PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

TITULO DE LA TESIS
ANÁLISIS DE VOTOS ELECTORALES USANDO MODELOS DE REGRESIÓN PARA DATOS DE CONTEO

Tesis para optar el grado de Magíster en Estadística

AUTOR

Norma Contreras Vilca

ASESOR

Dr. Jorge Luis Bazán Guzmán

JURADO

Dr. Cristian Luis Bayes Rodríguez

Dr. Jorge Luis Bazán Guzmán

Dra. Mery Elizabeth Doig Camino

LIMA-PERÚ

2012

Dedicatoria

A mis padres, por darme ejemplos dignos de superación y entrega, porque en gran parte gracias a ellos, hoy puedo ver alcanzada mi meta, y por que el orgullo que sienten por mí, fue lo que me hizo ir hasta el final.

Agradecimientos

En primer lugar agradezco a Dios por ser mi guía y por iluminar mi camino.

Seguidamente agradezco a mi asesor, Dr. Jorge Luis Bazán Guzmán, por la orientación y los conocimientos impartidos para realizar esta investigación.

Asimismo a mi familia y amigos, mil palabras no bastarían para agradecerles su apoyo, su comprensión y sus consejos en los momentos difíciles. De igual manera a los docentes Dr. Cristian Bayes, Dr. Luis Valdivieso y Dra. Elizabeth Doig, por su apoyo y apreciaciones en la presente investigación.

En general, espero no defraudarlos y contar siempre con su valioso apoyo, sincero e incondicional.

Resumen

Se presentan dos modelos de regresión para datos de conteo: el modelo de regresión Poisson y modelo de regresión Binomial Negativa dentro del marco de los Modelos Lineales Generalizados.

Los modelos son aplicados inicialmente a un conjunto de datos conocido como «The Aircraft Damage» presentado en Montgomery (2006) referido al número de daños en las aeronaves durante la guerra de Vietnam.

La principal aplicación de este trabajo será el análisis de los votos obtenidos por el candidato Ollanta Humala Tasso en los resultados de las «Elecciones Generales y Parlamento Andino 2011», analizamos los datos de la primera vuelta a nivel de regiones considerando diversos predictores.

Ambos conjunto de datos, presentan sobredispersión, esto es una varianza mayor que la media, bajo estas condiciones el modelo de Regresión Binomial Negativa resulta más adecuado que el modelo de Regresión Poisson.

Adicionalmente, se realizaron estudios de diagnósticos que confirman la elección del modelo Binomial Negativa como el más apropiado para estos datos.

Palabras-clave: Modelo Lineal Generalizado, Modelo de Regresión Poisson y Modelo de Regresión Binomial Negativa.

Abstract

We present two regressions of models for count data: Poisson Regression and Negative Binomial Regression within the framework of Generalized Linear Models.
The models are applied to a data initially known as The Aircraft Damage referred an Umber of damage located in the aircraft during the Vietnam War and Election Results.
The principal application for this work is to find a regression model to predict the number of votes obtained from the candidate Ollanta Humala Tasso in the Andean Parliament and General Elections, 2011 at the level of regions considering various predictors.
Both the data and Election Results Aircraft Damage occurred, have over dispersion, this is a variance greater than average in certain conditions Regression model Negative Binomial result or As appropriate for the Regression Model Poisson.
Additionally, we performed studies diagnostic confirming the election Negative Binomial model as most appropriate for these data
Keywords: Generalized Linear Model, Poisson Regression and Negative Binomial Regression Model.

Índice general

Lista de Abreviaturas

MLG	Modelo Lineal Generalizado.
MRP	Modelo de Regresión Poisson.
MRBN	Modelo de Regresión Binomial Negativa.
AIC	Criterio de Información de Akaike.
IDH	Índice de Desarrollo Humano.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

Los Modelos Lineales Generalizados (MLG) propuestos por Nelder y Wedderburn (1972) surgen por la necesidad de expresar en forma cuantitativa relaciones entre un conjunto de variables, en la que una de ellas se denomina variable respuesta y las restantes son denominadas covariables, cuando el supuesto de normalidad no es sostenible. Para los MLG la distribución de componentes aleatorias no es necesariamente homocedástica; es decir, no se requiere de un supuesto de homogeneidad de varianzas y tampoco de normalidad, como ocurre en el Modelo Lineal General. Los MLG permiten que el componente aleatorio pueda provenir de la familia exponencial, la cual unifica a los modelos con variables de respuesta categórica y númerica; casos particulares de esta familia son las distribuciones: binomial, poisson, hipergeométrica, binomial negativa, gamma y normal, entre otros.

El MLG difiere del Modelo Lineal General en que la variable respuesta sea un miembro de la familia exponencial donde la respuesta puede ser, heteroscedástica. Así, la dispersión puede variar con la media que a su vez varía con las variables explicativas.

En resumen los Modelos Lineales Generalizados se caracterizan por lo siguiente:

- Los valores observados y_i son independientes.
- No se requiere el supuesto de homogeneidad de variancias. En algunos modelos como el de Regresión Poisson o la de Bernoulli tienen un solo parámetro ajustado, la media μ, de forma que al varie μ varíe también la variancia.

Generalmente, la variable respuesta de interés en el análisis político está representada por datos de conteo; es decir, el número de votos alcanzado por un determinado candidato en una circunscripción electoral, como en las regiones o distritos del Perú. En la mayoría de los casos los datos de conteo no siguen una distribución normal.

El propósito de esta investigación es analizar la relación entre el número de electores que votan por un determinado candidato en una circunscripción electoral y los factores asociados que puede influir en esas cantidades, considerando los modelos de regresión de conteo: Poisson y Binomial Negativa.

El Modelo de Regresión Poisson es un MLG y es el modelo de referencia en estudios de variables de conteo (Cameron y Trivedi (1998); Winkelmann (2000)).

El Modelo de Regresión Poisson (MRP) ha sido usado extensamente en diversas áreas de investigación, pero muy poco en el Análisis Político. En nuestro medio, es casi nula la bibliografía sobre el número de votos obtenidos en un proceso electoral con estudios relacionados a propiedades estadísticas y estimaciones.

El modelo MRP es adecuado cuando los datos no presentan sobredispersión; es decir, cuando la varianza muestral es igual a la media. Se dice que existe sobredispersión cuando la varianza exhibida por los datos es mucho más grande que la que predice el modelo. El Modelo de Regresión Binomial Negativa (MRBN) es casi siempre pensada como el modelo alternativo al Modelo de Regresión Poisson cuando hay sobredispersión en los datos.

1.2. Objetivo de la Tesis

El objetivo general de la tesis es estudiar y presentar las propiedades de los modelos de conteo como parte de los MLG considerando aplicaciones a Resultados Electorales.

- Revisar la literatura acerca de los modelos de regresión de conteo: Poisson y Binomial Negativa como parte de los MLG.
- Evaluar y presentar propiedades de los modelos de regresión de conteo: Poisson y Binomial Negativa.
- Presentar e implementar los métodos de estimación clásica para los modelos de regresión de conteo.
- Aplicación del modelo de regresión de conteo para el análisis de resultados electorales peruanos incluyendo estudios de diagnósticos.

1.3. Organización del Trabajo

El presente trabajo de investigación se encuentra organizado en capítulos que describiremos a continuación.

En el presente capítulo se expone los objetivos de la investigación que se desea realizar. En el capítulo 2, se presentan los conceptos previos para el desarrollo de los modelos de conteo, una revisión sobre los Modelos Lineales Generalizados, los conceptos de la familia exponencial, función verosimilitud, función enlace, variable offset y la función desvío. En el capítulo 3, se explican los modelos de regresión Poisson y Binomial Negativa, para datos de conteo y se detalla el concepto de equidispersión. En el capítulo 4, se describe a detalle las aplicaciones de los modelos de regresión Poisson y Binomial Negativa para los conjuntos de datos «The Aircraft Damage» y del «Voto obtenido por el candidato Ollanta Humala» en las Elecciones Generales y Parlamento Andino 2011 y se presentan los resultados obtenidos en la aplicación los que determinan el mejor modelo. Finalmente, en el capítulo 5 se discute

algunas conclusiones obtenidas en este trabajo. Se analizan las ventajas y desventajas de los modelos propuestos.

En el anexo A se presentan resultados obtenidos por el candidato Ollanta Humala y las variables del contexto social de los electores. En el anexo B se expresan los programas utilizados en R.

Capítulo 2

Modelos Lineales Generalizados

2.1. Conceptos

Familia Exponencial

Sea Y_i una variable aleatoria. La función de densidad o probabilidad de esta variable pertenece a la familia exponencial de distribución si y solo si tiene la siguiente forma Paula (2010):

$$f(y_i;\theta_i,\phi) = exp[\phi^{-1}\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)] \tag{2.1}$$

Donde:

- θ_i es el parámetro canónico
- ϕ es el parámetro de dispersión
- $b(\theta_i)$ y $c(y_i,\phi)$ son funciones conocidas y determinan la función de probabilidad como la binomial, normal ó gamma.

En términos de $b(\theta_i)$ se puede expresar la media y varianza de la siguiente manera:

$$E(y_i) = \mu_i = b^{'}(\theta_i) \tag{2.2}$$

$$Var(y_i) = \phi b^{''}(\theta_i) \tag{2.3}$$

$b^{'}(\theta_i)$ y $b^{''}(\theta_i)$ son respectivamente la primera y segunda derivadas de $b(\theta_i)$ con respecto a θ_i. La función $b^{''}(\theta_i)$ a menudo se expresa en función de μ_i, y se denomina la función varianza.

Función Varianza

La función varianza juega un papel importante en la familia exponencial, ya que caracteriza a la distribución. La función $b^{''}(\theta_i)$ puede ser escrita en función de la media μ_i de la siguiente manera:

$$b^{''}(\theta_i) = \frac{\partial b^{'}(\theta_i)}{\partial \theta_i} = \frac{\partial \mu_i}{\partial \theta_i} \equiv V(\mu_i)$$

Luego podemos escribir:

$$Var(y_i) = \phi^{-1}V(\mu_i)$$

donde $V(\mu_i)$ es llamada la función de varianza e indica la relación entre la media y la varianza.

Para mostrar la relación de la media y varianza, se define:

- $f^{'}(y_i;\theta_i,\phi)$ primera derivada y
- $f^{''}(y_i;\theta_i,\phi)$ segunda derivada de $f(y_i;\theta_i,\phi)$ en (2.1) con respecto a θ_i.

Reemplazando en (2.1):
Si $\phi^* = \phi^{-1}$

$$\begin{aligned} f(y_i;\theta_i,\phi) &= exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi^*} + c(y_i,\phi)\} \\ &= exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi^*}\}exp\{c(y_i,\phi)\} \\ &= c^*(y_i,\phi)exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi^*}\} \end{aligned}$$

Primera derivada:

$$\begin{aligned} f^{'}(y_i;\theta_i,\phi) &= c^*(y_i,\phi)exp\{\frac{y_i\theta_i - b(\theta_i)}{\phi^*}\}\frac{d}{d\theta_i}\{\frac{y_i\theta_i - b(\theta_i)}{\phi^*}\} \\ &= f(y_i;\theta_i,\phi)(\frac{y_i - b^{'}(\theta_i)}{\phi^*}) \end{aligned}$$

Segunda derivada:

$$\begin{aligned} f^{''}(y_i;\theta_i,\phi) &= \frac{d}{d\theta_i}\{f(y_i;\theta_i,\phi)(\frac{y_i - b^{'}(\theta_i)}{\phi^*})\} \\ &= (\frac{d}{d\theta_i}f(y_i;\theta_i,\phi))(\frac{y_i - b^{'}(\theta_i)}{\phi^*}) + (f(y_i;\theta_i,\phi)(\frac{d}{d\theta_i}(\frac{y_i - b^{'}(\theta_i)}{\phi^*})) \\ &= f(y_i;\theta_i,\phi)(\frac{y_i - b^{'}(\theta_i)}{\phi^*})^2 + f(y_i;\theta_i,\phi)(\frac{-b^{''}(\theta_i)}{\phi^*}) \\ &= f(y_i;\theta_i,\phi)(\frac{y_i - b^{'}(\theta_i)}{\phi^*})^2 - f(y_i;\theta_i,\phi)(\frac{b^{''}(\theta_i)}{\phi^*}) \end{aligned}$$

Luego

$$f^{'}(y_i;\theta_i,\phi) = f(y_i;\theta_i,\phi)\{\frac{y_i - b^{'}(\theta_i)}{\phi^*}\} \tag{2.4}$$

$$f^{''}(y_i;\theta_i,\phi) = f(y_i;\theta_i,\phi)\{\frac{y_i - b^{''}(\theta_i)}{\phi^*}\}^2 - f(y_i;\theta_i,\phi^*)\{\frac{b^{''}(\theta_i)}{\phi^*}\} \tag{2.5}$$

Integrando a ambos lados de las ecuaciones (2.4) y (2.5) se obtiene las expresiones (2.6) y (2.7) respectivamente con respecto a y_i, que nos permite llegar a (2.2) y (2.3).

$$0 = \frac{E(y_i) - b^{'}(\theta_i)}{\phi^*} \tag{2.6}$$

$$0 = \frac{E[\{(y_i) - b^{'}(\theta_i)\}^2]}{\phi^{*2}} - \frac{b^{''}(\theta_i)}{\phi^*} \tag{2.7}$$

Los lados izquierdos son ceros definido por Jong y Heller (2008), puesto que:

$$f^{'}(y_i;\theta_i,\phi)dy_i = \frac{\partial}{\partial\theta_i}\int f(y_i;\theta_i,\phi)dy_i$$

$$f^{''}(y_i;\theta_i,\phi)dy_i = \frac{\partial^2}{\partial\theta_i^2}\int f(y_i;\theta_i,\phi)dy_i$$

donde $\int f(y_i;\theta_i,\phi)dy_i = 1$, la demostración se puede dar por terminada, asumiendo que la integración y diferenciación pueden ser intercambiadas.

