadores de ajuste en el modelo *hurdle logit* NB2

Criterio	Interpretación
Estadístico Chi Cuadrado de Pearson	El estadístico Chi cuadrado de Pearson toma el valor de 332.2985, de modo que $P\left(\chi^2_{(242)} > 332.2985\right) = 0.0001$, sin embargo, el cociente de este estadístico entre sus grados de libertad es 1.3731, un valor algo cercano a uno si se compara con los obtenidos en otros modelos.
Pseudo R^2 basado en la verosimilitud	El Pseudo-R^2 toma el valor de 0.1417
AIC	El AIC del modelo en estudio es 807.4923, mientras que el del modelo de sólo intercepto es 928.1759.
Residuales de Pearson	Se observan sólo 9 residuales de Pearson por encima de las dos desviaciones estándar. Ningún residual por debajo de las 2 desviaciones estándar.

FUENTE: Elaboración propia

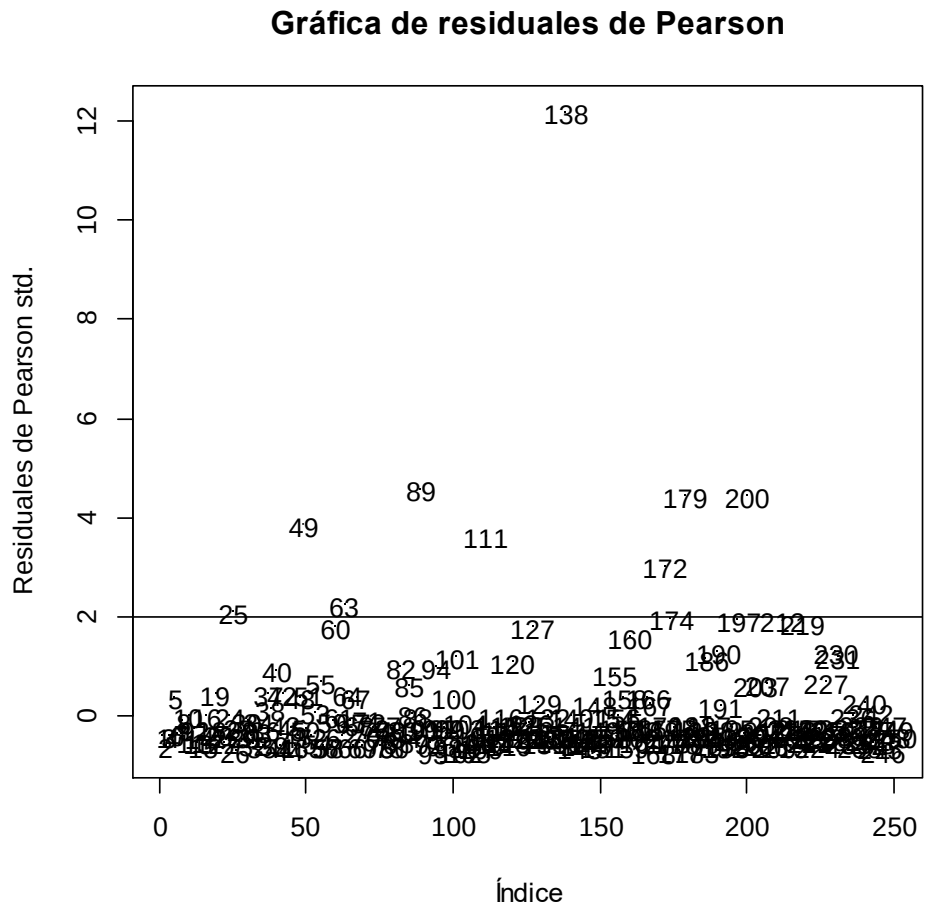

Figura 22. Ap2: Gráfica de Residuales de Pearson para el modelo *hurdle logit* NB2

4.2.9 COMPARACIÓN DE MODELOS

Cuadro N° 33: Ap2: Indicadores de ajuste en todos los modelos

	Poisson	NB2	Poisson Inflado en cero	NB2 Inflado en cero	Hurdle (*logit*) Poisson	Hurdle (*logit*) NB2
Deviance	*1323.7*	**210.65**	---	---	---	---
$\frac{\chi^2_{Pearson}}{gl}$	*11.4563*	2.4658	3.6576	**1.3701**	2.7667	**1.3731**
Pseudo R^2 basado en la verosimilitud	**0.4961**	*0.1275*	0.357	0.1328	0.3348	0.1417
AIC	*1670.728*	820.44	1417.256	**817.5438**	1517.354	**807.4923**
$\lvert r_{Pearson} \rvert > 2$	*40*	**12**	19	**11**	16	**9**
Número de parámetros estimados	5	5	11	6	9	8
Número de variables predictoras	3	3	3	2	3	3

FUENTE: Elaboración propia

En el Cuadro N° 33 se presentan en negrita y mayor tamaño los *mejores* indicadores obtenidos para los modelos y en cursiva y menor tamaño los peores indicadores el modelo, de acuerdo a ello, el modelo ***hurdle logit* NB2** tiene los mejores indicadores (cociente $\chi^2_{Pearson}/gl$ cercano a uno, menor AIC y menor cantidad de residuales de Pearson más allá de las dos desviaciones estándar), aunque el modelo **NB2 inflado en cero** tiene una performance bastante similar. Por otro lado, el modelo que peor ajusta a los datos es el modelo **Poisson**.

No obstante, estas comparaciones deben ser contrastadas mediante pruebas de hipótesis utilizando alguno de los tests ya propuestos (razón de verosimilitudes o el test de Vuong). En la siguiente página se muestra un diagrama que resume las comparaciones entre todos los modelos propuestos, los cuales resultaron en su totalidad no anidados; las líneas azules indican que los modelos son similares mientras que las líneas rojas señalan que es preferible elegir el modelo con menor AIC. Para cada comparación se muestra la diferencia de AICs y el pvalor obtenido en la prueba de Vuong, donde la hipótesis nula indica similitud de los modelos en comparación. Se asume que los modelos *hurdle logit* Poisson y *hurdle* Poisson - Poisson son equivalentes en cuanto a la cantidad de información que brindan ya que presentan AICs muy similares, lo mismo sucede entre los modelos *hurdle logit* NB2 y *hurdle* Poisson NB2.

Finalmente esta información obtenida en la Figura 23 se puede consolidar aún más en el siguiente cuadro de modelos estimados ordenados de mayor a menor AIC:

Cuadro N° 34: Ap2: Resumen para la comparación de modelos

Poisson	Hurdle *Logit* Poisson	Poisson inflado en cero	NB2	NB2 inflado en cero	Hurdle *Logit* NB2
a	a		c	c	
b		b	d		d
				e	e

FUENTE: Elaboración propia

En este cuadro, dos letras iguales indican que los modelos son indistinguibles (uno no es mejor al otro). Según la disposición de los modelos por AIC se observa que cualquiera de los modelos de la familia binomial negativa ajusta mejor que los modelos Poisson, lo cual no discrepa de lo mostrado en el Cuadro N° 33 donde los modelos NB2 son los que muestran mejores indicadores de ajuste.

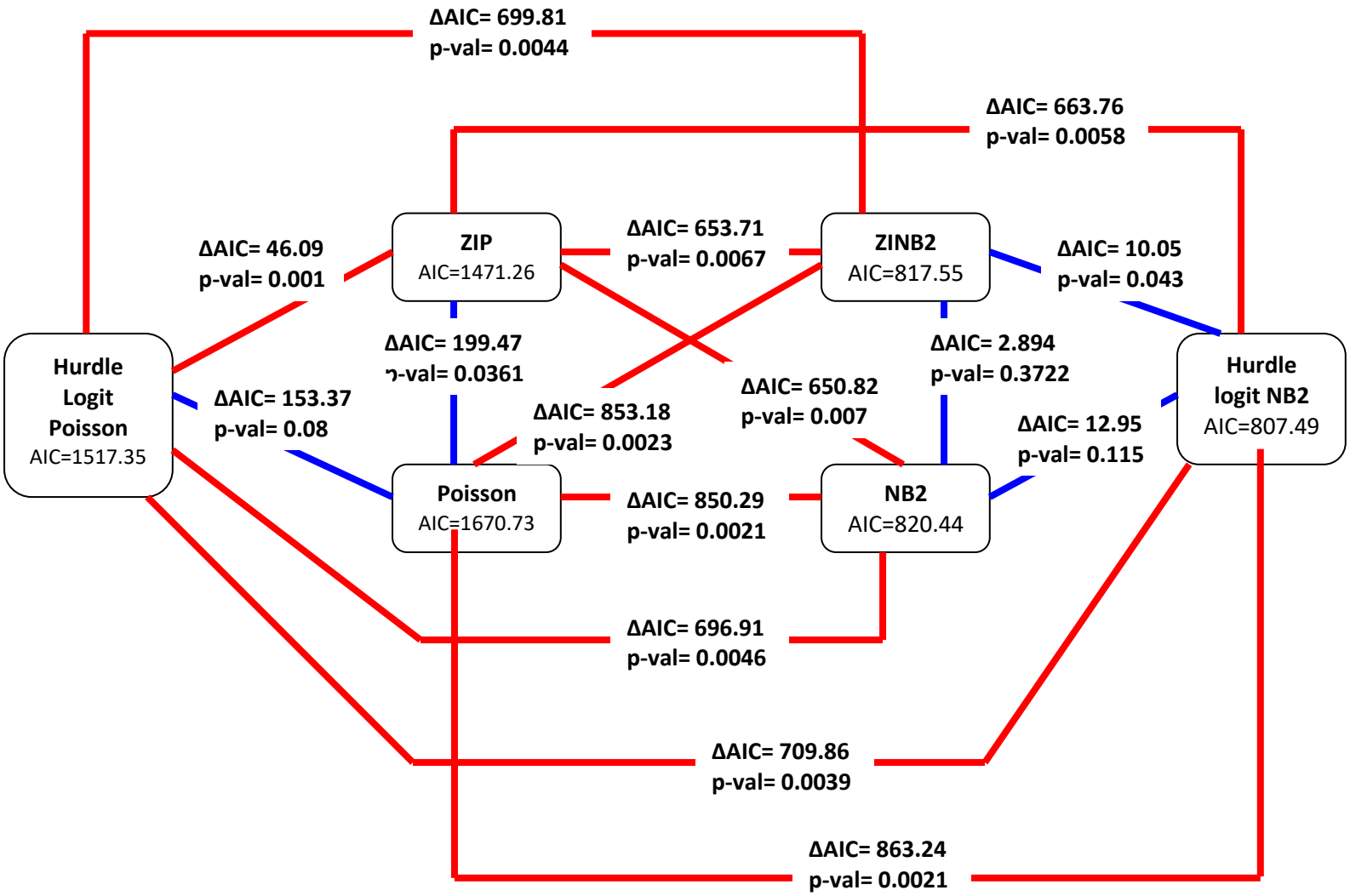

Figura 23. Ap2: Comparación de los modelos propuestos: Clásicos, inflados en cero y hurdle