2.1.1. Elementos del Modelo Lineal Generalizado

Dada una variable respuesta y_i, la construcción de un MLG está compuesto por los siguientes elementos:

- **Componente Aleatorio**: Dado $Y_1, ..., Y_n$ un conjunto de variables respuesta, caracterizada por los parámetros θ_i y ϕ, pertenece a la familia exponencial si presenta la forma:

$$f(y_i;\theta_i,\phi) = exp[\phi^{-1}\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)]$$

- **Componente sistemático**: Especifica las variables explicativas $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ que ingresan en forma de efecto fijos de un modelo lineal, y se relacionan como:

$$\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip},\ \ i = 1, ..., n$$

  Esta combinación lineal de las variables explicativas se denominan predictor lineal y se puede generalizar para el término:

$$\eta_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}$$

  donde β_j es el j-ésimo coeficiente de regresión y x_{ij} es el j-ésimo predictor, en el i-ésimo individuo, para $i = 1, ..., n$ y $j = 1, .., p$.

  Los elementos η_i pueden ser expresado como un vector de la siguiente manera $(\eta_1, ..., \eta_n)$.

- **Función Enlace**: Los dos componentes son combinada en el modelo mediante la elección de un enlace denotado como $g(.)$, de manera que relaciona μ_i con el predictor lineal η_i, a través de la función.

$$g(\mu_i) = \eta_i$$

De este modo, para $i = 1, ..., n$ y el valor esperado de la variable respuesta:

$$E(y_i \mid \boldsymbol{x}_i) = \mu_i$$

Cada distribución tiene una función enlace especial que se denomina enlace canónico, para la cual existe un estadístico suficiente y se da cuando $\eta_i = g(\mu_i) = \theta_i$.

Los enlaces más conocidos para $g(\mu_i)$ son:

Cuadro 2.1: Enlaces de los Modelos Lineales Generalizados

Función	**Enlace**
Logaritmo	$\log\mu_i = \eta_i$
Identidad	$\mu_i = \eta_i$
Raíz Cuadrada	$\sqrt{\mu_i} = \eta_i$
Logit	$\log(\frac{\mu_i}{n-\mu_i}) = \eta_i$
Recíproca	$\frac{1}{\mu_i} = \eta_i$
Exponencial	$\mu^n = \eta_i$
Inverso	$\frac{-1}{\mu_i} = \eta_i$
Normal Inversa	$\phi^{-1}(\mu_i)$

La elección del enlace dependerá de la familia de distribución, del tipo de respuesta y de la aplicación en que se emplea.

2.2. Estimación Clásica en los Modelos Lineales Generalizados

Dentro de los Modelos Lineales Generalizados para estimar los parámetros desconocidos, se utilizan varios métodos, los más comunes son el método de Mínimos Cuadrados Ordinarios y el método de Máxima Verosimilitud (Inferencia Clásica), pero también se tiene el método bayesiano. La estimación de parámetros más utilizadas para el modelo de regresión lineal es el método de Mínimo Cuadrado Ordinarios; éste no resulta adecuado cuando el componente aleatorio del modelo no es normal, en este caso se debe utilizar el método de Máxima Verosimilitud.

2.2.1. Función de Verosimilitud

La estimación de máxima verosimilitud se basa en la elección de estimaciones de los parámetros que maximizan la probabilidad de haber observado la muestra $\boldsymbol{y} = (y_1, ..., y_n)^T$ un conjunto de n observaciones aleatorias independientes cuya función de densidad de probabilidad $f(y_i; \theta_i, \phi)$ depende de un vector de parámetros θ_i y ϕ. Si los y_i son independientes, entonces su función de probabilidad conjunta es:

$$f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} f_i(y_i; \theta_i, \phi)$$

Se escribe la función log-verosimilitud de la siguiente forma:

$$L(\theta_i, \phi) = lnf(y_i; \theta_i, \phi) \equiv \sum_{i=1}^{n} lnf_i(y_i; \theta_i, \phi)$$

Si $f(y_i; \theta_i, \phi)$ pertenece a la familia exponencial de probabilidad entonces $L(\theta_i, \phi)$ tiene la forma siguiente:

$$\begin{aligned} L(\theta_i, \phi) &= \sum_{j=1}^{n} \{ln(c(y_j, \phi)) + \frac{y_j\theta_i - b(\theta_i)}{\phi}\} \\ &= \sum_{j=1}^{n} ln(c(y_j, \phi)) + \frac{n\{\overline{y}\theta_i - b(\theta_i)\}}{\phi} \\ &= \frac{n\{\overline{y}\theta_i - b(\theta_i)\}}{\phi} + \sum_{j=1}^{n} ln(c(y_j, \phi)) \end{aligned}$$

Se desea encontrar los estimadores de θ_i que maximizen a L, por lo que se puede tomar derivada de primer orden a ambos lados de la igualdad para encontrar el estimado.

$$\frac{\partial l(\theta_i, \phi)}{\partial \theta_i} = \frac{n\{\overline{y} - b^{'}(\theta_i)\}}{\phi} = 0$$

Como $c(y_i, \phi)$ no depende de θ_i, su derivada es 0

$$b^{'}(\theta_i) = \bar{y}$$

Donde la estimación máxima verosimilitud de θ_i se obtiene mediante la búsqueda θ_i de tal manera que:

$$E(b^{'}(\theta_i)) \equiv E(\mu_i)$$

es igual a la media de la muestra $\bar{y}$.

Entonces para cualquier distribución de la familia exponencial se tiene:

$$\hat{\mu}_i = \overline{y}$$

Estas ecuaciones de estimación no se pueden resolver directamente y el principal interés es la estimación de $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_P)^T$ y el parámetro de dispersión ϕ.

Para el algoritmo de estimación se utiliza el método de Score de Fisher.

2.2.2. Función Score e Información de Fisher

Este método implica una sustitución de la matriz de derivadas parciales de segundo orden por la matriz de valores esperados de derivadas parciales; es decir, la matriz de información observada por la matriz de información de Fisher.

Función Score para $\boldsymbol{\beta}$

Se considera la partición $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$, como en Paula (2010), que denota el logaritmo de la función de verosimilitud por $L(\boldsymbol{\theta})$.

Para obtener la función score para los parámetros $\boldsymbol{\beta}$ se calcula inicialmente derivadas:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_j} &= \sum_{i=1}^{n} \phi\{y_i \frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j} - \frac{db(\theta_i)}{d\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_j}\} \\
&= \sum_{i=1}^{n} \phi\{y_i V_i^{-1}(\frac{d\mu_i}{d\eta_i})x_{ij} - \mu_i V_i^{-1}(\frac{d\mu_i}{d\eta_i})x_{ij} \\
&= \sum_{i=1}^{n} \phi\{V_i^{-1}(\frac{d\mu_i}{d\eta_i})(y_i - \mu_i)x_{ij}\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (i) \\
&= \sum_{i=1}^{n} \phi\{\sqrt{\frac{\omega_i}{V_i}}(y_i - \mu_i)x_{ij}\} \\
&= \phi X W^{1/2} V^{1/2}(\boldsymbol{y} - \boldsymbol{\mu})
\end{aligned}$$

De (i):

$$\begin{aligned}
V_i^{-1}(\frac{d\mu_i}{d\eta_i}) &= \frac{\sqrt{(d\mu_i/\eta_i)^2}}{\sqrt{V_i^2}} \\
&= \frac{1}{\sqrt{V_i}}\sqrt{\frac{(d\mu_i/\eta_i)^2}{V_i}} \\
&= \frac{1}{\sqrt{V_i}}\sqrt{w_i} \\
&= \sqrt{\frac{w_i}{V_i}}
\end{aligned}$$

donde

$$w_i = \frac{(d\mu_i/d\eta_i)^2}{V_i}$$

Luego se escribe la función score en forma de matriz, descrito por Paula (2010):

$$\begin{aligned} \boldsymbol{U}_\beta(\boldsymbol{\theta}) &= \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ &= \phi \boldsymbol{X}^T \boldsymbol{W}^{1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned}$$

Donde:

- $\boldsymbol{X}$ es una matriz $n \times p+1$ de rango completo cuyas filas son denotadas por $\boldsymbol{x}_i^T$, $i = 1, ..., n$,
- $\boldsymbol{W} = diag\{w_1, ..., w_n\}$ es una matriz de ponderaciones,
- $\boldsymbol{V} = diag\{V_1, ..., V_n\}$,
- $\boldsymbol{y} = (y_1, ..., y_n)^T$ y
- $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$.

Matriz de Información de Fisher para $\boldsymbol{\beta}$

Para obtener la matriz de información de Fisher se necesita calcular la segunda derivada:

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_l} = \phi \sum_{i=1}^{n} (y_i - \mu_i) \frac{d^2\theta_i}{d\mu_i^2} (\frac{d\mu_i}{d\eta_i})^2 x_{ij} x_{il} + \phi \sum_{i=1}^{n} (y_i - \mu_i) \frac{d\theta_i}{d\mu_i} \frac{d^2\mu_i}{d\eta_i^2} x_{ij} x_{il} - \phi \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} (\frac{d\mu_i}{d\eta_i})^2 x_{ij} x_{il}$$

Cuyos valores esperados están dados por:

$$\begin{aligned} E\{\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_l}\} &= -\phi \sum_{i=1}^{n} \frac{d\theta_i}{d\mu_i} (\frac{d\mu_i}{d\eta_i})^2 x_{ij} x_{il} \\ &= -\phi \sum_{i=1}^{n} \frac{(d\mu_i/d\eta_i)^2}{V_i} x_{ij} x_{il} \\ &= -\phi \sum_{i=1}^{n} \omega_i x_{ij} x_{il} \end{aligned}$$

Luego, podemos escribir la información de Fisher para $\boldsymbol{\beta}$ en forma matricial y denotarlo como:

$$\begin{aligned} \boldsymbol{K}_{\beta\beta}(\boldsymbol{\theta}) &= E\{-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\} \\ &= \phi \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \end{aligned}$$

En particular, para el enlace canónico ($\theta_i = \eta_i$), estas cantidades toman formas simplificadas:

$$\boldsymbol{U}_{\beta} = \phi \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu})$$

$$\boldsymbol{K}_{\beta\beta} = \phi \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}$$

Si particionamos el vector de parámetros $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, la función score y la matriz de información de Fisher para el parámetros $\boldsymbol{\beta}_1$, se tiene respectivamente:

$$\boldsymbol{U}_{\boldsymbol{\beta}_1} = \phi \boldsymbol{X}_1^T \boldsymbol{W}^{1/2} \boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu})$$

$$\boldsymbol{K}_{\beta_1\beta_1} = \phi \boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1$$

Función Score para ϕ

La función score para el parámetro ϕ como en Paula (2010) está dada por:

$$\boldsymbol{U}_{\phi}(\boldsymbol{\theta}) = \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}}$$