Cuadro N° 35: Ap2: Resumen de modelos obtenidos para la aplicación dos

	POISSON	**NB2**	**ZIP**		**ZINB2**		***Hurdle logit* Poisson**		***Hurdle logit* NB2**	
			C.C.	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-1.1607	-1.625	-0.053	1.8726	-1.3119	-4.2514	0.1439	-0.073	-1.4334	-2.3087
	***	***	n.s.	***	***	**	n.s.	n.s.	*	***
Persons	0.83192	1.0608	0.5437	-0.956	1.0783		0.4789		1.026	1.1104
	***	***	***	***	***		***		***	***
Child	-1.17	-1.7805	1.0831	2.0386	-1.2339	2.8694	-0.0203	-1.129	-1.1642	-2.138
	***	***	*	***	***	***	n.s	***	***	***
Camper	-0.163	0.6211	-0.475	-1.126			-0.538	0.7775		1.0179
	n.s.	**	n.s.	**			n.s.	**		**
Persons*Child			-0.3423							
			**							
Persons*Camper	0.33754		0.4255				0.4535			
	***		***				***			
Child*Camper			-1.1975				-1.374			
			***				***			
Theta	---	0.4635	---	---	1.8584	---	---	---	3.0938	---
Deviance	1323.7	210.65	---	---	---	---	---	---	---	---
AIC	**1670.728**	**820.44**	**1471.256**		**817.5458**		**1517.354**		**807.4923**	
LogVero	**-830.364**	**-405.222**	**-724.6**		**-402.77**		**-794.7**		**-395.7**	

FUENTE: Elaboración propia

Como se vio anteriormente, una de las consecuencias de trabajar con un modelo Poisson cuando éste no se ajusta correctamente es la subestimación de los errores estándar de los coeficientes estimados, como se muestra a continuación para la segunda aplicación:

Cuadro N° 36: Errores estándar de los modelos estimados

Variable Predictora	Poisson	Hurdle *Logit* Poisson	Poisson Inflado en Cero	NB2	NB2 inflado en cero	Hurdle *Logit* NB2
Intercepto	0.250	0.289	0.303	0.330	0.287	0.573
persons	0.078	0.089	0.092	0.114	0.111	0.155
child	0.081	0.188	0.433	0.185	0.271	0.312
camper	0.300	0.344	0.345	0.235	---	---
persons*camper	0.090	0.103	0.102	---	---	---
child*¨camper	---	0.218	0.224	---	---	---
persons*child	---	---	0.125	---	---	---

FUENTE: Elaboración propia

Para el caso de los modelos inflados en cero y *hurdle* sólo se muestran los coeficientes estimados del submodelo Poisson o binomial negativo según sea el caso. Se aprecia que los modelos de tipo Poisson estiman una mayor cantidad de coeficientes y/o los errores estándar asociados a éstos son menores que en los modelos NB2.

V. CONCLUSIONES

1. De acuerdo a la significancia obtenida en las pruebas de Bohning, Cameron y Trivedi y Dean y Lawless, es posible afirmar que las variables "Número de cigarros consumidos semanalmente" y "Número de peces capturados por pescador" presentan sobredispersión.

2. El modelo Poisson presentó, en ambas aplicaciones, los peores indicadores de bondad de ajuste. Entonces, debido al problema de sobredispersión en la variable respuesta, el modelo Poisson no es adecuado ante la preponderancia de ceros.

3. Los errores estándar en los modelos Poisson fueron subestimados, lo cual originó una inferencia inválida ya que un mayor número de variables predictoras fueron incluidas en esos modelos, es decir, mediante el modelo Poisson se estimó al menos tantos parámetros como el modelo NB2, con el modelo ZIP al menos tantos como el modelo ZIBN y con el modelo *hurdle* Poisson al menos tantos como el modelo hurdle NB2.

4. A pesar de lo señalado en el punto anterior, en la aplicación uno, los modelos Poisson inflado en cero y *hurdle logit* Poisson arrojaron una tasa de clasificación correcta más alta (63.15 y 70.3 por ciento respectivamente) que los modelos NB2 inflado en cero (43.2 por ciento) y *hurdle* NB2 (57.52 por ciento), debido a que éstos últimos no eran los modelos que se ajustaban mejor a los datos.

5. No es posible señalar un modelo como el mejor (o la única alternativa) ante la presencia de sobredispersión, ya que en ambas aplicaciones se obtuvo una terna distinta de los "*mejores modelos*", para la aplicación uno esta estuvo conformada por los modelos NB2, Poisson inflado en cero y *hurdle logit* Poisson, mientras que en la segunda aplicación estuvo conformada por los modelos binomial negativo: NB2, NB2 inflado en cero y *hurdle logit* NB2.

6. En cada terna de los mejores modelos ya presentados el ajuste era similar, entonces, si bien el criterio estadístico lleva a elegir el modelo más parsimonioso, que sería el NB2 para la primera aplicación y el modelo *hurdle logit* NB2 para la segunda, también podrían ser elegidos los modelos inflados en cero y los modelos *hurdle* dependiendo de la interpretación que se desee dar a los coeficientes estimados, ya que el modelo NB2 sólo estima la media de la variable respuesta, mientras que los modelos inflados en cero y *hurdle* estiman, además de la media de la variable respuesta, proporciones de ceros estructurales y ceros, respectivamente, que pueden ser de interés para el investigador

7. Para la aplicación de consumo de cigarros se tiene que los factores de riesgo de mayor incidencia en el consumo de cigarros son el consumo regular de bebidas alcohólicas, el entorno de compañeros que fuma, la asistencia a charlas de prevención sobre el consumo de cigarros y la edad del ingresante, en ese orden. Además, el consumo regular de bebidas alcohólicas es el factor determinante en la decisión del ingresante acerca de iniciarse o no en el consumo de cigarros (prevalencia de vida), así como en el consumo actual de éstos (prevalencia actual).

8. Para la segunda aplicación, sobre la tasa de pesca, las variables “Número de acompañantes” e “Ir en casa rodante al lago estatal”, son factores que incrementan la tasa de peces capturados por pescador, mientras que el número de niños que acuden al lago con el pescador origina lo opuesto. Además de ello, esta última variable también incrementa la chance de que el pescador decida no pescar (cero estructural) o que, cuando pesque, no logre capturar ningún pez (cero en un modelo *hurdle*)

VI. RECOMENDACIONES

La metodología propuesta no pretende ser la única y definitiva respecto al tema de modelos para datos de conteo, ya que existe una gran variedad de maneras de abordar el problema. Se plantean a continuación algunas recomendaciones y propuestas de investigación a futuro:

- Simular conjuntos de datos para determinar el modelo que mejor ajuste bajo criterios específicos como el porcentaje de ceros, asimetría, etc.
- Flexibilizar la relación media varianza en el modelo Poisson mediante un enfoque semiparamétrico (modelo Quassi Poisson) u otra función de densidad asociada a la tasa o promedio en vez de la distribución Gamma, por ejemplo la distribución de Lindley.
- Utilizar la distribución de Conway – Maxwell – Poisson o comúnmente conocida como Poisson generalizada, la cual añade a la distribución de Poisson un parámetro de dispersión que puede modelar el exceso o la escasez de ésta.
- Obtener modelos de regresión para datos de conteo desde el punto de vista bayesiano, asumiendo distribuciones a priori para los parámetros en estudio y hacer uso de técnicas de selección de variables vía el análisis de la densidad posterior de cada uno de los parámetros (coeficientes de regresión).
- Ahondar en mayor detalle acerca del muestreo y la obtención de un tamaño de muestra óptimo para modelos de regresión para datos de conteo.
- Utilizar conjuntos de datos que presenten una variable respuesta con infradispersión (lo opuesto a la sobredispersión) y ajustarlos a un modelo *hurdle*.

VII. REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. 2002. Categorical Data Analysis. 2 ed. Wiley Series in Probability and Statistics

Azen, R; Walker, CM. 2010. Categorical Data Analysis for the Behavioral and Social Sciences. Taylor & Francis Group.

Böhning, D. 1994. A Note on a Test for Poisson Overdispersion. Biometrika. 81(2), p. 418-419.

Böhning, D.; Dietz, E.; Schlattmann; P. 1997. Zero-Inflated Count Models and their Applications in Public Health and Social Science. Berlin.

Brazilian Symposium of Probability and Statistics (13, 1998, Caxambu). 2007. A Short Course for SINAPE 1998. Hinde, J., Demétrio, C.

Breslow, N. 1996. Generalized Linear Models. Checking Assumptions and Strengthening Conclusions. Statistica Applicata v. 8, p. 23-41.

Camacho, E. 2012. Una Versión Estocástica del Algoritmo EM. Tesis para obtener el título de licenciada en Ciencias Matemáticas. Universidad Centrooccidental “Lisandro Alvarado”, Barquimiseto, Venezuela.

Cameron, A.; Windmeijer, F. 1996. R-Squared Measures for Count Data Regression with Applications to Health-Care Utilization. Journal of Business & Economic Statistics. 14(2). p. 209-220.

Cameron, A.; Trivedi, P. 1998, Regression analysis of count data. Primera Edición. Cambridge University Press.

Cameron, A.; Trivedi, P. 1999. Essentials of Count Data Regression. Badi H. Baltagi ed. A companion to Theoretical Econometrics. p. 331-348

Castro Salcedo, L. 2011. Consumo de tabaco en el Perú. Perú. Consultado 06 dic. 2012. Disponible en http://luz-mariela-castro-salcedo.suite101.net/consumo-de-tabaco-en-el-peru-a35851

Casualty Actuarial Society Forum (2007, Virginia). 2007. Handling Overdispersion with Negative Binomial and Generalized Poisson Regression Models. Ismail, N.; Aziz, A. p. 103-158.