Del punto (2.1)

$$\begin{aligned} L(f(y_i;\theta_i,\phi)) &= \prod_i^n f(y_i;\theta,\phi) \\ &= exp[\sum_{i=1}^n [\phi^{-1}\{y_i\theta_i - b(\theta_i)\} + c(y_i,\phi)]] \end{aligned}$$

Tomando logaritmo:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \ln(l(f(y_i;\theta_i,\phi))) \\ &= \sum_{i=1}^n [\phi\{y_i\theta_i - b(\theta_i)\}] + \sum_{i=1}^n [c(y_i,\phi)] \end{aligned}$$

derivando:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}} = \sum_{i=1}^n \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^n c^{'}(y_i,\phi)$$

donde:

$$c'(y_i,\phi) = \frac{\partial c(y_i,\phi)}{\partial \phi}$$

Matriz de Información de Fisher para ϕ

Para obtener la información de Fisher para ϕ se tiene que calcular:

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \phi^2} = \sum_{i=1}^{n} c''(y_i, \phi)$$

donde:

$$c''(y_i, \phi) = \frac{\partial^2 c(y_i, \phi)}{\partial \phi^2}$$

Por lo tanto, la información de Fisher para ϕ es dada por:

$$\begin{aligned} \boldsymbol{K}_{\phi\phi}(\boldsymbol{\theta}) &= E\{-\frac{\partial L(\boldsymbol{\theta})}{\partial \phi^2}\} \\ &= -\sum_{i=1}^{n} E\{c''(Y_i, \phi)\} \end{aligned}$$

2.2.3. Estimación de los Parámetros

Estimación de $\boldsymbol{\beta}$

Mediante el proceso iterativo de Newton-Raphson se obtiene la estimación de Máxima Verosimilitud de $\boldsymbol{\beta}$ y se define mediante la expansión de la función score $\boldsymbol{U}_\beta$ en torno a un valor inicial $\boldsymbol{\beta}^{(0)}$, Paula (2010) tal que:

$$\boldsymbol{U}_\beta \cong \boldsymbol{U}_\beta^{(0)} + \boldsymbol{U}_\beta^{'(0)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})$$

donde, $\boldsymbol{U}_\beta^{'}$ denota la primera derivada de $\boldsymbol{U}_\beta$ respecto a $\boldsymbol{\beta}^T$, siendo $\boldsymbol{U}_\beta^{'}(0)$ y $\boldsymbol{U}_\beta^{(0)}$ respectivamente, estas cantidades son evaluadas en $\boldsymbol{\beta}^{(0)}$. Por lo tanto, repetir el procedimiento anterior, genera el proceso iterativo siguiente:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \{(-\boldsymbol{U}_\beta^{'})^{-1}\}^{(m)} \boldsymbol{U}_\beta^{(m)}$$

donde $m = 0, 1, ...$ Como la matriz $-\boldsymbol{U}_\beta^{'}$ puede ser no positiva definida, la aplicación del método Score de Fisher sustituye la matriz $-\boldsymbol{U}_\beta^{'}$ por el correspondiente valor esperado $\boldsymbol{K}_{\beta\beta}$. Esto da como resultado el siguiente proceso iterativo:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \{(-\boldsymbol{K}_{\beta\beta}^{-1})\}^{(m)} \boldsymbol{U}_\beta^{(m)}$$

donde $m = 0, 1, ...$

Trabajando el lado derecho de la expresión anterior, se llega a mínimos cuadrados iterativos reponderados como sigue:

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{X})^{(-1)} \boldsymbol{X}^T \boldsymbol{W}^{(m)} \boldsymbol{z}^{(m)} \tag{2.8}$$

Donde:

- $m = 0, 1, ...$
- $\boldsymbol{z} = \boldsymbol{\eta} + \boldsymbol{W}^{-1/2} \boldsymbol{V}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu})$

Además $\boldsymbol{z}$ desempeña el papel de una variable dependiente modificada y $\boldsymbol{W}$ es una matriz de pesos que cambia en cada paso del proceso iterativo.

La convergencia de (2.8) generalmente se produce en un número finito de pasos, independientemente de los valores iniciales utilizados. Es usual tomar como valor inicial $\boldsymbol{\eta}^{(0)} = \boldsymbol{g}(\boldsymbol{y})$ para (2.8).

Por ejemplo con la binomial logística, obtenemos $w = n\mu(1-\mu)$ y la modificación de la variable dependiente dada por $z = \eta + (y + n\mu)/n\mu(1-\mu)$.

Recordando para el modelo lineal normal no es necesario recurrir al proceso iterativo (2.8) para obtener la estimación de probabilidad máxima. En este caso, $\hat{\boldsymbol{\beta}}$ toma la forma de:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^1 \boldsymbol{X}^T \boldsymbol{y}$$

Se puede observar que el lado derecho de (2.8) no depende de ϕ. Por lo tanto para obtener $\hat{\boldsymbol{\beta}}$ no es preciso conocer ϕ.

Estimación de ϕ

Igualando la función score U_ϕ a cero, se llega a la siguiente solución:

$$\sum_{i=1}^{n} c'(y_i, \hat{\phi}) = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) - \sum_{i=1}^{n} \{y_i \tilde{\theta}_i - b(\tilde{\theta_i})\}$$

donde $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ es la función desvío del modelo a estimar.

Se ha encontrado que la estimación de máxima verosimilitud para ϕ para el caso normal y normal inversa, igualando U_ϕ a cero, está dada por:

$$\hat{\phi} = \frac{n}{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}$$

2.3. Enlace Canónico

Asumiendo ϕ conocido, la función log verosimilitud para un MLG con respuesta independiente se puede expresar como:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi^{-1}\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi)$$

Un caso particularmente importante se produce cuando el parámetro canónico (θ_i) coincide con la predicción lineal; es decir, cuando $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. En este caso, $L(\boldsymbol{\beta})$ viene dada por:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi^{-1}\{y_i \sum_{j=1}^{p} x_{ij}\beta_j - b(\sum_{j=1}^{p} x_{ij}\beta_i)\} + \sum_{i=1}^{n} c(y_i, \phi)$$

La creación del estadístico;

$$S_j = \phi \sum_{j=1}^{n} y_j x_{ij}$$

donde $L(\boldsymbol{\beta})$ es expresado de la siguiente forma:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} S_j\beta_j - \phi \sum_{j=1}^{n} b(\sum_{j=1}^{p} x_{ij}\beta_j) + \sum_{i=1}^{n} c(y_i, \phi)$$

Luego, por el teorema de factorización del estadístico $\boldsymbol{S} = (S_1, ..., S_p)^T$ es suficiente para el mínimo vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Los enlaces que corresponden al estadístico son llamadas enlaces canónicos y juegan un papel importante en la teoría de los MLGs.

Una de las ventajas de usar enlaces canónicos es que garantizan la concavidad de la función log-verosimilitud $L(\boldsymbol{\beta})$ y por tanto se obtienen resultados asintóticos fácilmente. La concavidad de la función log-verosimilitud $L(\boldsymbol{\beta})$ garantiza la unicidad de la estimación de máxima verosimilitud $\hat{\beta}$ cuando ésta existe.

2.4. Función Desvío

La bondad de ajuste en un MLG es evaluado a través de la función desvío:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\{L(\boldsymbol{y}; \boldsymbol{y}) - L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})\}$$

Suponiendo que el logaritmo de la función verosimilitud está definida como en Paula (2010):

$$L(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} L(\mu_i; y_i)$$

donde:

$$\mu_i = g^{-1}(\eta_i)$$

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Luego el modelo con un parámetro por observación se llama un modelo saturado.

Para el modelo saturado ($p = n$) la función $L(\boldsymbol{\mu}; \boldsymbol{y})$ está estimada por:

$$L(\boldsymbol{y}; \boldsymbol{y}) = \sum_{i=1}^{n} L(y_i; y_i)$$

Es decir, la estimación de máxima verosimilitud de μ_i está dada por $\tilde{\mu_i} = y_i$. Cuando $p < n$, denotamos la estimación de $L(\boldsymbol{\mu}; \boldsymbol{y})$ por $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$. En este caso, la estimación de máxima verosimilitud μ_i viene dada por $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$, donde:

$$\hat{\eta_i} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$$

Entonces la calidad del ajuste de MLG se evalúa a través de la función desvío:

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \phi D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\{L(\boldsymbol{y}; \boldsymbol{y}) - L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})\}$$

que es la diferencia entre el logaritmo de la función de verosimilitud del modelo saturado (con n parámetros) y el modelo a estimar (con p parámetros) evaluados en una estimación máxima verosimilitud $\hat{\boldsymbol{\beta}}$. Un valor pequeño para una función desvío indica un menor número de parámetros, obtenemos un ajuste tan bueno como cuando se ajuste un modelo saturado.

Sea:

$$\hat{\theta_i} = \theta_i(\hat{\mu_i})$$

$$\tilde{\theta_i} = \theta_i(\tilde{\mu_i})$$

Estimaciones de máxima verosimilitud de $\boldsymbol{\theta}$ para los modelos con p parámetros ($p < n$) y modelos saturado ($p = n$), respectivamente, tenemos que la función $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ alternativamente está dada por:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\{y_i(\tilde{\theta_i} - \hat{\theta_i}) + (b(\hat{\theta_i}) - b(\hat{\theta_i}))\}$$

Donde el desvío es siempre mayor o igual a cero. Para probar la adecuación a un MLG, el valor del desvío debe ser comparado con el percentil de alguna distribución de probabilidad referente. En la práctica, la función desvío se compara con los percentiles de una distribución χ^2_{n-p} McCullagh y Nelder (1991).

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}$$

2.5. La Variable Offset

En aquellos casos en que los conteos de las observaciones se dan en períodos de tiempo, tamaño de población, espacios no homogéneos entre los valores de las variables explicativas se requiere una corrección, es recomendable incluir en el modelo un término adicional: la variable de control, también denominada offset que se simboliza por t.

Si μ_i es la media de conteo de y_i, luego la presencia del ratio μ_i/t de interés Jong y Heller (2008) y

$$g(\frac{\mu_i}{t}) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

cuando $g(.)$ es la función Log, esto se convierte en:

$$ln(\frac{\mu_i}{t}) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

entonces:

$$\ln(\mu_i) = \ln(t) + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

donde la variable t es llamada de exposición y $\ln(t)$ es llamada offset. Un offset efectivamente es otra variable x en la regresión, con un coeficiente β igual 1. t_i es un vector de columnas que contiene las variables de exposición para cada unidad de observación. Con la variable offset, y tiene un valor esperado directamente proporcional a la exposición:

$$\mu_i = te^{x_i^T \beta}$$

Entonces el offset es utilizado para hacer la correción mencionada anteriormente.

2.6. Selección del Modelo

Existen varios criterios para seleccionar el mejor modelo alternativo, entre los principales criterios para la comparación tenemos el Criterio de Información de Akaike - AIC, propuesto por Akaike (1974), es un índice que evalúa tanto el ajuste del modelo a los datos como la complejidad del modelo. La idea es seleccionar un modelo que es parsimonioso, que tenga un número reducido de parámetros. Cuando el logaritmo de la función verosimilitud $L(\hat{\boldsymbol{\beta}})$ crece o aumenta el número de parámetros del modelo, una propuesta razonable sería encontrar un modelo con menor valor para la función:

$$AIC = -L(\hat{\boldsymbol{\beta}}) + 2p$$

Este método se extiende directamente para los MLG. Donde el método de Akaike puede ser expresado de una forma más simple con la función desvío del modelo. En este caso el criterio consiste en encontrar un modelo tal que el valor sea mínima:

$$AIC = -D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) + 2p$$

donde $D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ es la función desvío del modelo y p es el número de parámetros.