Cattaneo, M. Factores de Riesgo al Consumo de Sustancias Psicoactivas. Consultado 07 ene. 2013. Disponible en http://www.fiso-web.org/imagenes/publicaciones/archivos/2540.pdf

Centro de Información y Educación para la Prevención del Abuso de Drogas (CEDRO), Centros para el Control y la Prevención de Enfermedades, Organización Panamericana de la Salud, Organización Mundial de la Salud. 2007. Encuesta Mundial de Profesionales de la Salud (GHPS): Uso de Tabaco en estudiantes de Tercer Año de Medicina, Enfermería y Farmacia. Perú.

Dean, C; Lawless, JF. 1989. Detecting Overdispersion in Poisson Regression Models. Journal of the American Statistical Association. 84 (406). p. 467 – 472.

Dobson, A. 2002. An introduction to generalized linear models. 2 ed. Chapman and Hall/CRC

Fox, J. 1997. Applied Regression Analysis, Linear Models, and Related Methods. SAGE Publications. p. 570

Gao, K.; Khoshgoftaar, T. 2007. A Comprehensive Empricial Study of Count Models for Software Fault Prediction, IEEE Transactions on reliability. 56(2). p. 223-236.

Garay, A.; Hashimoto, E.; Ortega, E.; Lachos, V. 2010. On Estimation and Influence Diagnostics of Zero-Inflated Negative Regression Models. Computational Statistics & Data Analysis. 55(3). p. 1304-1318.

Greene, W. 1994. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. NYU Working Paper número EC-94-10.

Gurmu, S.; Trivedi, P. 1996. Excess Zeros in Count Models for Recreational Trips, Journal of Business & Economic Statistics. 14(4). p. 469-477.

Hardin, JW; Hilbe, JM. 2012. Generalized Estimating Equations. 2 ed. CRC Press.

Hilbe, JM. 2011. Negative Binomial Regression. 2 ed. Cambridge University Press

Institute for digital research and education. SPSS Data Analysis Examples. Consultada 06 abr. 2013. Disponible en http://www.ats.ucla.edu/stat/spss/dae/neg_binom.htm

Institue for digital research and education. R Data Analysis Examples. Consultada 01 set. 2013. Disponible en http://www.ats.ucla.edu/stat/r/dae/zinbreg.htm

Jornada "Investigaciones en Facultad" de Ciencias Económicas y Estadística (2010, 14, Argentina). 2010. Modelos Alternativos para el análisis de datos de conteo con exceso de ceros. Hachuel, L.; Boggio, G.; Harvey, G.

Journal of Tropical Pediatrics. Research Methods II. Chapter 13: Poisson regression analysis. Consultado 16 mar. 2013. Disponible en:
http://www.oxfordjournals.org/our_journals/tropej/online/ma_chap13.pdf

Levine, N; Lord, D; Park, BJ. 2010. CrimeStat Version 3.3 Update Notes: Part2: Regression Modelling. Disponible en: http://www.icpsr.umich.edu/CrimeStat/files/CrimeStat3.3updatenotesPartII.pdf

Miller, J.; Miller, D. 2008. No Zero Left Behind: Comparing the Fit for Zero-Inflation Models as a Function of Skew and Proportion of Zeros. Consultada 15 abr. 2013. Disponible en http://interstat.statjournals.net/YEAR/2008/articles/0810011.pdf

Moghimbeigi, A.; Eshraghian, M.; Mohammad, K.; Nourijelyani, K.; Husseini, M. 2009. Determinants Numbers of Cigarette Smoked with Iranian Adolescents: A Multilevel Zero Inflated Poisson Regression Model, Iranian Journal of Public Health. 38(4). p. 91-96

Moon, T. 1996. The Expectation – Maximization Algorithm, IEEE Signal Processing Magazine. 13(6). p. 47-60.

Morel, JG; Neerchal, NK. Overdispersion Models in SAS. 2012. SAS Institute Inc.

Orme, JG; Orme, TC. 2009. Multiple Regression with Discrete Dependent Variables. Oxford University Press.

Proceedings if the Annual Meeting of American Statistical Association. (2001, 08). Variable Selection for Poisson Regression Model. Famoye, F.; Rothe, DE.

Proceedings International Symposium on Software Reliability Engineering (2001, 12, Hong Kong). 2001. An Application of Zero-Inflated Poisson Regression for Software Fault Prediction. Khoshgoftaar, TM; Szabo, RM. p. 66-73.

Rabines Juárez, A. 2002. Factores de riesgo para el consumo de tabaco en una población de adolescentes escolarizados. Tesis para optar por el título de Médico-Cirujano. Universidad Nacional Mayor de San Marcos. Lima, Perú.

Ramaswamy, V; Anderson, EW; DeSarbo, WS. 1994. A Disaggregate Negative Binomial Regression Procedure for Count Data Analysis. Management Science. 40 (3). p. 405-417.

SAS Institute Inc. 2008. Count Models in SAS. SAS Global Forum. Disponible en http://www2.sas.com/proceedings/forum2008/371-2008.pdf

Signorini, D. 1991. Sample Size for Poisson Regression. Biometrika. 78(2). p. 446-450.

Stata Corporation. 2003. From the help desk: hurdle models. The Stata Journal. 3(2). p. 178-184. Consultado 10 feb. 2013. Disponible en http://ageconsearch.umn.edu/bitstream/116071/2/sjart_st0040.pdf

Sumathi, K.; Aruna, R. 2006. On estimation and tests for Zero Inflated Regression Models. India. Consultado 12 nov. 2012. Disponible en http://interstat.statjournals.net/YEAR/2009/articles/0908004.pdf

Vuong, QH. 1989. Likelihood Ratio Test for Model Selection and Non-Nested Hypotheses. Econometrica. 57(2). p.307-333.

Winkelmann, R. 2008. Econometric Analysis of Count Data. 5 ed. Editorial Springer.

Zeileis, A.; Kleiber, C.; Jackman S. 2008. Regression Models for Count Data in R. Consultada 19 ago 2013. Disponible en http://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

Zewotir, T.; Ramroop, S. 2009. Application of Negative Binomial Regression for Assessing Public Awareness of the Health Effects of Nicotine and Cigarettes, African Safety Promotion: A Journal of Injury and Violence Prevention. 7(1). p. 14-29.

Zorn, CJ. 1996. Evaluating Zero-Inflated and Hurdle Poisson Specifications. Midwest Political Science Asociation. 18(20). p. 1-16.

ANEXOS

ANEXO 1: FUNCIONES DE LOG VEROSIMILITUD EN LOS MODELOS DE REGRESIÓN PARA DATOS DE CONTEO

Modelo de regresión Poisson

- Función de log-verosimilitud

$$L\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}\right) = \prod_{i=1}^{n} \frac{\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\left(\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)^{y_i}}{y_i\,!}$$

- Gradiente

$$\frac{\partial \ln L\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}\right)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left(y_i - \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\mathbf{x}_i' = \sum_{i=1}^{n}\left(y_i - \mu_i\right)\mathbf{x}_i'$$

- Hessiana

$$\frac{\partial^2 \ln L\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n}\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\mathbf{x}_i\mathbf{x}_i'$$

Modelo de regresión NB-C

- Función de log-verosimilitud

$$L(\boldsymbol{\beta},\alpha \mid \mathbf{y},\mathbf{X}) = \prod_{i=1}^{n} \exp\left(\ln\Gamma\left(y_i+\alpha^{-1}\right) - \ln\Gamma\left(y_i+1\right) - \ln\Gamma\left(\alpha^{-1}\right) + y_i\mathbf{x}_i'\boldsymbol{\beta} + \alpha^{-1}\ln\left(1-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right)$$

- Gradiente

$$\frac{\partial \ln L(\boldsymbol{\beta},\alpha \mid \mathbf{y},\mathbf{X})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\left(y_i - \frac{\alpha^{-1}}{\exp\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)-1}\right)\mathbf{x}_i' = \sum_{i=1}^{n}\left(y_i-\mu_i\right)\mathbf{x}_i'$$

$$\frac{\partial \ln L(\boldsymbol{\beta},\alpha \mid \mathbf{y},\mathbf{X})}{\partial \alpha} = \sum_{i=1}^{n}\left(-\frac{1}{\alpha^2}\left[\ln\left(1-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right) + \Psi\left(y_i+\frac{1}{\alpha}\right) - \Psi\left(\frac{1}{\alpha}\right)\right]\right)$$

Donde $\Psi(\bullet)$ se define como la función digamma:

$$\Psi(z) = \frac{\ln\Gamma\left(z+\delta\right) - \ln\Gamma\left(z-\delta\right)}{2\delta}$$

- Hessiana

$$\begin{bmatrix} \sum_{i=1}^{n}\left[-\left(\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\left(\mathbf{x}_i'\right)\left(\mathbf{x}_i'\right)}{\alpha\left(\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)-1\right)^2}\right)\right] & \sum_{i=1}^{n}\left(\frac{\mathbf{x}_i'}{\alpha^2\left(\exp\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)-1\right)}\right) \\ \sum_{i=1}^{n}\left(\frac{\mathbf{x}_i'}{\alpha^2\left(\exp\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)-1\right)}\right) & \sum_{i=1}^{n}\left(\frac{2\alpha\ln\left(1-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)+2\alpha\Psi\left(y_i+\frac{1}{\alpha}\right)+\Psi'\left(y_i+\frac{1}{\alpha}\right)-2\alpha\Psi\left(\frac{1}{\alpha}\right)-\Psi'\left(\frac{1}{\alpha}\right)}{\alpha^4}\right) \end{bmatrix}$$

Modelo de regresión NB2

- Función de log-verosimilitud

$$L(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X})=\prod_{i=1}^{n}\exp\left(\ln\Gamma\left(y_i+\alpha^{-1}\right)-\ln\Gamma\left(y_i+1\right)-\ln\Gamma\left(\alpha^{-1}\right)+y_i\ln\left(\frac{\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)-\alpha^{-1}\ln\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right)$$

- Gradiente

$$\frac{\partial\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\left[\left(\frac{y_i-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}\right)\mathbf{x}_i'\right]$$

$$\frac{\partial\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\alpha}=\sum_{i=1}^{n}\left[\frac{1}{\alpha^2}\left[-\Psi\left(y_i+\alpha^{-1}\right)+\Psi\left(\alpha^{-1}\right)+\ln\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right]+\frac{\left(y_i-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}{\alpha\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}\right]$$

- Hessiana

$$\begin{bmatrix}\dfrac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & \dfrac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}\partial\alpha}\\ \dfrac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}\partial\alpha} & \dfrac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\alpha^2}\end{bmatrix}$$, donde:

$$\frac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}=\sum_{i=1}^{n}\left[\left(\frac{-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\left(1+\alpha y_i\right)}{\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)^2}\right)\left(\mathbf{x}_i\mathbf{x}_i'\right)\right]$$

$$\frac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\alpha^2}=$$

$$=\sum_{i=1}^{n}\left[\frac{1}{\alpha^3}\left(M+\frac{\left(y_i-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\left(\alpha\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)+\alpha^2\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)-\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)}{\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)^2}+2\ln\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right)\right]$$

Donde: $M=2\Psi\left(y_i+\frac{1}{\alpha}\right)-\frac{1}{\alpha}\Psi'\left(y_i+\frac{1}{\alpha}\right)-2\Psi\left(\frac{1}{\alpha}\right)-\frac{1}{\alpha}\Psi'\left(\frac{1}{\alpha}\right)$

$$\frac{\partial^2\ln L\left(\boldsymbol{\beta},\alpha\,|\,\mathbf{y},\mathbf{X}\right)}{\partial\boldsymbol{\beta}\partial\alpha}=-\sum_{i=1}^{n}\left[\frac{\left(y_i-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\mathbf{x}_i'}{\left(1+\alpha\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)^2}\right]$$

Modelo de regresión Poisson inflado en cero

- Función de log verosimilitud

$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{x}) = A + B + C$, donde:

$$A = \sum_{y_i = 0} \ln\left(\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right) + \exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right)$$

$$B = \sum_{y_i > 0} \left[-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + y_i \mathbf{x}_i'\boldsymbol{\beta} - \ln y_i !\right]$$

$$C = \sum_{i=1}^{n} \ln\left(\frac{1}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}\right)$$

Modelo de regresión NB2 inflado en cero

- Función de log verosimilitud

$\ln L(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{x}) = A + B + C$, donde:

$$A = \sum_{y_i = 0} \ln\left(\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right) + \left(\frac{\phi}{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + \phi}\right)^{\phi}\right)$$

$$B = \sum_{y_i > 0} \left[\ln\binom{y_i + \phi - 1}{y_i} + y_i \ln\left(\frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + \phi}\right) + \phi \ln\left(\frac{\phi}{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + \phi}\right)\right]$$

$$C = \sum_{i=1}^{n} \ln\left(\frac{1}{1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)}\right)$$

Modelo de regresión *hurdle* Poisson

- Función de log verosimilitud: $\ln L\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \mathbf{Z},\mathbf{X},\mathbf{y}\right) = l_1\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) + l_2\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$,donde:

  Para el modelo *hurdle logit* Poisson:

$$l_1\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_1} \mathbf{z}_i'\boldsymbol{\gamma} - \sum_{i\in\Omega} \ln\left(1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)\right)$$

$$l_2\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_1} \left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln\left(1-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right) - \ln y_i!\right)$$

  En caso se plantee el modelo *hurdle* Poisson - Poisson:

$$l_1\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_0} \ln\left(1-\exp\left(-\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)\right)\right) - \sum_{i\in\Omega_1} \exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)$$

$$l_2\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_1} \left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) + y_i\mathbf{x}_i'\boldsymbol{\beta} - \ln\left(1-\exp\left(-\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right)\right) - \ln y_i!\right)$$

Modelo de regresión *hurdle* NB2

- Función de log verosimilitud: $\ln L\left(\boldsymbol{\beta},\boldsymbol{\gamma} \mid \mathbf{Z},\mathbf{X},\mathbf{y}\right) = l_1\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) + l_2\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right)$,donde:

$$l_1\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_1} \mathbf{z}_i'\boldsymbol{\gamma} - \sum_{i\in\Omega} \ln\left(1+\exp\left(\mathbf{z}_i'\boldsymbol{\gamma}\right)\right)$$

$$l_2\left(\boldsymbol{\beta},\boldsymbol{\gamma}\right) = \sum_{i\in\Omega_1} \left[\ln\binom{\alpha^{-1}+y_i-1}{y_i} + y_i\ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) + \alpha^{-1}\ln\left(\frac{1}{1+\alpha\mu_i}\right) - \ln\left(1-\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\right)\right]$$

ANEXO 2: APLICACIÓN UNO: MODELO DE REGRESIÓN POISSON

- **Selección de variables**

Paso	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Intercepto	-1.055	-2.3389	-2.8332	-2.4197	-2.7284	-7.8736	-7.6029	-1.5832	-1.9424	0.28447	**-4.5704**
	***	***	***	***	***	***	***	***	***	n.s.	***
Bebidas	2.074	1.6528	2.9667	2.8927	3.3154	3.0427	2.31404	2.30951	2.29038	2.3272	**2.3313**
	***	****	***	***	***	***	***	***	***	***	***
FumaComp		1.8693	2.4227	2.5533	2.4535	2.4743	2.09868	- 4.896	- 4.887	-3.799	**1.9804**
		****	***	***	***	***	***	.	n.s.	n.s.	***
Charlas				-1.114	-0.2198	-0.3739	-0.3362	-0.477	-0.4318	-0.586	**-0.495**
				***	n.s.	n.s.	n.s.	*	.	*	*
Edad						0.2746	0.27458	-0.0353	-0.029	-0.139	**0.1159**
						***	***	n.s.	n.s.	n.s.	.
FumaCasa									0.57488	-11.36	**-11.77**
									***	***	***
Beb*Fcomp			-1.422	-1.306	-1.084	-0.8232					
			**	*	*	n.s.					
Beb*Charla					-1.6812	-1.4689	-1.5292	-1.4055	-1.5696	-1.606	**-1.681**
					***	***	***	***	***	***	***
Edad*Fco								0.36505	0.36371	0.3056	
								*	*	.	
Edad*Fcasa										0.6197	**0.6418**
										***	***
Deviance	686.18	614.12	607.45	551.60	523.33	483.26	485.26	477.72	464.09	418.71	**422.64**
AIC	815.92	745.87	741.19	687.35	661.08	623.01	623.27	617.47	605.84	562.46	**564.38**
LogVero	-405.96	-369.94	-366.6	-338.67	-324.54	-304.51	-305.64	-301.73	-294.92	-272.23	-274.19

- **Salidas en R del modelo final**

```
Call:
glm(formula = CigarrosSemanales ~ Bebidas * Charlas + FumaCompañeros + Edad
* FumaCasa, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6218  -0.9094  -0.4083  -0.3009   7.3418

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.57044    1.16275  -3.931 8.47e-05 ***
BebidasSí             2.33126    0.20476  11.385  < 2e-16 ***
CharlasSí            -0.49467    0.24364  -2.030   0.0423 *
FumaCompañerosSí      1.98044    0.27627   7.168 7.59e-13 ***
Edad                  0.11587    0.05926   1.955   0.0505 .
FumaCasaSí          -11.76830    1.83226  -6.423 1.34e-10 ***
BebidasSí:CharlasSí  -1.68052    0.33322  -5.043 4.58e-07 ***
Edad:FumaCasaSí       0.64179    0.09412   6.819 9.18e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 871.21  on 265  degrees of freedom
Residual deviance: 422.64  on 258  degrees of freedom
AIC: 564.38

Number of Fisher Scoring iterations: 6
```

- **Expresiones para interpretar interacciones**

Al estimar la tasa semanal de consumo de cigarros para un ingresante cuando:

Consume no consume bebidas alcohólicas regularmente $(X_1 = 0)$ y asiste a charlas $(X_2 = 1)$:

$$\hat{\mu}_i = \exp(-5.065 + 1.98X_{3i} + 0.116X_{4i} - 11.768X_{5i} + 0.6418X_{4i}X_{5i})$$

Consume no consume bebidas alcohólicas regularmente $(X_1 = 0)$ ni asiste a charlas $(X_2 = 0)$:

$$\hat{\mu}_i = \exp(-4.57 + 1.98X_{3i} + 0.116X_{4i} - 11.768X_{5i} + 0.6418X_{4i}X_{5i})$$

Consume consume bebidas alcohólicas regularmente $(X_1 = 1)$ y asiste a charlas $(X_2 = 1)$:

$$\hat{\mu}_i = \exp(-2.734 + 1.98X_{3i} + 0.116X_{4i} - 11.768X_{5i} - 1.681 + 0.6418X_{4i}X_{5i})$$

Consume consume bebidas alcohólicas regularmente $(X_1 = 1)$ y no asiste a charlas $(X_2 = 0)$:

$$\hat{\mu}_i = \exp(-2.239 + 1.98X_{3i} + 0.116X_{4i} - 11.768X_{5i} + 0.6418X_{4i}X_{5i})$$

Para la variable X_4: *Edad* , cuando no fuman en casa del ingresante $X_5 = 0$:

$$\hat{\mu}_i = \exp(-4.57 + 2.331X_{1i} - 0.495X_{2i} + 1.98X_{3i} + 0.116X_{4i} - 1.681X_{1i}X_{2i})$$

Luego para aquellos ingresantes cuyos familiares sí fuman, es decir cuando $X_5 = 1$:

$$\hat{\mu}_i = \exp(-16.338 + 2.331X_{1i} - 0.495X_{2i} + 1.98X_{3i} + 0.7578X_{4i} - 1.681X_{1i}X_{2i})$$

- **Salidas en R de los tests de sobredispersión**

Test de Bohning

```
obs<-CigarrosSemanales
exp<-predict(mod_pois,type="response")
epi.bohning(obs,exp)

  test.statistic p.value
1       26.52687       0
```

Chi Cuadrado de Pearson

```
res.p<-residuals(mod_pois,type="pearson")
chicuadrado<-sum(res.p^2)
chicuadrado
[1] 831.4383
chicuadrado/(266-8)
[1] 3.222629
```

Test de Dean y Lawless

```
T<-sum((obs-exp)^2-exp)/sqrt(2*sum(exp^2))
T
[1] 18.4889
1-pnorm(T)
[1] 0
```

Test de Cameron y Trivedi (1985)

```
dispersiontest(mod_pois)

        Overdispersion test
data:  mod_pois
z = 2.1783, p-value = 0.01469
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  3.200678
```