2.7. Análisis de Diagnóstico

Seleccionado el modelo, es importante hacer un análisis de diagnóstico para verificar el ajuste de los datos a un MLG. Para este proceso, se seguirá la metodología propuesta por Paula (2010) que consiste en:

- **Puntos Leverage:**

  Considerando la expresión para $\hat{\boldsymbol{\beta}}$ obtenida en el proceso de convergencia interactivo dada en (2.8), Paula (2010) se tiene:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{W}} \hat{\boldsymbol{z}} \tag{2.9}$$

  con $\hat{\boldsymbol{z}} = \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{W}}^{-1/2} \hat{\boldsymbol{V}}^{-1/2} (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$

  Por lo tanto, $\hat{\boldsymbol{\beta}}$ puede ser interpretado como una solución de mínimos cuadrados de la regresión lineal de $\hat{\boldsymbol{W}}^{1/2} \boldsymbol{z}$ frente a la columna $\hat{\boldsymbol{W}}^{1/2} \boldsymbol{X}$. La matriz de proyección $\hat{\boldsymbol{H}}$ de mínimos cuadrados de Regresión Lineal de $\hat{\boldsymbol{z}}$ versus a $\boldsymbol{X}$ con ponderación $\hat{\boldsymbol{W}}$ para los MLG se define como:

$$\hat{\boldsymbol{H}} = \hat{\boldsymbol{W}}^{1/2} \boldsymbol{X} (\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{W}}^{1/2} \tag{2.10}$$

  Que sugiere utilizar los elemento de la diagonal $\hat{h}_{ii}$ de la matriz sombrero $\hat{\boldsymbol{H}}$, para detectar presencia de puntos leverage.

  Donde:

  - $\hat{h}_{ii} = \partial \hat{y}_i / \partial y_i$
  - $\hat{\boldsymbol{H}}$ es simétrica é idempotente

  Por ser idempotente, se tiene: $Rango(\hat{\boldsymbol{H}}) = traza(\hat{\boldsymbol{H}}) = \sum_i^n h_{ii} = p$. Luego, se sugiere que los puntos $h_{ii} \geq \frac{2p}{N}$ donde pueden ser considerados puntos palanca o de alto leverage.

- **Residuos para Puntos Aberrantes**: Es importante precisar otro tipo de residuos que son definidos como:

  - **Residuos en base a desvío**: Los residuos más utilizado en los Modelos Lineales Generalizados se definen a partir de los componentes de la función de desvío. La versión estándar (Ver McCullagh (1987)) es la siguiente:

$$\begin{aligned} t_{D_i} &= \frac{d^*(y_i;\hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}} \\ &= \frac{\phi^{1/2} d(y_i;\hat{\mu}_i)}{\sqrt{1-\hat{h}_{ii}}} \end{aligned}$$

    donde $d(y_i;\hat{\mu}_i) = \pm\sqrt{2}\{y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\}^{1/2}$. Con el signo de $d(y_i;\hat{\mu}_i)$ la misma de $(y_i - \hat{\mu}_i)$.

  - **Residuo de Pearson**: El residual de Pearson es el residual más lógico e intuitivo. Este residual corrige la heterocedasticidad debido a que incorpora la varianza de μ, sin embargo una desventaja es que su distribución es bastante asimétrica para modelos no normales. El residuo de Pearson está definido como:

$$rp_i = \phi^{1/2} r_i^*$$

  donde $\boldsymbol{r}_i^* = \hat{\boldsymbol{V}}^{1/2}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

- **Influencia ó Distancia de Cook**: Suponiendo ϕ conocido, la distancia en verosimilitud, cuando eliminamos la i-ésima observación es denotada por:

$$LD_i = 2\{L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})\}$$

  Es por tanto una medida que verifica una influencia de la eliminación de la i-ésima observación en $\boldsymbol{\beta}$. Puesto que es imposible obtener una forma análitica para LD_i, es usual utilizar una segunda aproximación por series de Taylor en torno de $\hat{\boldsymbol{\beta}}$. Esta extensión conduce al siguiente resultado:

$$LD_i \cong (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\{-\ddot{L}_{\beta\beta}(\hat{\boldsymbol{\beta}})\}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\}$$

  Sustituyendo $-\ddot{\boldsymbol{L}}_{\beta\beta}(\hat{\boldsymbol{\beta}})$ por el correspondiente valor esperado y $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ obtenemos:

$$LD_i \cong \phi(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) \tag{2.11}$$

  Así tenemos una buena aproximación para LD_i cuando $L(\boldsymbol{\beta})$ es aproximadamente cuadrática en torno a $\hat{\boldsymbol{\beta}}$.

Generalmente no es posible obtener una forma cerrada para $\hat{\boldsymbol{\beta}}_{(i)}$ se obtiene una aproximación en un paso (Ver, Cook y Weisberg (1982)) que consiste en tomar la primera iteracción del proceso iterativo por el método Score de Fisher cuando se comienza en $\hat{\boldsymbol{\beta}}$

Esta aproximación es introducida por Pregibon (1981):

$$\boldsymbol{\beta}^1_{(i)} = \hat{\boldsymbol{\beta}} + \{-\ddot{\boldsymbol{L}}_{\beta\beta}(\hat{\boldsymbol{\beta}})\}^{-1}\boldsymbol{L}_{(i)}(\hat{\boldsymbol{\beta}})$$

Donde $\boldsymbol{L}_{(i)}(\hat{\boldsymbol{\beta}})$ es la función de logaritmo de maxima verosimilitud sin la i-ésimo observación. Sustituyendo nuevamente $-\ddot{\boldsymbol{L}}_{\beta\beta}(\hat{\boldsymbol{\beta}})$ por $\boldsymbol{K}(\hat{\boldsymbol{\beta}})$ se obtiene:

$$\boldsymbol{\beta}^1_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{r}_{P_i}\sqrt{\hat{w}_i\varphi^{-1}}}{(1-\hat{h}_{ii})}(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{2.12}$$

Finalmente, sustituyendo en la expresión (2.11), la distancia de Cook es definida por:

$$LD_i \cong \{\frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})}\}t^2_{S_i} \tag{2.13}$$

donde:

$$t_{S_i} = \frac{\phi^{1/2}(y_i - \hat{\mu}_i)}{\sqrt{\hat{\boldsymbol{V}}_i^{1/2}(1-\hat{h}_{ii})}}$$

Incorpora el $i-$ésimo h_{ii} elementos de la matriz sombrero $\hat{\boldsymbol{H}}$

- **Gráfico de probabilidad normal con Envelope**

  Para evaluar el ajuste de un modelo, también se puede utilizar el gráfico de probabilidad normal o semi-normal con envelope simulado.

  El gráfico de $t_{D(i)}$, versus a los valores esperados de las estadísticas de la normal estandar, $Z_{(i)}$ es dado por:

$$E(Z_{(i)}) \cong \phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$

  donde $\phi(.)$ es la función de distribución acumulada de $N(0,1)$.

  También existe el gráfico de probabilidad medio-normal con banda de ajuste simulada, definido como el gráfico de $E = |t^*_{(i)}|$ frente a los valores esperados de $E = (|Z_{(i)}|)$. Se tiene la aproximación:

$$E(|Z_{(i)}|) \cong \phi^{-1}\left(\frac{n+i+1/2}{2n+9/8}\right)$$

  Si se gráfica A_i versus $E(|Z_{(i)}|)$. Puede ser informativo sobre la presencia de puntos aberrantes y/o influyentes.

Adicionalmente al análisis de diagnóstico, de manera complementaria se tiene:

- Residuos estandarizados versus a variables explicativas: Representa los residuos frente a los valores ajustados ayuda a identificar si la falta de linealidad o la heterocedasticidad es debido a algún punto aberrante. Si un punto está relativamente por encima o muy por debajo de la recta horizontal, es un valor atípico.
- Normalidad de los errores (q-q plot): El gráfico cuantil - cuantil sirve para ver si los residuos tiene distribución gaussiana (normal). En el caso perfecto, todos los puntos estarían en línea recta. Los puntos que más se desvia de la línea recta aparecen con etiquetas identificadas.
- Raíz de valor absoluto de residuo frente a valores ajustados: éste gráfico ayuda para el diagnóstico de la homocedasticidad, pero dificulta el diagnóstico de linealidad; esto es debido a las transformaciones que se someten los residuos, por lo que no ofrece ninguna información relevante para el análisis de los residuos.
- Valores atípico frente a leverage: gráfico de valores atípicos, el leverage es una medida de influencia que tiene un punto en el cálculo de los coeficientes del modelo. El leverage se basa en la aportación del punto a las varianzas de las variables independientes. Los puntos poseen una influencia notable si el residuo correspondiente se separa mucho del cero. Se suele considerar muy influyente si supera la distancia de Cook igual a 1.

Capítulo 3

Modelos de Regresión para Datos de Conteo

3.1. Modelo de Regresión Poisson

Cuando la variable respuesta es de conteo. Es conveniente utilizar la distribución de Poisson. Con el Modelo de Regresión de Poisson (MRP), la media de μ se explica en términos de las variables y através de un enlace adecuado.

3.1.1. Distribución Poisson

Sea Y una variable aleatoria discreta que indica el número de veces que cierto evento ocurre, tal que la función de probabilidad de Y_i está dada por:

$$f(y_i) = P(Y_i = y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}, \quad \text{para } y = 1, 2, 3...$$

donde:

- y_i: es el número de ocurrencias de un evento
- μ: es un parámetro positivo que representa el número de veces que se espera que ocurra el evento durante un periodo.

La función acumulada de la distribución Poisson se expresa por:

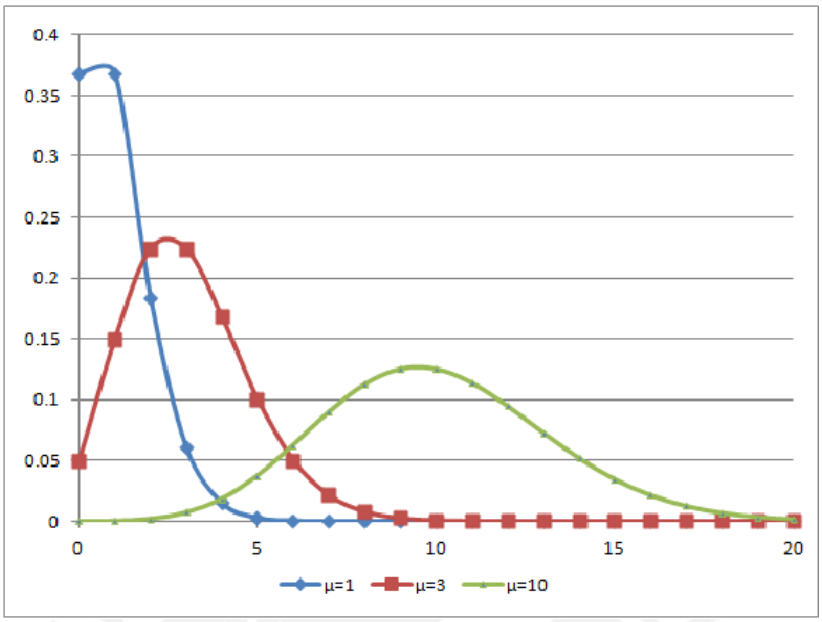
$$F(y|\mu) = \frac{\Gamma([y_i + 1], \mu)}{[y_i]!}$$

para $y_i \geq 0$ donde $\Gamma(x, y)$ es la función Gamma incompleta

Propiedades de la distribución de Poisson:

1. Si μ crece, la masa de la distribución se desplaza hacia la derecha. Entonces: $E(y_i) = \mu$. El parámetro μ es conocido como "tasa" dado que es el número esperado de veces que un evento ha ocurrido por unidad de tiempo.
2. La varianza es igual a la esperanza en la distribución de Poisson. Esta propiedad se conoce como equidispersión: $E(y_i) = Var(y_i) = \mu$
3. A medida que μ crece, $P(Y_i = 0)$ decrece.

Figura 3.1: Distribución de Poisson



4. A medida que μ crece, la distribución de Poisson se aproxima a la distribución normal.

La función de probabilidad puede tomar diversas formas y valores para los parámetros que caracteriza esta distribución. En la Figura 3.1 se presenta como si fuera densidades, pero se debe considerar que se trata de valores discretos para la función de probabilidad de la distribución Poisson para diferentes valores de μ_i

3.1.2. La Distribución Poisson como Familia Exponencial

Sea $Y_1, ..., Y_n$ variables aleatorias independientes e idénticamente distribuidos. La función de probabilidad de este vector pertenece a la familia exponencial y se puede escribir de la siguiente forma:

$$\begin{aligned} lnf(y_i) &= exp\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\} \\ &= -\ln(y_i!) + \frac{y_i\theta_i - b(\theta_i)}{\phi} \end{aligned}$$

Donde:

- $\phi = 1$, parámetro de escala
- $\theta_i = \ln(\mu_i)$
- $b(\theta_i) = e^{\theta_i}$
- $c(y_i, \phi) = -\ln(y_i!)$

Esto muestra que la distribución de Poisson es de la familia exponencial, donde:

$$b'(\theta_i) = e^{\theta_i} = \mu_i = E(y_i)$$

$$b''(\theta_i) = \mu_i = Var(y_i)$$

Es decir, en el modelo Poisson la media y la varianza son iguales entre si e igual a μ_i.