Test de Cameron y Trivedi (1986)

```
ty<-(obs-exp)^2
tx<-exp
summary(lm(ty~tx-1))

Call:
lm(formula = ty ~ tx - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-42.833  -2.846  -0.702  -0.306 154.290

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
tx   6.0644     0.4841   12.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.46 on 265 degrees of freedom
Multiple R-squared: 0.3719,     Adjusted R-squared: 0.3695
F-statistic: 156.9 on 1 and 265 DF,  p-value: < 2.2e-16
```

Test de Cameron y Trivedi (1990)

```
ty<-(obs-exp)^2-exp
tx<-exp^2
summary(lm(ty~tx-1))

Call:
lm(formula = ty ~ tx - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-57.717  -0.364  -0.090  -0.044 163.516

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
tx  0.53627    0.05384    9.96   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.66 on 265 degrees of freedom
Multiple R-squared: 0.2724,     Adjusted R-squared: 0.2696
F-statistic: 99.21 on 1 and 265 DF,  p-value: < 2.2e-16
```

ANEXO 3: APLICACIÓN UNO: MODELO DE REGRESIÓN NB2

- **Selección de variables**

Paso	1	2	3	**4**
Intercepto	-2.1282	-2.6236	-7.9942	**-8.2606**
	***	***	***	*******
FumaComp	2.3341	2.1374	1.9031	**1.6443**
	***	***	***	*******
Bebidas		1.8951	1.8938	**2.1794**
		***	***	*******
Edad			0.2935	**0.3278**
			**	******
Charlas				**-1.0708**
				******
Theta	0.1137	0.1514	0.1751	**0.1903**
Deviance	108.31	110.18	113.05	**111.45**
AIC	417.27	402.33	398.05	**393.35**
LogVero	-205.633	-197.163	-194.027	**-190.675**

- **Salidas en R del modelo final**

```
Call:
glm.nb(formula = CigarrosSemanales ~ Bebidas + FumaCompañeros +  Edad +
Charlas, init.theta = 0.1902645053, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1510  -0.6548  -0.3915  -0.2097   2.0399

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.2606     1.9779  -4.177 2.96e-05 ***
BebidasSí          1.6443     0.4592   3.581 0.000343 ***
FumaCompañerosSí   2.1794     0.4683   4.654 3.26e-06 ***
Edad               0.3278     0.1012   3.239 0.001198 **
CharlasSí         -1.0708     0.4026  -2.660 0.007820 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative binomial(0.1903) family taken to be 1)

    Null deviance: 185.19  on 265  degrees of freedom
Residual deviance: 111.45  on 261  degrees of freedom
AIC: 393.35

Number of Fisher Scoring iterations: 1


              Theta:  0.1903
          Std. Err.:  0.0444

 2 x log-likelihood:  -381.3510
```

ANEXO 4: APLICACIÓN UNO: MODELO DE REGRESIÓN POISSON INFLADO EN CERO

- **Selección de variables**

Paso	**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-5.65	1.893	-7.479	1.813	-8.962	1.635	-2.561	1.843	-2.489	2.84	-7.36	2.778	-6.32	2.8	-6.24	2.477	-6.07	1.53
	***	***	***	***	***	***	n.s.	***	n.s.	***	***	***	***	***	***	***	***	***
Bebidas	0.704	-1.904	6.394	-1.778	6.965	-1.681	1.585	-2.26	1.575	-2	1.46	-1.97	1.47	-2	1.355	-1.96	1.48	-1.95
	***	***	**	***	**	***	***	***	***	***	***	***	***	***	***	***	***	***
Edad	0.352		0.444		0.536		0.178		0.175		0.449		0.386		0.339		0.286	
	***		***		***		.		n.s.		***		***		***		***	
Charlas					-0.929		-5.518		-5.56		-0.1		-0.06		-0.18		0.059	
					***		*		*		n.s.		n.s.		n.s.		n.s.	
Depart. Nacim.											-0.75		-0.69		-0.518			
											***		**		*			
Fuman casa													0.444		0.649		0.815	
													*		***		***	
Fuman Comp.										-1.59		-1.57		-1.55	0.895	-1.19	1.478	
										***		***		***	*	*	***	
Edad * Bebidas			-0.294		-0.316													
			**		*													
Charlas * Edad							0.3013		0.3032									
							*		*									
Bebidas*Charlas							-2.121		-2.14		-1.9		-2.05		-1.97		-2.24	
							***		***		***		***		***		***	
AIC	**488.5758**		**483.8537**		**453.6669**		**413.6513**		**401.8342**		**394.6357**		**390.2808**		**384.4778**		**390.5283**	
Log Vero	**-239.3**		**-235.9**		**-219.8**		**-198.8**		**-191.9**		**-188.3**		**-185.1**		**-181.2**		**-186.3**	

- **Salidas en R del modelo final**

```
Call:
zeroinfl(formula = CigarrosSemanales ~ Bebidas * Charlas + Edad + Charlas
+ FumaCasa + FumaCompañeros | Bebidas)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.1721 -0.3771 -0.2841 -0.2106  7.6110

Count model coefficients (poisson with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.07040    1.13577  -5.345 9.06e-08 ***
BebidasSí            1.48495    0.21465   6.918 4.58e-12 ***
CharlasSí            0.05863    0.26083   0.225    0.822
Edad                 0.28550    0.05987   4.769 1.85e-06 ***
FumaCasaSí           0.81487    0.17369   4.692 2.71e-06 ***
FumaCompañerosSí     1.47789    0.34809   4.246 2.18e-05 ***
BebidasSí:CharlasSí -2.23966    0.36252  -6.178 6.49e-10 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5315     0.2533   6.045 1.49e-09 ***
BebidasSí    -1.9536     0.4668  -4.185 2.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 17
Log-likelihood: -186.3 on 9 Df
```

- **Expresiones para interpretar las interacciones**

Al estimar la tasa semanal de consumo de cigarros para un ingresante cuando:

Consume no consume bebidas alcohólicas regularmente $\left(X_1 = 0\right)$ y asiste a charlas $\left(X_2 = 1\right)$:

$$\hat{\mu}_i = \exp\left(-6.011 + 1.478X_{3i} + 0.286X_{4i} + 0.815X_{5i}\right)$$

Consume no consume bebidas alcohólicas regularmente $\left(X_1 = 0\right)$ ni asiste a charlas $\left(X_2 = 0\right)$:

$$\hat{\mu}_i = \exp\left(-6.07 + 1.478X_{3i} + 0.286X_{4i} + 0.815X_{5i}\right)$$

Consume consume bebidas alcohólicas regularmente $\left(X_1 = 1\right)$ y asiste a charlas $\left(X_2 = 1\right)$:

$$\hat{\mu}_i = \exp\left(-6.771 + 1.478X_{3i} + 0.286X_{4i} + 0.815X_{5i}\right)$$

Consume consume bebidas alcohólicas regularmente $\left(X_1 = 1\right)$ y no asiste a charlas $\left(X_2 = 0\right)$:

$$\hat{\mu}_i = \exp\left(-4.59 + 1.478X_{3i} + 0.286X_{4i} + 0.815X_{5i}\right)$$

ANEXO 5: APLICACIÓN UNO: MODELO DE REGRESIÓN NB2 INFLADO EN CERO

- **Selección de variables**

Paso	**1**		**2**		**3**		**4**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-2.128	-8.733	-2.6235	-9.106	-7.9944	-8.399	-8.2605	-8.291
	***		***		***		***	
Fuman Comp.	2.3341		2.1374		1.9031		2.1794	
	***		***		***		***	
Bebidas			1.895		1.8938		1.6422	
			***		***		***	
Edad					0.2935		0.3278	
					*		**	
Charlas							-1.0708	
							*	
Theta	-2.174	---	-1.8877	---	-1.7422	---	-1.6590	---
	***	---	***	---	***	---	***	---
AIC	**419.2664**		**404.3263**		**400.0549**		**395.3509**	
Log Vero	**-205.6**		**-197.2**		**-194**		**-190.7**	

- **Salidas en R del modelo final**

```
Call:
zeroinfl(formula = CigarrosSemanales ~ FumaCompañeros + Bebidas + Edad +
Charlas | 1, dist = c("negbin"))

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4294 -0.3586 -0.2512 -0.1441  6.3593

Count model coefficients (negbin with log link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.2605     2.1811  -3.787 0.000152 ***
FumaCompañerosSí   2.1794     0.4722   4.616 3.92e-06 ***
BebidasSí          1.6442     0.4595   3.578 0.000346 ***
Edad               0.3278     0.1171   2.800 0.005118 **
CharlasSí         -1.0708     0.4225  -2.535 0.011252 *
Log(theta)        -1.6590     0.2380  -6.970 3.17e-12 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.291    119.552  -0.069    0.945
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.1903
Number of iterations in BFGS optimization: 85
Log-likelihood: -190.7 on 7 Df
```

ANEXO 6: APLICACIÓN UNO: MODELO DE REGRESIÓN *HURDLE* POISSON

- **Selección de variables del modelo *hurdle logit* Poison**

Paso	**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	1.1	-2.09	-5.288	-2.09	-7.157	-2.09	-8.71	-2.09	-6.636	-2.09	-6.22	-2.09	-6.67	-3.02	-5.59	-3.02
	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Bebidas	0.635	2.038	0.687	2.038	6.17	2.038	7.181	2.038	3.671	2.038	1.665	2.038	1.621	1.705	1.553	1.705
	***	***	***	***	**	***	**	***		***	***	***	***	***	***	***
Edad			0.334		0.428		0.523		0.389		0.368		0.36		0.27	
			***		***		***		***		***		***		***	
Charlas							-0.887		0.482		0.507		0.405		0.256	
							***		.		.					
Fuman Comp.													0.707	1.509	1.217	1.509
													*	***	***	***
Fuman casa															0.808	
															***	
Edad * Bebidas					-0.284		-0.328		0.465							
					*		**									
Bebidas*Charlas									-2.369		-2.436		-2.33		-2.501	
									***		***		***		***	
AIC	**525.04**		**491.253**		**486.8628**		**460.2688**		**421.6517**		**420.183**		**405.2389**		**388.9188**	
Log Vero	**-258.52**		**-240.6265**		**-237.4314**		**-223.1344**		**-202.8258**		**-203.0915**		**-193.642**		**-184.4594**	