3.1.3. Modelo de Regresión Poisson

Decimos que una variable Y_i sigue el modelo de Regresión Poisson si se cumple que:

$$Y_i \sim P(\mu_i), \quad i = 1, 2, 3, ...n$$

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Donde:

- $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ es el vector de covariables explicativas.
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$, es el vector de parámetros desconocidos.

Los elementos del Modelo de Regresión Poisson son:

- **Componente Aleatorio**: Dado $Y_1, ..., Y_n$ un vector de variable respuesta positiva y $\boldsymbol{x}_i = (x_{i1}, ...x_{ip})^T$ un vector de covariables explicativas con parámetro μ_i especifica que:

$$Y_i \sim P(\mu_i), \quad i = 1, 2, 3, ...n$$

- **Componente sistemático**: Dado μ_i, y el llamado predictor lineal simbolizado por:

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \\ &= \beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta_j} x_{ij} \\ &= x_i^T \boldsymbol{\beta} \end{aligned}$$

- **Función Enlace**: Ambos componentes desarrollados anteriormente son combinados en el modelo, mediante la elección de una función enlace:

$$g(\mu_i) = \eta_i$$

Las más usada para MRP, son:

Cuadro 3.1: Enlaces para el Modelo de Regresión Poisson

Función	Enlace
Logaritmo	$\log\mu_i = \eta_i$
Identidad	$\mu_i = \eta_i$
Raíz Cuadrada	$\sqrt{\mu_i} = \eta_i$

Cuando el enlace logaritmo $\hat{\mu}_i = exp(x_i^T\hat{\boldsymbol{\beta}})$ es positivo.

3.1.4. Función Desvío

Se tiene $\theta_i = \log(\mu_i)$, lo que implica $\tilde{\theta}_i = \log(y_i)$ para $y_i > 0$ y $\hat{\theta}_i = \log(\hat{\mu}_i)$. Por lo tanto Paula (2010):

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\}$$

Si $y_i = 0$, el i-ésimo término de $D(y;\hat{\mu})$ es $2\hat{\mu}_i$. Por lo tanto, tenemos el siguiente resultado para el modelo Poisson de la función desvío:

$$d^2(y_i,\hat{\mu}_i) = \begin{cases} 2\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\} & ,\text{si } y_i > 0, \\ 2\hat{\mu}_i & ,\text{si } y_i = 0. \end{cases}$$

3.1.5. Estimación Máxima Verosimilitud del Modelo de Regresión Poisson

Sea $Y_1, ..., Y_n$ un conjunto con n observaciones aleatorias e independientes donde el predictor es x, entonces la función de verosimilitud es:

$$\frac{\prod_{i=1}^{n}\mu_i^{y_i} exp(-\sum_{i=1}^{n}\mu_i)}{\prod_{i=1}^{n} y_i!}$$

donde $\mu_i = g^{-1}(\boldsymbol{x^T},\boldsymbol{\beta})$.

Una vez que la función enlace se ha seleccionado, se puede maximizar. El logaritmo de la Función Verosimilitud está dado por:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(\mu_i) - \sum_{i=1}^{n}\ln(\mu_i) - \sum_{i=1}^{n}\ln(y_i!)$$

El valor que maximice $L(\boldsymbol{\beta})$, es el vector de coeficientes estimado $\hat{\boldsymbol{\beta}}$.

3.2. Equidispersión

El Modelo de Regresión Poisson se presenta como un modelo con indudables mejoras para representar datos de conteos, sin embargo éste puede resultar inapropiado debido a incumplimiento de ciertos supuestos, cuyo origen es diverso Winkelmann (2000).

La distribución de Poisson se caracteriza por la *equidispersión*, esto es:

$$Var(y) = E(y) = \mu$$

La equidispersión constituye un supuesto básico de diversos MLG. Una violación del supuesto de la variancia, es suficiente para violar el supuesto distribución de Poisson. Sin embargo, un problema que se da con cierta frecuencia en este modelo es que la relación media-varianza no es equitativa. Las desviaciones en relación a la equidispersión pueden resultar en:

- Sobredispersión: $Var(y) > E(y)$ es decir si $\sigma^2 > 1$.
- Infradispersión o Subdispersión: $Var(y) < E(y)$ es decir si $\sigma^2 < 1$

Tal como señalan Krzanowki (1998) y Winkelmann (2000), es mucho más frecuente una situación de sobredispersión que de infradispersión.

Cuando existe exceso de variación en los datos, las estimaciones de los errores estándar pueden resultar sesgadas, pudiendo presentarse errores en las inferencias a partir de los parámetros del modelo de regresión Krzanowki (1998). Fenómeno que ocurre en aplicaciones con distribuciones con varianza poco flexible como la Poisson o Binomial.

Entre las diversas causas de sobredispersión, se tiene:

- Alta variabilidad en los datos
- Los datos no provienen de una distribución Poisson
- Los eventos no ocurren independientemente a través del tiempo
- Falta de estabilidad; es decir la probabilidad de ocurrencia de un evento puede ser independiente de la ocurrencia de la media μ Winkelmann (2000) como omitir variables explicativas o que entran al modelo a través de alguna transformación en lugar de linealmente.
- Errores al elegir la función enlace.
- Es la heterogeneidad de la muestra que puede ser debido a la variabilidad entre experimentos.

Existen diversas propuestas para detectar sobredispersión, una de ella es el **Índice de dispersión** (I_n): Lindsey (1995B) propone aplicar (I_n), como un indicador para evaluar el supuesto de equidispersión. Se define como la razón entre la varianza y la esperanza matemática.

$$I_n = \frac{Var(y)}{E(y)} \tag{3.1}$$

Teóricamente, $Var(y) = E(y)$, el Indice de dispersión debería ser igual a 1. Entonces:

- Posiblemente exista sobredispersión, si $I_n > 1$,
- Indica infradispersión, si $I_n < 1$.

La presencia de sobredispersión como de infradispersión dependerá de la magnitud del valor del Indice de dispersión.

Otro indicador simple y sencillo para determinar sobredispersión es: Si la varianza estimada es más del doble de la media estimada, probablemente los datos permanezcan sobredispersos aún después de la inclusión de regresores. Cameron y Trivedi (1986).

3.3. Modelo de Regresión Binomial Negativa

El Modelo de Regresión Binomial Negativa (MRBN) es casi siempre pensada como el modelo alternativo al Modelo de Regresión Poisson que no impone igualdad entre la media y la varianza, cuando hay sobredispersión en los datos.

3.3.1. Distribución Binomial Negativa

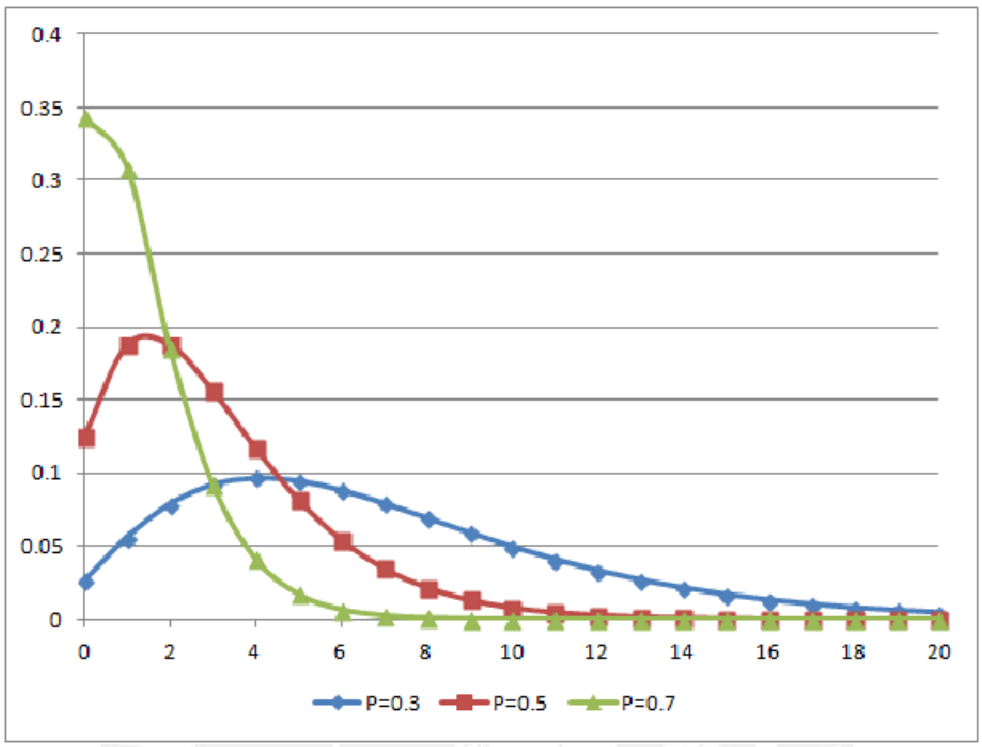
La densidad de la distribución binomial negativa es dada por:

$$f(y_i) = \frac{\Gamma(\phi + y_i)}{\Gamma(y_i + 1)\Gamma(\phi)}(1 - \mu_i)^{y_i}\mu_i^{\phi}$$

Propiedades de la distribución Binomial Negativa:

- $E(Y) = \phi\frac{1-\mu}{\mu}$
- $Var(Y) = \phi\frac{1-\mu}{\mu^2}$

Figura 3.2: Distribución Binomial Negativa (0.5,10)



La función de probabilidad puede tomar diversas formas y valores de los parámetros que caracteriza a la distribución Binomial Negativa. En la figura 3.2 se presenta como si fueran densidades, pero se trata de valores discretos, donde $\boldsymbol{Y}_i \sim BN(0{,}5, 10)$.

3.3.2. La Distribución Binomial Negativa como Familia Exponencial

La función de probabilidad de la binomial negativa es dada por:

$$f(y_i;\mu,\phi) = \frac{\Gamma(\phi+y_i)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i}\left(\frac{\phi}{\mu_i+\phi}\right)^{\phi}$$

donde $y = 0, 1, ...$, con parámetros μ_i y ϕ, con $\mu_i > 0$ y $\phi > 0$.

Si $1/\phi \to 0$, entonces $Var(Y_i) \to \mu_i$ y la distribución binomial negativa converge a una distribución de Poisson.

Cuando ϕ es fijo esta densidad pertenece a la familia exponencial y podríamos hablar de un MLG binomial negativa.

Entonces, si denotamos $Y|z \sim P(z)$ y $Z \sim G(\mu,\phi)$ donde ϕ no depende de μ Paula (2010). En este caso:

$$E(Z) = \mu$$

$$Var(Z) = \frac{\mu^2}{\phi}$$

Se tiene que:

$$f(y|z) = \frac{e^{-z}z^y}{y!}$$

$$g(z;\mu,\phi) = \frac{1}{\Gamma(\phi)}(\frac{z\phi}{\mu})^{\phi}e^{-}\frac{\phi z}{\mu}\frac{1}{z}$$

La función de probabilidad Y viene dada por:

$$\begin{aligned} P(\boldsymbol{Y}=y) &= \int_0^{\infty} f(y|z)g(z;\mu,\phi)dz \\ &= \frac{1}{y!\phi}(\frac{\phi}{\mu})^{\phi}\int_0^{\infty} e^{-z(1+\phi/\mu)}z^{\phi+y-1}dz \end{aligned}$$

Transformando la variable:

$$t = z(1+)\frac{\phi}{\mu}$$

Tenemos:

$$\frac{dz}{dt} = (1+\frac{\phi}{\mu})^1$$

De aquí se deduce que:

$$\begin{aligned} P(\boldsymbol{Y}=y) &= \frac{1}{y!\Gamma(\phi)}(\frac{\phi}{\mu})^{\phi}(1+\frac{\phi}{\mu})^{-(\phi+y)}\int_0^{\infty} e^{-t}t^{\phi+y-1}dt \\ &= \frac{\Gamma(\phi+y)\mu^y\phi^{\phi}}{\Gamma(\phi)\Gamma(y+1)(\mu+\phi)^{\phi+y}} \\ &= \frac{\Gamma(\phi+y)}{\Gamma(y+1)\Gamma(\phi)}(\frac{\mu}{\mu+\phi})^y(\frac{\phi}{\mu+\phi})^{\phi} \\ &= \frac{\Gamma(\phi+y)}{\Gamma(y+1)\Gamma(\phi)}(1-\pi)^{\phi}\pi^y \end{aligned}$$

En el que $\pi = \mu/(\mu+\phi)$.