- **Salidas en R del modelo *hurdle logit* Poisson final**

```
Call:
hurdle(formula = CigarrosSemanales ~ Edad + Bebidas + Bebidas * Charlas +
FumaCompañeros + FumaCasa | Bebidas + FumaCompañeros)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.0496 -0.4052 -0.2029 -0.1959  7.2197

Count model coefficients (truncated poisson with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.5899     1.1597  -4.820 1.43e-06 ***
Edad                  0.2696     0.0600   4.493 7.03e-06 ***
BebidasSí             1.5533     0.2308   6.731 1.68e-11 ***
CharlasSí             0.2555     0.2778   0.919 0.357868
FumaCompañerosSí      1.2166     0.3606   3.374 0.000741 ***
FumaCasaSí            0.8075     0.1811   4.460 8.20e-06 ***
BebidasSí:CharlasSí  -2.5015     0.3937  -6.354 2.10e-10 ***
Zero hurdle model coefficients (binomial with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.0229     0.4009  -7.540 4.71e-14 ***
BebidasSí          1.7048     0.3999   4.263 2.02e-05 ***
FumaCompañerosSí   1.5089     0.4483   3.366 0.000763 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 17
Log-likelihood: -184.5 on 10 Df
```

- **Expresiones para interpretar las interacciones**

Al estimar la tasa semanal de consumo de cigarros para un ingresante cuando:

Consume no consume bebidas alcohólicas regularmente $\left(X_1 = 0\right)$ y asiste a charlas $\left(X_2 = 1\right)$:

$$\hat{\mu}_i = \exp\left(-5.3339 + 1.22X_{3i} + 0.27X_{4i} + 0.808X_{5i}\right)$$

Consume no consume bebidas alcohólicas regularmente $\left(X_1 = 0\right)$ ni asiste a charlas $\left(X_2 = 0\right)$:

$$\hat{\mu}_i = \exp\left(-5.5899 + 1.22X_{3i} + 0.27X_{4i} + 0.808X_{5i}\right)$$

Consume consume bebidas alcohólicas regularmente $\left(X_1 = 1\right)$ y asiste a charlas $\left(X_2 = 1\right)$:

$$\hat{\mu}_i = \exp\left(-6.2839 + 1.22X_{3i} + 0.27X_{4i} + 0.808X_{5i}\right)$$

Consume consume bebidas alcohólicas regularmente $\left(X_1 = 1\right)$ y no asiste a charlas $\left(X_2 = 0\right)$:

$$\hat{\mu}_i = \exp\left(-4.0399 + 1.22X_{3i} + 0.27X_{4i} + 0.808X_{5i}\right)$$

ANEXO 7: APLICACIÓN UNO: MODELO DE REGRESIÓN *HURDLE* NB2

- **Selección de variables del modelo *hurdle logit* NB2**

Paso	**1**		**2**		**3**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-6.294	-2.089	-6.294	-3.023	-6.294	-2.6986
		***		***		***
Bebidas		2.0381		1.7048		1.7787
		***		***		***
Fuman Comp.				1.5089		1.6576
				***		***
Charlas						-0.8373
						*
Log(Theta)	-8.626		-8.626		-8.626	
AIC	**413.8651**		**402.2991**		**399.2992**	
Log vero.	**-202.9**		**-196.1**		**-193.6**	

- **Salidas en R del modelo *hurdle logit* NB2 final**

```
Call:
hurdle(formula = CigarrosSemanales ~ 1 | Bebidas + FumaCompañeros, dist =
c("negbin"))
Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5091 -0.2697 -0.1332 -0.1332  6.7666

Count model coefficients (truncated negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.294     83.187  -0.076    0.940
Log(theta)    -8.626     83.216  -0.104    0.917
Zero hurdle model coefficients (binomial with logit link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.0229     0.4009  -7.540 4.71e-14 ***
BebidasSí          1.7048     0.3999   4.263 2.02e-05 ***
FumaCompañerosSí   1.5089     0.4483   3.366 0.000763 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 2e-04
Number of iterations in BFGS optimization: 1413
Log-likelihood: -196.1 on 5 Df
```

ANEXO 8: APLICACIÓN UNO: COMPARACIÓN DE MODELOS: SALIDAS DE R

Comparación del modelo Poisson con el NB2:

```
> vuong(mod_pois,mod_nb2)
Vuong Non-Nested Hypothesis Test-Statistic: -2.77014
(test-statistic is asymptotically distributed N(0,1) under the null that
the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.002801609
```

Comparación del modelo Poisson con el modelo Poisson inflado en cero

```
> vuong(mod_pois,mod_zip)
Vuong Non-Nested Hypothesis Test-Statistic: -3.174027
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.0007516977
```

Comparación del modelo Poisson con el modelo NB2 inflado en cero:

```
> vuong(mod_pois,mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.770136
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.002801644
```

Comparación del modelo Poisson con el modelo hurdle *logit* Poisson:

```
> vuong(mod_pois,mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: -3.25967
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.00055771
```

Comparación del modelo Poisson con modelo hurdle *logit* NB2

```
> vuong(mod_pois,mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.528684
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.005724548
```

Comparación del modelo NB2 con el modelo Poisson inflado en cero

```
> vuong(mod_nb2,mod_zip)
Vuong Non-Nested Hypothesis Test-Statistic: -0.5523
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.2903714
```

Comparación del modelo NB2 con el modelo NB2 inflado en cero

```
> vuong(mod_nb2,mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: 0.3283711
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.3713155
```

Comparación del modelo NB2 con el modelo hurdle *logit* Poisson

```
> vuong(mod_nb2,mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: -0.7259296
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.233941
```

Comparación del modelo NB2 con el modelo hurdle *logit* NB2

```
> vuong(mod_nb2,mod_hnb2)
Vuong Non-Nested Hypothesis Test-Statistic: 1.264891
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.1029552
```

Comparación del modelo Poisson inflado en cero con el modelo NB2 inflado en cero

```
> vuong(mod_zip,mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: 0.5523108
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.2903677
```

Comparación del modelo Poisson inflado en cero con el modelo hurdle *logit* Poisson

```
> vuong(mod_zip,mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: -0.5978622
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.2749659
```

Comparación del modelo Poisson inflado en cero con el modelo hurdle *logit* NB2

```
> vuong(mod_zip,mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: 1.161218
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.1227767
```

Comparación del modelo NB2 inflado en cero con el modelo hurdle *logit* Poisson

```
> vuong(mod_zibn,mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: -0.7259386
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.2339382
```

Comparación del modelo NB2 inflado en cero con el modelo hurdle *logit* NB2

```
> vuong(mod_zibn,mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: 1.305073
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.09593402
```

Comparación del modelo *hurdle logit* Poisson con el modelo hurdle *logit* NB2

```
> vuong(mod_hp,mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: 1.389336
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.08236524
```

ANEXO 9: APLICACIÓN DOS: MODELO DE REGRESIÓN POISSON

- **Selección de variables**

Paso	1	2	3	4	5	6
Intercepto	-1.284	-1.3925	-1.5275	-2.07	-1.9818	-1.1607
	***	***	***	***	***	***
Persons	0.8264	1.12085	1.16	1.12	1.09126	0.83192
	***	***	***	***	***	***
Child		-1.7733	-0.5265	-0.8077	-1.69	-1.17
		***	n.s.	*	***	***
Camper				0.9178	0.93094	-0.163
				***	***	n.s.
Persons*Child			-0.3512	-0.2504		
			***	*		
Persons*Camper						0.33754
						***
Deviance	2417	1466.5	1457.4	1332.4	1337.1	1323.7
AIC	2758.1	1809.6	1802.462	1679.4	1682.145	1670.728
LogVero	-1377.04	-901.788	-897.231	-834.72	-837.073	-830.364

- **Salidas de R**

```
Call:
glm(formula = count ~ persons + child + camper + persons * camper,
    family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9652  -1.3090  -1.0156   0.1488  15.8421

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.16107    0.24973  -4.649 3.33e-06 ***
persons         0.83192    0.07785  10.686  < 2e-16 ***
child          -1.66697    0.08113 -20.547  < 2e-16 ***
camper         -0.16299    0.29966  -0.544 0.586499
persons:camper  0.33754    0.09037   3.735 0.000188 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2958.4  on 249  degrees of freedom
Residual deviance: 1323.7  on 245  degrees of freedom
AIC: 1670.7

Number of Fisher Scoring iterations: 6
```

- **Expresiones para interpretar las interacciones**

  Al estimar la tasa de peces capturados por un pescador asistente al lago cuando:

  No va en casa rosante $(X_3 = 0)$: $\hat{\mu}_i = \exp(-1.161 + 0.832 X_{1i} - 1.17 X_{2i})$

  Sí va en casa rodante $(X_3 = 1)$: $\hat{\mu}_i = \exp(-1.324 + 1.17 X_{1i} - 1.17 X_{2i})$

- **Salidas en R de los tests de sobredispersión**

  Test de Bohning

```
obs<-count
exp<-predict(ap2mod_pois,type="response")
epi.bohning(obs,exp)
  test.statistic p.value
1       116.5504       0
```

  Chi Cuadrado de Pearson

```
res.p<-residuals(mod_pois,type="pearson")
chicuadrado<-sum(res.p^2)
chicuadrado
[1] 2806.787
chicuadrado/(250-5)
[1] 11.45627
```

  Test de Dean y Lawless

```
T<-sum((obs-exp)^2-exp)/sqrt(2*sum(exp^2))
T
[1] 139.2557
1-pnorm(T)
[1] 0
```

Test de Cameron y Trivedi (1985)

```
dispersiontest(mod_pois)

        Overdispersion test
data:  mod_pois
z = 2.1783, p-value = 0.01469
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  3.200678
```

Test de Cameron y Trivedi (1986)

```
ty<-(obs-exp)^2
tx<-exp
summary(lm(ty~tx-1))

Call:
lm(formula = ty ~ tx - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1260.9   -74.1   -37.4   -21.5 13221.3

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
tx   44.472      7.698   5.777 2.26e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 873 on 249 degrees of freedom
Multiple R-squared: 0.1182,     Adjusted R-squared: 0.1146
F-statistic: 33.37 on 1 and 249 DF,  p-value: 2.261e-08
```

Test de Cameron y Trivedi (1990)

```
ty<-(obs-exp)^2-exp
tx<-exp^2
summary(lm(ty~tx-1))

Call:
lm(formula = ty ~ tx - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1468.6    -6.4    -1.4    -0.4 13013.6

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
tx   1.7732     0.2909   6.095 4.14e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 864.9 on 249 degrees of freedom
Multiple R-squared: 0.1298,     Adjusted R-squared: 0.1263
F-statistic: 37.15 on 1 and 249 DF,  p-value: 4.135e-09
```