Por lo tanto Y sigue una distribución binominal negativa de media μ y parámetro de dispersión ϕ.

Entonces la función de probabilidad de esta distribución, se puede escribir de la siguiente forma:

$$logf(y) = exp\{ylog(\frac{\mu}{\mu+\phi}) - \phi log(\frac{\mu+\phi}{\phi}) + log\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\}$$

donde:

- $\phi = 1$ parámetro de escala
- $\theta = log(\frac{\mu}{\mu+\phi})$
- $b(\theta) = \phi log(\frac{\mu+\phi}{\phi})$
- $b(\theta) = \phi log(1 - e^{\theta})$
- $c(y_i, \phi) = log[\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}]$

Además se deduce lo siguiente:

$$E(\boldsymbol{Y_i}) = \mu_i$$

$$Var(\boldsymbol{Y_i}) = \mu_i + \frac{\mu_i^2}{\phi}$$

3.3.3. Modelo de Regresión Binomial Negativa

Decimos que una variable Y_i sigue el modelo de Regresión Binomial Negativa, si cumple que:

$$\boldsymbol{Y}_i \sim BN(\mu_i, \phi), \quad i = 0, 1, 2, 3, ...,$$

$$g(\mu_i) = x_i^T \boldsymbol{\beta}$$

donde:

- $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ es el vector de covariables explicativas.
- $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$, es el vector de parámetros desconocidos.

Elementos del Modelo de Regresión Binomial Negativa:

- **Componente Aleatorio**: Sea $Y_1, ..., Y_n$ una variable aleatoria independiente que indica el número de sucesos necesarios para obtener r-éxitos. Es decir, el número de éxito está predeterminado y la aleatoriedad es el número de sucesos, de modo que:

$$\boldsymbol{Y}_i \sim BN(\mu_i, \phi), \quad i = 0, 1, 2, 3, ...,$$

- **Componente sistemático**: Dado μ_i, y el llamado predictor lineal simbolizado por:

$$\eta_i = x_i^T \boldsymbol{\beta}$$

- **Función Enlace**: Los dos componentes son combinados en el modelo, mediante la elección de la función enlace:

$$g(\mu_i) = \eta_i$$

Donde la $g(.)$ es una función enlace.

Algunas enlaces usados para MRBN, son:

Cuadro 3.2: Enlaces para el Modelo de Regresión Binomial Negativa

Función	Enlace
Logaritmo	$\log\mu_i = \eta_i$
Identidad	$\mu_i = \eta_i$
Raíz Cuadrada	$\sqrt{\mu_i} = \eta_i$

3.3.4. Función Desvío

Si se asume ϕ fijo, la función desvío es dada por Paula (2010):

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}[\phi log\{\frac{\hat{\mu}_i+\phi}{y_i+\phi}\} + y_i log\{\frac{y_i(\hat{\mu}_i+\phi)}{\hat{\mu}_i(y_i+\phi)}\}]$$

donde $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Bajo la hipótesis de que el modelo adoptado es correcto $D^*(\mathrm{y};\hat{\boldsymbol{\mu}})$ para ϕ y μ_i grande. Sigue una distribución $X^2_{(n-p)}$ con $(n-p)$ grado de libertad.

3.4. Estimación Máxima Verosimilitud para Modelo de Regresión Binomial Negativa

Se considera la partición $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$, que denota el logaritmo de la función verosimilitud por: Paula (2010)

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n}[\log\{\frac{\Gamma(\phi+y_i)}{\Gamma(y_i+1)\Gamma(\phi)}\} + \phi\log\phi + y_i\log\mu_i - (\phi+y_i)\log(\mu_i+\phi)]$$

donde $\mu_i = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$, es una función score para $\boldsymbol{\beta}$.

Calculamos inicialmente las derivadas para la función score para $\boldsymbol{\beta}$:

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_j} &= \sum_{i=1}^{n}\{\frac{y_i}{\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{\beta_j} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\} \\
&= \sum_{i=1}^{n}\{\frac{y_i}{\mu_i}\frac{d\mu_i}{d\eta_i}x_{ij} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ij}\} \\
&= \sum_{i=1}^{n}\{\frac{\phi(d\mu_i/d\eta_i}{\mu_i(\phi+\mu_i)}(y_i-\mu_i)x_{ij}\} \\
&= \sum_{i=1}^{n} w_i f_i^{-1}(y_i-\mu_i)x_{ij}
\end{aligned}$$

Donde:

$$w_i = \frac{(d\mu_i/d\eta_i)^2}{(\mu_i^2\phi^{-1}+\mu_i)}$$

$$f_i = \frac{d\mu_i}{\eta_i}$$

Luego podemos expresar la función score en forma matricial para $\boldsymbol{\beta}$:

$$\boldsymbol{U}_\beta(\boldsymbol{\theta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{F}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{3.2}$$

Donde:

- $\boldsymbol{X}$ es una matriz con modelo lineal: $\boldsymbol{x}_i^T, i = 1, ..., n$,
- $\boldsymbol{W} = diag\{w_1, ..., w_n\}$ con $w_i = \frac{(d\mu_i/d\eta_i)^2}{(\mu_i^2\phi^{-1}+\mu_i)}$
- $\boldsymbol{F} = diag\{f_1, ..., f_n\}$ con $f_i = \frac{d\mu_i}{d\eta_i}$
- $\boldsymbol{y} = (y_1, ..., y_n)^T$
- $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$

Lo mismo podemos expresar para la función score de ϕ, dada por:

$$U_\phi(\boldsymbol{\theta}) = \sum_{i=1}^{n}[\psi(\phi+y_i) - \psi(\phi) - \frac{(y_i+\phi)}{(\phi+\mu_i)} + log\{\frac{\phi}{(\phi+\mu_i)}\} + 1] \tag{3.3}$$

donde $\psi(.)$ es una función digama.

Para obtener la matriz de información de Fisher calculamos las derivadas:

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_l} = -\sum_{i=1}^{n}\{\frac{(\phi+y_i)}{(\phi+\mu_i)^2} - \frac{y_i}{\mu_i}\}(\frac{d\mu_i}{d\eta_i})^2 x_{ij}x_{il} + \sum_{i=1}^{n}\{\frac{y_i}{\mu_i} - \frac{(\phi+y_i)}{(\phi+\mu_i)}\}\frac{d^2\mu_i}{d\eta_i^2}x_{ij}x_{il}$$

Cuyos valores esperados son dados por:

$$\begin{aligned} E\{\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\beta_j\partial\beta_l}\} &= -\sum_{i=1}^{n}\{\frac{\phi(d\mu/d\eta_i)^2}{(\phi+\mu_i)}x_{ij}x_{il} \\ &= -\sum_{i=1}^{n} w_i x_{ij} x_{il} \end{aligned}$$

Luego, podemos expresar la información de Fisher para $\boldsymbol{\beta}$, en forma matricial:

$$\boldsymbol{K}_{\beta\beta}(\boldsymbol{\theta}) = E\{\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\beta\partial\beta^T}\} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$$

Lawless (1982) muestra que la información de Fisher para ϕ se puede expresar como:

$$\boldsymbol{K}_{\beta\beta}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\{\sum_{j=1}^{\infty}(\phi+j)^2 Pr(Y_i \geq j) - \phi^{-}1\mu_i/(\mu_i+\phi)$$

donde β y ϕ son parámetros ortogonales. Por lo tanto, la matriz de información de Fisher para θ asume la forma de bloque diagonal:

$$\boldsymbol{K}_{\theta\theta} = \begin{bmatrix} \boldsymbol{K}_{\beta\beta} & 0 \\ 0 & \boldsymbol{K}_{\phi\phi} \end{bmatrix}$$

La estimación de máxima verosimilitud para $\boldsymbol{\theta}$ y ϕ puede ser obtenida a través de un algoritmo de mínimos cuadrados ponderados para obtener $\hat{\boldsymbol{\theta}}$ desarrollado a partir del punto (3.2) y el método de Newton-Raphson para obtener $\hat{\boldsymbol{\phi}}$ desarrollado a partir del punto (3.3) que se describe a continuación:

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{y}^{*(m)}$$

y

$$\phi^{(m+1)} = \phi^{(m)} - \{\frac{\boldsymbol{U}_{\phi}^{m}}{\ddot{\boldsymbol{L}}_{\phi\phi}^{(m)}}\}$$

para $m = 0, 1, 2, ...$, en la que:

$$\boldsymbol{y}^{*} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{F}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$$

3.5. Implementación Computacional

3.5.1. Ajuste del Modelo

El Modelo de Regresión Poisson y el modelo de Regresión Binomial Negativa en su estimación clásica son casos particulares de la estimación presentada en el capítulo 2 del presente estudio para los MLG.

La implementación computacional para la estimación clásica se realiza a través de la librería **Mass** y **glm2** del programa R Development Core Team (2011), mediante las funciones:

glm2:

Estima los modelos lineales generalizados, pero con un método de ajuste modificado predeterminado que proporciona una mayor estabilidad para ciertos modelos que pueden fallar al converger con la función glm.

glm.bn:

El paquete MASS, proporciona la función binomial negativa que directamente se puede enlazar en la función **glm()**, siempre que el argumento de θ_i sea especificado. θ_i no se conoce, pero se estima a partir de los datos, el modelo binomial negativa no es un caso especial de los MLG, sin embargo, un ajuste de los Modelos Lineales puede ser reutilizado en los MLG, métodologia de cálculo por iteracción de los β dado θ_i y viceversa. Esto conduce a estimaciones de los Modelos Lineales tanto para β y θ_i.

stepAIC:

Una manera de aplicar el criterio de Akaike - AIC, es partiendo del mayor modelo cuyos resultados se guarda en el objeto **fit.model**, para después utilizar el comando stepAIC. Cuando más pequeño son los criterios mejor son los ajustes.

3.5.2. Gráfico de Diagnóstico del modelo

Muestra la sensibilidad del modelo usado para el análisis de diagnóstico en los MLG, identificando puntos de leverage, influencia ó distancia de Cook, residuos(aberrantes) usando los residuos t_{D_i}, luego de ajustar el modelo, tratado en el capítulo 2 de la sección 2.7. La implementación computacional para la adecuación del modelo, se usa el programa diagnóstico desarrollados por Paula (2010), que presenta:

- **Punto Leverage**

  Con este gráfico se desea verificar si alguna observación son punto leverage.

  Inicialmente se ilustra como calcular h_{ii}. Los valores se almacenan en $fit.model$. La matriz diseño $\boldsymbol{X}$ se obtiene con el siguiente comando:

```
X=model.matrix(fit.model)
```

Donde $\boldsymbol{V}$ se puede mostrar la matriz $\hat{\boldsymbol{V}}$. Obtenemos la diagonal principal de $\boldsymbol{V}$ debe ser obtenido a partir de ajustes de los modelos que a su vez son extraídos a través del comando **fitted(fit.model)**. Como por ejemplo la matriz con las funciones de varianza estimada sería obtenido con un modelo de Poisson de la siguiente manera:

```
V = fitted(fit.model)
V = diag(V)
```

En particular una matriz $\hat{\boldsymbol{W}}$ también depende de los valores ajustados, sin embargo tanto, como en la matriz de peso, se puede obtener directamente mediante:

```
W=fit.model$weights
W=diag(V)
```

Una vez obtenida la matriz $\hat{\boldsymbol{W}}$ se puede obtener los elementos h_{ii} con la matriz:

```
H = solve(t(X)%*%W%*%X)
H = sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
```

Vector hat o leverages:

```
h = diag(H)
```

Grafico de índice para h_{ii}, a fin de detectar punto leverage.

```
Plot(h, xlab="Indice", ylab="leverage")
```

- **Residuos para puntos Aberrantes**

  Almacenando en **fit** la estimación de ϕ, éste componente de desvío de residuos estandarizados son obtenidos de la siguiente manera:

  Residuo en base a Desvío:

```
rd = resid(fit.model, type= "deviance")
td = rd*sqrt(fi/(1-h))
```

  Residuo de Pearson:

```
rp = resid(fit.model, type= "pearson")
rp = sqrt(fi)*rp
ts = rp/sqrt(1 - h)
```

  Recordando que los enlaces canónicos $\boldsymbol{W}$ y $\boldsymbol{V}$ coinciden.