ANEXO 10: APLICACIÓN DOS: MODELO DE REGRESIÓN NB2

- **Selección de variables**

Paso	**1**	**2**	**3**
Intercepto	1.7379	-1.2476	-1.625
	***	***	***
Child	-1.4423	-1.8378	-1.7805
	***	***	***
Persons		1.0845	1.0608
		***	***
Camper			0.6211
			**
Theta	0.2384	0.4348	0.4635
Deviance	201.58	209.53	210.65
AIC	894.83	825.04	820.44
LogVero	-444.4164	-408.5186	-405.222

- **Salidas en R del modelo final**

```
Call:
glm.nb(formula = count ~ child + persons + camper, init.theta =
0.4635287626,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6673  -0.9599  -0.6590  -0.0319   4.9433

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6250     0.3304  -4.918 8.74e-07 ***
child        -1.7805     0.1850  -9.623  < 2e-16 ***
persons       1.0608     0.1144   9.273  < 2e-16 ***
camper        0.6211     0.2348   2.645  0.00816 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.4635) family taken to be 1)

    Null deviance: 394.25  on 249  degrees of freedom
Residual deviance: 210.65  on 246  degrees of freedom
AIC: 820.44

Number of Fisher Scoring iterations: 1


              Theta:  0.4635
          Std. Err.:  0.0712

 2 x log-likelihood:  -810.4440
```

ANEXO 11: APLICACIÓN DOS: MODELO DE REGRESIÓN POISSON INFLADO EN CERO

- **Selección de variables**

Paso	**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-0.295	0.1652	-0.52	-0.812	-0.7218	-0.876	-1.265	-0.986	-1.4258	-1.043	-0.5873	-0.93	-0.231	1.0782	-0.053	1.8726
	.	n.s.	**	***	***	***	***	***	***	***	.	***	n.s.	*	n.s.	***
Persons	0.7505		0.8995		0.9572		0.9468		0.94303		0.6866		0.5908	-0.887	0.5437	-0.956
	***		***		***		***		***		***		***	***	***	***
Child			-1.099	1.1988	0.62907	1.3573	0.6042	1.3638	1.4552	1.5065	1.323	1.4179	1.1861	2.0077	1.0831	2.0386
			***	***	n.s.	***	n.s.	***	**	***	**	***	**	***	*	***
Camper							0.7743		0.9795		-0.081		-0.246		-0.475	-1.126
							***		***		n.s.		n.s.		n.s.	**
Persons*Child					-0.4756		-0.487		-0.4407		-0.3961		-0.368		-0.3423	
					***		***		***		**		**		**	
Child*Camper									-1.2039		-1.239		-1.195		-1.1975	
									***		***		***		***	
Persons*Camper											0.3202		0.365		0.4255	
											**		***		***	
AIC	**1857.173**		**1620.3**		**1609.456**		**1532.415**		**1509.485**		**1501.794**		**1480.204**		**1471.256**	
Log Vero	**-925.6**		**-805.1**		**-798.7**		**-759.2**		**-746.7**		**-741.9**		**-730.1**		**-724.6**	

- **Salidas en R del modelo final:**

```
Call:
zeroinfl(formula = count ~ persons * child + child * camper + persons *
camper | child + camper + persons)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-3.34595 -0.71134 -0.45832  0.04856 15.53778

Count model coefficients (poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.05273    0.30388  -0.174  0.86223
persons         0.54372    0.09193   5.914 3.34e-09 ***
child           1.08310    0.43292   2.502  0.01235 *
camper         -0.47478    0.34571  -1.373  0.16965
persons:child  -0.34233    0.12492  -2.740  0.00614 **
child:camper   -1.19746    0.22362  -5.355 8.56e-08 ***
persons:camper  0.42546    0.10206   4.169 3.06e-05 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.8726     0.4990   3.752 0.000175 ***
child         2.0386     0.3269   6.236 4.49e-10 ***
camper       -1.1257     0.3473  -3.241 0.001192 **
persons      -0.9557     0.1999  -4.780 1.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 19
Log-likelihood: -724.6 on 11 Df
```

- **Expresiones para interpretar las interacciones**

Al estimar la tasa de peces capturados por un pescador asistente al lago cuando:

No va en casa rosante $\left(X_3 = 0\right)$: $\mu_i = \exp\left(0.1439 + 0.4789X_{1i} - 0.0203X_{2i} - 0.538X_{3i}\right)$

Sí va en casa rodante $\left(X_3 = 1\right)$: $\mu_i = \exp\left(0.1439 + 0.9324X_{1i} - 1.3943X_{2i} - 0.538X_{3i}\right)$

ANEXO 12: APLICACIÓN DOS: MODELO DE REGRESIÓN NB2 INFLADO EN CERO

- **Selección de variables**

Paso	**1**		**2**		**3**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	1.738	-10.56	-1.2476	-11.47	-1.3119	-4.2514
	***	n.s.	***	n.s.	***	**
Child	-1.4424		-1.8378		-1.2339	2.8694
	***		***		***	***
Persons			1.0845		1.0783	
			***		***	
Log(Theta)	-1.4338	---	-0.8328	---	-0.6197	---
	***	---	***	---	***	---
AIC						
Log Vero	**-444.4**		**-408.5**		**-402.8**	

- **Salidas del modelo final en R:**

```
Call:
zeroinfl(formula = count ~ child + persons | child, dist = "negbin")
Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.69889 -0.56075 -0.36923 -0.05103 10.31866

Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3119     0.2872  -4.567 4.94e-06 ***
child        -1.2339     0.2706  -4.561 5.10e-06 ***
persons       1.0783     0.1108   9.735  < 2e-16 ***
Log(theta)   -0.6197     0.1816  -3.413 0.000643 ***
Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2514     1.3770  -3.087 0.002019 **
child         2.8694     0.8032   3.572 0.000354 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 0.5381
Number of iterations in BFGS optimization: 20
Log-likelihood: -402.8 on 6 Df
```

ANEXO 13: APLICACIÓN DOS: MODELO DE REGRESIÓN *HURDLE* POISSON

- **Selección de variables del modelo *hurdle logit* Poisson**

	1		**2**		**3**		**4**		**5**		**6**	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-0.284	-0.274	-0.32	0.3843	-0.499	0.3843	-1.01	-0.073	-0.0798	-0.073	0.1439	-0.073
	.	*	*	*	**	*	***	n.s.	n.s.	n.s.	n.s.	n.s.
Persons	0.7476		0.8459		0.8976		0.8872		0.5506		0.4789	
	***		***		***		***		***		***	
Child			-1.083	-1.111	0.4982	-1.111	0.4882	-1.129	1.0543	-1.129	-0.0203	-1.129
			***	***	n.s.	***	n.s.	***	*	***	n.s	***
Camper							0.7368	0.7775	-0.4178	0.7775	-0.538	0.7775
							***	**		**	n.s.	**
Persons*Child					-0.4385		-0.451		-0.3311			
					***		***		*			
Child*Camper									-1.224		-1.374	
									***		***	
Persons*Camper									0.411		0.4535	
									***		***	
AIC	**1859.337**		**1636.425**		**1627.5**		**1551.066**		**1513.331**		**1517.354**	
Log Vero	**-926.7**		**-813.3**		**-807.6**		**-767.5**		**-746.7**		**-749.7**	

- **Salidas del modelo *hurdle logit* Poisson en R:**

```
Call:
hurdle(formula = count ~ persons * camper + child * camper | child +
camper)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-1.27929 -0.76328 -0.40870 -0.04532 13.16042

Count model coefficients (truncated poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.14385    0.28940   0.497    0.619
persons         0.47890    0.08855   5.408 6.36e-08 ***
camper         -0.53865    0.34412  -1.565    0.118
child          -0.02030    0.18766  -0.108    0.914
persons:camper  0.45349    0.10256   4.421 9.80e-06 ***
camper:child   -1.37399    0.21766  -6.313 2.74e-10 ***
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.07316    0.23990  -0.305  0.76038
child       -1.12934    0.20896  -5.405 6.49e-08 ***
camper       0.77745    0.28802   2.699  0.00695 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 14
Log-likelihood: -749.7 on 9 Df
```

ANEXO 14: APLICACIÓN DOS: MODELO DE REGRESIÓN *HURDLE* NB2

- **Selección de variables del modelo *hurdle logit* NB2**

	1		2		3		4	
Componentes	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**	**C.C.**	**C.B.**
Intercepto	-1.198	-0.274	-1.4334	0.3843	-1.4334	-1.569	-1.4334	-2.3087
	*	*	*	*	*	***	*	***
Persons	0.9856		1.026		1.026	1.0292	1.026	1.1104
	***		***		***	***	***	***
Child			-1.1642	-1.111	-1.1642	-2.048	-1.1642	-2.138
			***	***	***	***	***	***
Camper								1.0179
								**
Log(theta)	-1.749	---	-1.1294	---	-1.1294	---	-1.1294	---
	*	---	*	---	*	---	*	---
AIC	**900.0442**		**854.7811**		**815.898**		**807.4923**	
Log Vero	**-446**		**-421.4**		**-400.9**		**-395.7**	

- **Salidas en R del modelo *hurdle logit* NB2**

```
Call:
hurdle(formula = count ~ persons + child | persons + child + camper, dist
= c("negbin"))

Pearson residuals:
     Min       1Q   Median       3Q      Max
-0.72427 -0.49730 -0.25410 -0.01439 12.19296

Count model coefficients (truncated negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4334     0.5719  -2.506 0.012196 *
persons       1.0260     0.1552   6.612  3.8e-11 ***
child        -1.1642     0.3177  -3.665 0.000247 ***
Log(theta)   -1.1294     0.5185  -2.178 0.029387 *
Zero hurdle model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3087     0.4612  -5.005 5.57e-07 ***
persons       1.1104     0.1911   5.811 6.19e-09 ***
child        -2.1380     0.3107  -6.882 5.90e-12 ***
camper        1.0179     0.3246   3.136  0.00171 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 0.3232
Number of iterations in BFGS optimization: 12
Log-likelihood: -395.7 on 8 Df
```

ANEXO 15: APLICACIÓN DOS: COMPARACIÓN DE MODELOS

Comparación del modelo Poisson con el NB2:

```
> vuong(ap2mod_pois,ap2mod_bn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.8579
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.002132271
```