- **Influencia ó distancia de Cook**

  Los puntos de influencia se detectan mediante el análogo del estadístico de Cook de los modelos lineales clásicos. La influencia puede ser medida a través del cambio en la estimación de los párametros cuando una i-ésima observación es retirada.

  El vector de la distancia de Cook es facilmente obtenido con el comando:

```
LD=h(ts^2)/(1-h)
plot(LD, ylab="Distancia de cook", xlab="Indice")
```

  La construcción de los gráficos desarrollados por Gilberto Paula se encuentra en:

  - $http: www.ime.usp.br/ \sim giapaula/diag - pois$
  - $http: www.ime.usp.br/ \sim giapaula/diag - bino$

  Se ejecuta a través de la secuencia de comandos:

```
fit.model <- ajuste
attach(dados)
source("diag_pois")
```

Gráfico de probabilidad normal con Envelope:

Otra técnica para evaluar el ajuste del modelo son las bandas de ajuste a través de simulaciones el cual se denomina Envelope. Consiste en generar residuos que tienen media cero y matriz de varianza - convarianza $(I_n - H)$. El procedimiento es:

1. Generar n observaciones $P(\mu)$ y almacenarlas en el vector $\boldsymbol{y} = (y_1, ..., y_n)^T$
2. Ajustar $\boldsymbol{y}$ frente $\boldsymbol{X}$ y obtener $r_i = y_i - \hat{y}_i$, $i = 1, ..., n$ tenemos que $E(r_i) = 0$, $Var(r_i) = 1 - h_{ii}$ y $Cov(r_i, r_j) = -h_{ij}$
3. Obtenemos $t_{D_i} = r_i / \{1 - h_{ii}\}^{1/2}, i = 1, ..., n$
4. Repetir los pasos (1)-(3) m veces. Luego tenemos, residuos que genera $t^*_{(ij)}, i = 1, ..., n$ y $j = 1, ..., m$
5. Colocamos cada grupo de n residuos en orden creciente, obteniendo $t^*_{(i)j}, i = 1, ..., n$ y $j = 1, ..., m$
6. Obtener los limites $t^*_{(i)I} = min_j t_{(i)j}$ y $t^*_{(i)S} = max_j t^*_{(i)j}$, asi, los limites correspondientes del i-ésimo resíduo seria dado por $t^*_{(i)I}$ y $t^*_{(i)S}$

Sugiere Atkinson (1985) generar $n = 19$ veces, tal que la probabilidad que el valor absoluto de un residual se encuentra fuera del envelope, se aproxima igual a $\frac{1}{20} = 0{,}05$.

La construcción de los gráficos desarrollados por Gilberto Paula, se encuentra en:

- $http : www.ime.usp.br/ \sim giapaula/envel - pois$
- $http : www.ime.usp.br/ \sim giapaula/envel - bino$

Se ejecuta a través de las secuencia de comandos:

```
fit.model <- ajuste
attach(dados)
source("envel_pois")
```

Además se usará de manera complementaria, implementada en R:

plot:

Los residuos pueden guiar sobre la adecuación del modelo. La función genérica plot(), muestra los gráficos residuales para un objeto del tipo "lm"ó "glm", que genera figuras de:

- Residuos estandarizados frente a variables explicativas
- Normalidad de los Errores (q-q plot)
- Raíz de valor absoluto de residuos
- Valores atípicos frente a leverage

Capítulo 4

Aplicación

4.1. The Aircraft Damage

Para ilustrar la metodología presentada en el Capítulo 3, se analiza el Modelo de Regresión Poisson para los datos de Aviones dañados de Montgomery (2006). Los datos consisten en 30 observaciones y considera las siguientes variables:

- Damage: número de daños encontrados en las aeronaves durante la guerra de Vietnam, en la armada de los Estados Unidos
- Type: variable binaria que indica el tipo de avión (0 para aviónes A-4 Skyhawk4, 1 para aviones A-6 Intruder)
- Bombload: carga de bombas en toneladas
- Airexp: totales de meses de experiencia de la tripulación

4.1.1. Estadística Descriptiva preliminar The Aircraft Damage

Previo al análisis del Modelo Lineal Generalizado para datos de conteo se llevó a cabo el análisis exploratorio, los resultados se presentan en el cuadro 4.1, donde se observa que el promedio de daños ubicados en las aeronaves es aproximadamente de 2 daños, con una tendencia a variar por debajo o encima. Además existen naves que no sufren ningún daño y otras que tuvieron un valor máximo de 7 daños.

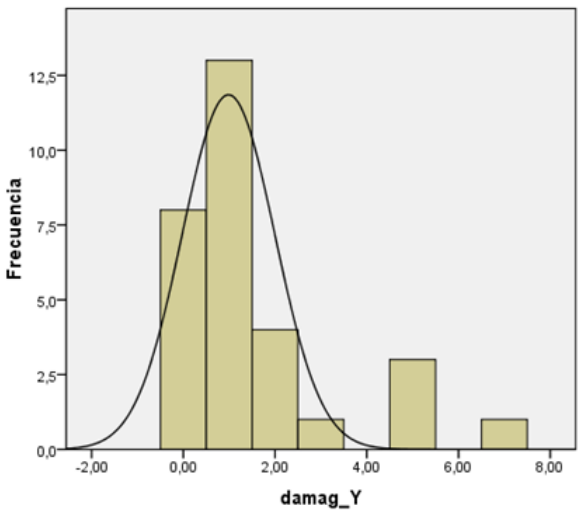
Con respecto a la carga de bombas en los aviones el promedio es de 8, con una tendencia a variar de 9 aproximadamente, con un valor mínimo de 4 y 14 como máximo.

El promedio de meses de experiencia de la tripulación de las aeronaves es de 81, con una tendencia a variar por debajo o encima de los 19 meses y la mayor cantidad de meses de experiencia es de 120, mientras que la más pequeña es de 50 con una amplitud de distribución de 70.

Cuadro 4.1: Estadística Descriptiva The Aircraft Damage - Preliminar

Estadísticas	damagY	type	bomb	air
Media	1.53	0.50	8.10	80.76
Mediana	1.00	0.50	7.50	80.25
Moda	1.00	0.00	7.00	50.00
Desv. Tip.	1.77	0.50	2.98	19.44
Varianza	3.15	0.25	8.99	377.93
Asimetría	1.72	0.00	0.66	0.28
Rango	7.00	1.00	10.00	70.00
Mínimo	0.00	0.00	4.00	50.00
Máximo	7.00	1.00	14.00	120.00

Figura 4.1: Distribución The Aircraft Damage



En la figura N° 4.1 se presenta la distribución de la variable número de daños encontrados en las aeronaves durante la guerra de Vietnam, además se observa una fuerte asimetría hacia la derecha, por existir mayor cantidad de aeronaves con daños encontrados,

El supuesto fundamental para la aplicación del Modelo de Regresión Poisson, es que exista equidispersión, el cual fue descrito en (3.2) del capítulo 3. Para determinar que no exista sobredispersión de la variable respuesta, se presenta a continuación el Indice de dispersión:

$$\begin{aligned} I_n &= \frac{S_y^2}{\overline{y}} \\ &= \frac{3{,}15}{1{,}53} \\ &= 2{,}06 \end{aligned}$$

Nótese que los datos presentan sobredispersión, según la ecuación anterior, donde el Indice de dispersión es mayor a 1. No obstante para ilustrar la metodológía del Modelo de Regresión Poisson, ignoramos la sobredispersión y estimaremos los parámetros mediante la Función de Verosimilitud.

4.1.2. Modelo de Regresión Poisson para datos The Aircraft Damage

Supongamos que el número de daños en las aeronaves en cada misión es independiente al de otras para un Modelo de Regresión Poisson con parámetros μ_i.

Sea Y número de daños ubicados en la aeronaves que se produce en 30 misiones, suponemos:

$$damage_i \sim Poisson(\mu_i)$$

Mediante el criterio de AIC se determina el mejor modelo o en su defecto el más apropiado para el conjunto de datos «The Aircraft Damage». Ver cuadro N° 4.2.

Cuadro 4.2: Valores AIC para los modelos de datos The Aircraft Damage

Modelo	**Null Deviance**	**Residual Deviance**	**Función Desvío**	**AIC**
Type	5388	38.28	28.95	95.98
Bombload	53.88	29.21	45.79	86.90
Airexp	53.88	50.54	6.20	108.20
Type + Bombload	53.88	28.63	46.86	88.33
Type + Airexp	53.88	32.19	40.26	91.89
Bombload + Airexp	53.88	27.22	49.48	86.92
Type + Bombload + Airexp	53.88	25.95	51.84	87.65

El modelo que presenta mejor ajuste al conjunto de datos de acuerdo a su AIC=86.90 es el modelo que considera a la variable «Bombload», esto debido a que es el valor más bajo entre todos los valores de AIC. (Véase Ntzoufras (2009))

El modelo a ser considerado es:

$$damage_i = \beta_0 + \beta_1 bombload_i, \quad i = 1, 2, ...,30 \tag{4.1}$$

En el cuadro N° 4.3 se presenta los estimadores de los coeficientes de regresión para el modelo de la ecuación (4.1).

Cuadro 4.3: Estimación del número de daños encontrados en las aeronaves para el modelo «Bombload» mediante el Modelo de Regresión Poisson con enlace Log lineal

Coeficiente	**Estimación**	**Error Estándar**	**z value**	$Pr(>\|z\|)$
(Intercept)	-1.70097	0.50685	-3.356	0.000791
bombload	0.23112	0.04677	4.942	7.72e-07

Se define desvianza nula como la desviación para el modelo que tiene solo la constante, la desvianza residual es la desviación del modelo que tiene la constante y la variable Bombload con valores 53,883 y 29,206 respectivamente. La diferencia entre los dos valores tiene una distribución chi-cuadrado con 29 grado de libertad. Determinado por Cayuela (2011) sobre la variabilidad, el modelo explica:

$$\begin{aligned} D &= \frac{DesvianzaNula - DesvianzaResidual}{DesvianzaNula} \times 100 \\ &= \frac{53{,}883 - 29{,}206}{53{,}883} \times 100 \\ &= 45{,}79 \end{aligned}$$

El modelo dado en la ecuación (4.1) para la regresión Poisson con enlace logaritmo explica un 45,79 % el número de daños debido a la carga en el avión, asimismo se observa que las variables son significativas en la estimación.

Diagnóstico para el modelo de la ecuación (4.1) mediante el Modelo de Regresión Poisson:

Seleccionado el modelo, se procede a validar el MLG, asumiendo una familia Poisson y se realizan gráficos de diagnóstico. El modelo explica el número de daños respecto a la carga en el avión. (Ver figura N°4.2).

Considerando el análisis de diagnósticos en la figura N°4.2 a) se presenta los valores $\hat{h}_{ii}$ en cualquiera de los 8 grupos y se puede observar que destaca un punto. En la Figura N°4.2 b) se denota al menos 2 puntos con mayor influencia en $\hat{\beta}$ destando el punto 25. De la figura N°4.2 c) se muestra la influencia del punto 25 encontrandose fuera de la banda. Por lo tanto existe evidencia de observaciones influyentes en el ajuste.

Ajustando el modelo sin el punto 25, en la figura N°4.3 se sigue observando otro punto 29 como influyente, el gráfico de distancia de Cook y el grafico de análisis de residuos se observan varios puntos fuera de la banda, notándose que el modelo no mejora a pesar de eliminar un punto influyente, lo que confirma que el Modelo de Regresión de Poisson no ajusta convenientemente a los datos.