Comparación del modelo Poisson con el modelo Poisson inflado en cero

```
> vuong(ap2mod_pois,ap2mod_zip)
Vuong Non-Nested Hypothesis Test-Statistic: -1.797617
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.03611889
```

Comparación del modelo Poisson con el modelo NB2 inflado en cero:

```
> vuong(ap2mod_pois,ap2mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.828868
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.002335647
```

Comparación del modelo Poisson con el modelo hurdle *logit* Poisson:

```
> vuong(ap2mod_pois,ap2mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: -1.384206
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.08314772
```

Comparación del modelo Poisson con modelo hurdle *logit* NB2

```
> vuong(ap2mod_pois,ap2mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.86686
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.002072831
```

Comparación del modelo NB2 con el modelo Poisson inflado en cero

```
> vuong(ap2mod_bn,ap2mod_zip)
Vuong Non-Nested Hypothesis Test-Statistic: 2.455706
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.007030414
```

Comparación del modelo NB2 con el modelo NB2 inflado en cero

```
> vuong(ap2mod_bn,ap2mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: -0.3260894
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.3721784
```

Comparación del modelo NB2 con el modelo hurdle *logit* Poisson

```
> vuong(ap2mod_bn,ap2mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: 2.602533
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.004626894
```

Comparación del modelo NB2 con el modelo hurdle *logit* NB2

```
> vuong(ap2mod_bn,ap2mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -1.200764
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.1149213
```

Comparación del modelo Poisson inflado en cero con el modelo NB2 inflado en cero

```
> vuong(ap2mod_zip,ap2mod_zibn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.474095
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.006678703
```

Comparación del modelo Poisson inflado en cero con el modelo hurdle *logit* Poisson

```
> vuong(ap2mod_zip,ap2mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: 3.074697
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.001053581
```

Comparación del modelo Poisson inflado en cero con el modelo hurdle *logit* NB2

```
> vuong(ap2mod_zip,ap2mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.521806
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.005837707
```

Comparación del modelo NB2 inflado en cero con el modelo hurdle *logit* Poisson

```
> vuong(ap2mod_zibn,ap2mod_hp)
Vuong Non-Nested Hypothesis Test-Statistic: 2.618735
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.004412825
```

Comparación del modelo NB2 inflado en cero con el modelo hurdle *logit* NB2

```
> vuong(ap2mod_zibn,ap2mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -1.713174
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.04334027
```

Comparación del modelo *hurdle logit* Poisson con el modelo hurdle *logit* NB2

```
> vuong(ap2mod_hp,ap2mod_hbn)
Vuong Non-Nested Hypothesis Test-Statistic: -2.663837
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model2 > model1, with p-value 0.003862746
```

ANEXO 16: NOTACIÓN Y SIMBOLOGÍA

$\mathbf{y}$: Vector de observaciones de la variable respuesta

y_i : i-ésima observación de la variable respuesta.

$\mathbf{X}$: Matriz de variables predictoras *

$\mathbf{x}_i$: Vector de variables predictoras *

x_{ij} : i-ésima observación para la j-ésima variable predictora *

$\mathbf{Z}$:Matriz de variables predictoras**

$\mathbf{z}_i$:Vector de variables predictoras para el i-ésimo elemento **

z_{ij} : i-ésima observación para la j-ésima variable predictora **

$\boldsymbol{\beta}$: Vector de coeficientes *

β_j : j-ésimo coeficiente *

$\hat{\boldsymbol{\beta}}$: Vector de coeficientes estimados*

$\hat{\beta}_j$: j-ésimo coeficiente estimado *

$\boldsymbol{\gamma}$: Vector de coeficientes **

γ_j : j-ésimo coeficiente **

$\hat{\boldsymbol{\gamma}}$: Vector de coeficientes estimados **

$\hat{\gamma}$: j-ésimo coeficiente estimado **

α: Parámetro de dispersión en los modelos de tipo binomial negativo

$\hat{\alpha}$: Parámetro estimado de dispersión en los modelos de tipo binomial negativo

$a(\phi)$: Parámetro de escala

θ: Parámetro exponencial

$\boldsymbol{\theta}$: Vector de parámetros en el modelo de regresión para datos de conteo ***

$\hat{\boldsymbol{\theta}}$: Vector de parámetros estimados en el modelo de regresión para datos de conteo. ***

μ: Media de la variable respuesta en un modelo de regresión Poisson o NB.

$\hat{\mu}_i$: i-ésimo valor ajustado / estimado

π_i: Probabilidad de ocurrencia de ceros estructurales (modelos inflados en cero) / Probabilidad de ocurrencia de valores positivos (modelos *hurdle*)

$\hat{\pi}_i$: Probabilidad estimada de ocurrencia de ceros estructurales (modelos inflados en cero) / Probabilidad estimada de ocurrencia de valores positivos (modelos *hurdle*)

$\boldsymbol{\varepsilon}$: Vector de residuales.

p: Número de variables predictoras

n: Tamaño de muestra

$E[.]$: Valor esperado

$V[.]$: Varianza

η: Predictor lineal

$g(.)$: Función de enlace

$f(.)$: Función de probabilidad (para una v.a. discreta) o de densidad (para una v.a. continua)

$f_{cero}(.)$: Función de probabilidad para modelar la probabilidad de ocurrencia de ceros (modelos *hurdle*).

$f_{conteo}(.)$: Función de probabilidad para modelar datos de conteo.

$L(.)$: Función de verosimilitud

$l(.) = \ln L(.)$: Función de log-verosimilitud

$P(Y = y)$: Probabilidad de que la variable aleatoria Y tome el valor de y

$\frac{\partial^k f(x)}{x^k} = f^{(k)}(x)$: Derivada de k-ésimo orden respecto a x, si $k = 1$, $f^{(1)}(x) = f'(x)$

$\square$: Conjunto de los números enteros

$\square^+$: Conjunto de los números enteros positivos

$\Gamma(.)$: Función Gamma

$\Psi(.)$: Función Digamma

χ^2_{Pois}: Estadístico Chi Cuadrado de Pearson para el MRP

χ^2_{NB2} :Estadístico Chi Cuadrado de Pearson para el NB2

$\mathbf{U}$: Gradiente o score

$\mathbf{H}$: Matriz observada de información de Fisher

$\mathbf{I}$: Valor esperado de la matriz de información de Fisher

$\mathbf{I}(.)$: Matriz de información de Fisher

$\mathbf{I}_n$: Matriz de identidad de orden n.

* Válida para los modelos: MRP, NB-C, NB2; para el modelo que explica los conteos aleatorios en los modelos inflados en cero, y para el modelo truncado en cero para el caso de los modelos *hurdle*

** Válida para el modelo que explica la proporción de ceros estructurales en los modelos inflados en cero, y para el modelo que explica la proporción de ceros en los modelos *hurdle*.

*** Para el MRP: $\boldsymbol{\theta} = \boldsymbol{\beta}$

Para el NB-C, NB2: $\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \alpha]'$

Para el ZIP y *hurdle* Poisson: $\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \boldsymbol{\gamma}]'$

Para el ZIBN y *hurdle* BN: $\boldsymbol{\theta} = [\boldsymbol{\beta} \quad \boldsymbol{\gamma} \quad \alpha]'$

Esta aclaración es válida también para $\hat{\boldsymbol{\theta}}$ y $\hat{\boldsymbol{\theta}}^{(k)}$.

ANEXO 17: COMANDOS EN R

Comando en R	Paquete	Utilidad
glm	stats	Construir modelos lineales generalizados, para esta investigación se empleó para construir modelos Poisson
glm.nb	MASS	Construir modelos de regresión NB2. Contiene una subrutina para estimar el parámetro de dispersión
zeroinfl	pscl	Construir modelos inflados en cero
hurdle	pscl	Construir modelos hurdle
logLik	stats	Obtener la log verosimilitud para un determinado modelo estimado
voung	pscl	Obtener el estadístico de prueba del test de Vuong
AIC	stats	Obtener el Criterio de Información de Aikake
epi.bohning	epiR	Obtener el estadístico de prueba del test de Bohning
*outlier*Test	car	Determinar los *outliers* en un modelo lineal generalizado, según los residuales estudentizados más grandes
biserial.cor	ltm	Calcula el coeficiente de correlación biserial
polyserial	polycor	Calcula el coeficiente de correlación poliserial

ANEXO 18: ENCUESTA

ENCUESTA

Buenos días / tardes la aplicación de esta encuesta tiene como objetivo la ejecución de un trabajo de Tesis, para lo cual se requiere su colaboración. La encuesta es anónima y los datos sólo serán utilizados con fines académicos. Muchas gracias por su tiempo. **IMPORTANTE: No existe respuesta correcta o incorrecta.**

1. Sexo: ( ) F ( ) M

2. Edad: _____ años

3. Departamento de nacimiento: _____

4. Distrito de residencia: _____

5. Carrera: _____

6. Situación sentimental:
( ) Soltero(a) sin pareja
( ) Soltero(a) con pareja
( ) Otro – Indicar: _____

7. ¿Ha fumado cigarros alguna vez? ( ) Sí ( ) No
¿Por qué motivo? _____
Si respondió **NO**, pasar a la **pregunta 8.**
Si respondió **SI**, contestar las **siguientes preguntas**:

7.1. ¿A qué edad fumó por primera vez? A los _____ años
7.2. ¿Ha fumado cigarros en el último año? ( ) Sí ( ) No
7.3. ¿Ha fumado cigarros en el último mes? ( ) Sí ( ) No
7.4. ¿Fuma cigarros actualmente? ( ) Sí ( ) No
7.5. ¿Cuántas veces ha intentado dejar de fumar? _____
7.6. ¿Cuántos cigarros fuma a la semana aproximadamente? _____

8. ¿Tiene familiares que fuman regularmente en casa? ( ) Sí ( ) No

9. ¿Tiene compañeros que fuman regularmente en su entorno? ( ) Sí ( ) No

10. ¿Ha recibido charlas sobre el consumo de cigarros y sus consecuencias? ( ) Sí ( ) No

11. ¿A qué enfermedad asociaría usted el consumo excesivo de cigarros? _____

12. ¿Consume bebidas alcohólicas de modo regular? (Al menos 3 fines de semana por mes)
( ) Sí ( ) No

13. Marca con una X, lo(s) producto(s) que alguna vez has consumido:
( ) Puros ( ) Tabaco de Pipa ( ) Tabaco en polvo
( ) Tabaco de mascar (rapé) ( ) Cigarros electrónicos

1.