Figura 4.2: Diagnóstico para el modelo de la ecuación (4.1) mediante el Modelo de Regresión Poisson con enlace log Lineal

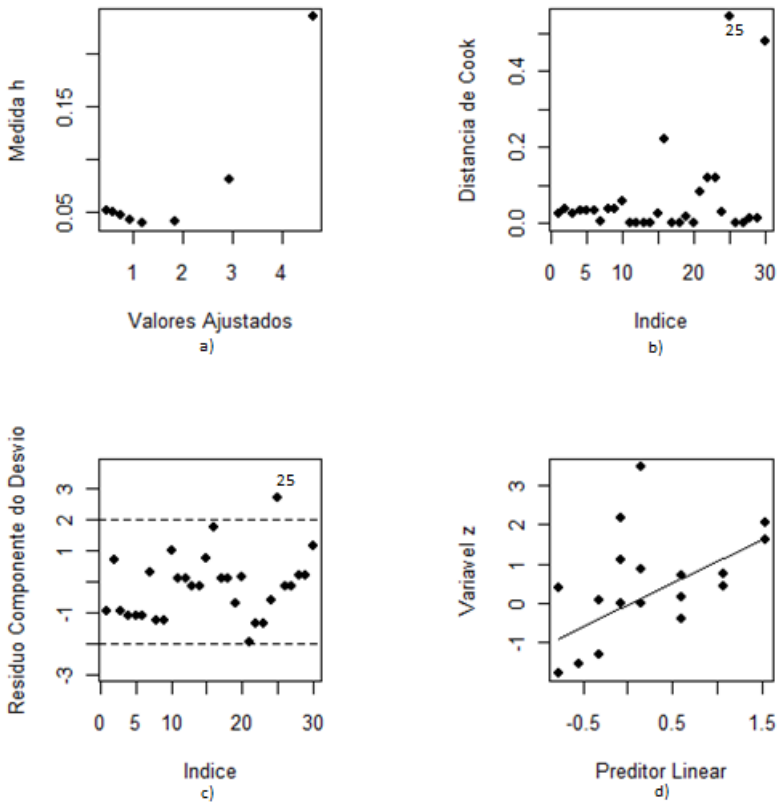
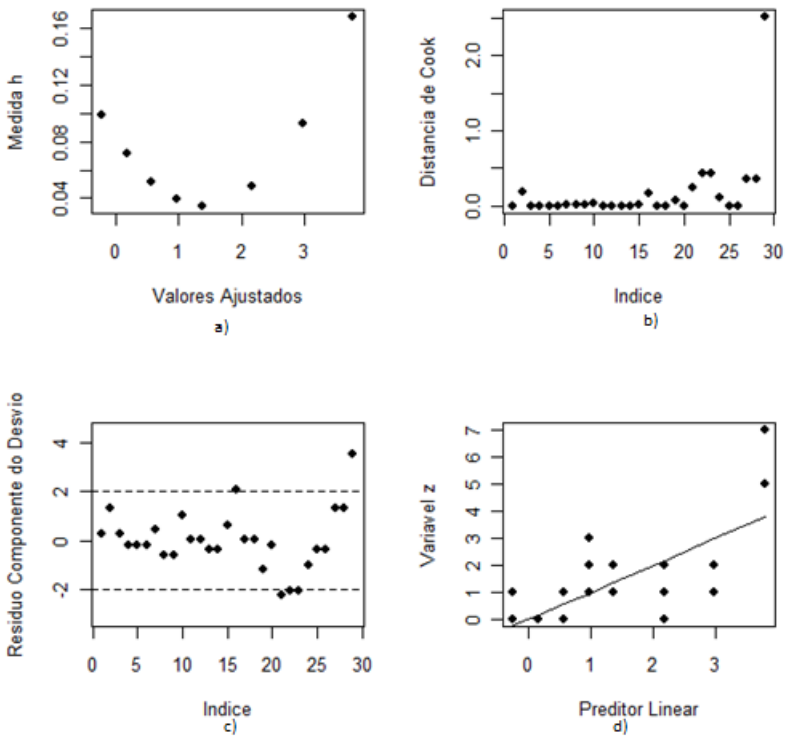


Figura 4.3: Diagnóstico para el modelo de la ecuación (4.1) sin el punto 25 mediante el Modelo log Lineal de la Regresión Poisson



4.1.3. Modelo de Regresión Binomial Negativa para datos The Aircraft Damage

Como se ha indicado, el modelo de Regresión Binomial Negativa es adecuado cuando los datos cumplen todos los requisitos del modelo de Poisson y además presentan sobredispersión, evaluamos este modelo para los datos The Aircraft Damage.

Cuadro 4.4: Estimación de los números de daños encontrados en las aeronaves para el modelo de la ecuación (4.1) mediante el Modelo de Regresión Binomial Negativa con enlace log Lineal

Coeficiente	**Estimación**	**Error Estándar**	**z value**	$Pr(>\|z\|)$
(Intercept)	-1.70093	0.50689	-3.356	0.000792
bombload	0.23112	0.04677	4.942	7.75e-07

Definido anteriormente al modelo de la ecuación (4.1) se presenta en el cuadro N° 4.4 la estimación mediante el modelo Regresión Binomial Negativa con enlace logaritmo, presentando un valor AIC=86.902 y la variabilidad del modelo es determinada mediante el desvío explicada:

$$\begin{aligned} D &= \frac{53{,}875 - 29{,}202}{53{,}875} \times 100 \\ &= 45{,}80 \end{aligned}$$

El modelo explica un 45,80 % el número de daños debido a la carga en el avión. Asimismo la variable resultó ser significativa para la estimación del modelo «Bombload».

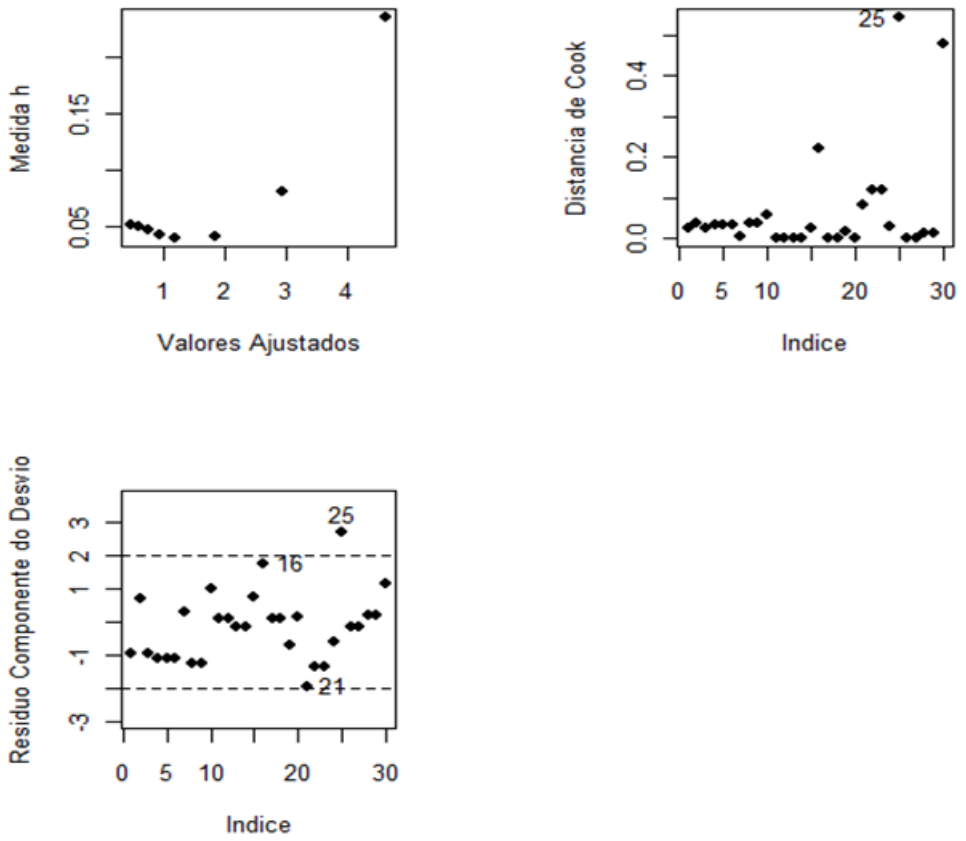
Diagnóstico para el modelo de la ecuación (4.1) mediante el Modelo de Regresión Binomial Negativa:

Se procede a validar el MLG, asumiendo un Modelo de Regresión Binomial Negativa con gráficos de diagnóstico. El modelo explica el número de daños respecto a la carga en el avión. (Ver figura N°4.4).

Mediante el análisis de diagnóstico en la figura N°4.4 a) del gráfico de $\hat{h}_{ii}$ indica que existe una observación con alto leverage y que podría ser una observación influyente. En la figura N°4.4 b) se denonta al menos 3 puntos con mayor influencia en $\hat{\beta}$ siendo el punto 25 con mayor presencia. De la figura N°4.4 c) de los residuales con bandas simuladas se confirma la presencia de datos atípicos como el punto 25.

Ajustando el modelo eliminando el punto 25, en la figura N°4.5 se observa que todavía existe un punto influyente en el gráfico de la distancia de Cook en el gráfico de residuos no se puede determinar que existe puntos aberrantes observándose en el gráfico los datos dentro de la banda. El modelo mejora notablemente retirando el punto 25.

Figura 4.4: Diagnóstico para el modelo de la ecuación (4.1) mediante el Modelo de Regresion Binomial Negativa con enlace log Lineal



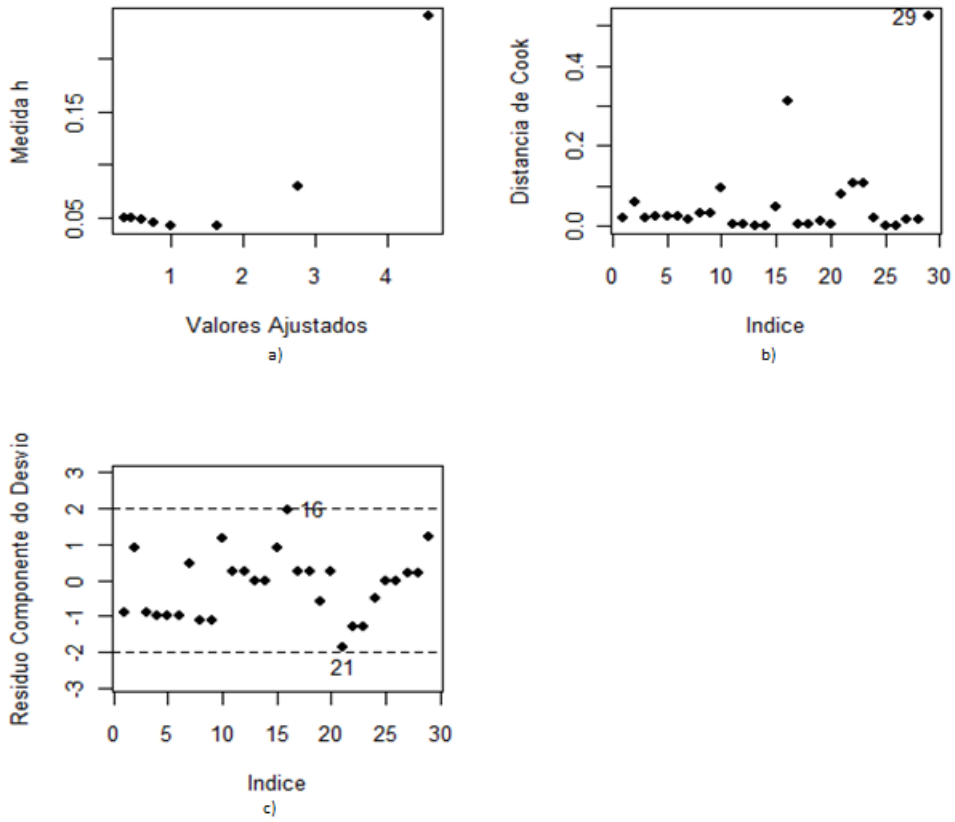
Cuadro 4.5: Comparación final entre ambos modelos de regresión para el modelo de la ecuación (4.1), sin el punto 25

	Regresión de Poisson			**Regresión Binomial Negativa**		
Variable	**Estimación**	**Error Estándar**	$Pr(>\|z\|)$	**Estimación**	**Error Estándar**	$Pr(>\|z\|)$
(Intercept)	-1.84259	0.62592	0.00659	-2.03263	0.55492	0.000249
bombload	0.40185	0.07244	7.03e-06	0.25408	0.04046	366.e-07
AIC		95.3.14			77.937	
Desvío Exp.		53.26			55.42	

Se observa en los cuadros N° 4.3 y 4.4 que los dos modelos ajustado son muy similares y presenta igual porcentaje de desvío para los Modelo de Regresión Poisson y el Modelo de Regresión Binomial Negativa.

Sin embargo se puede observar el cuadro N° 4.5, en relación a los modelos de Regresión Poisson y Binomial Negativa estimados para datos de conteo, se ajusta mejor el MRBN eliminando el valor 25, que se determinó como un valor atípico. El Modelo de Regresión

Figura 4.5: Diagnóstico para el Modelo de la ecuación (4.1) eliminado la etiqueta 25 - Modelo log Lineal de Regresión Binomial Negativa ajustado



Binomial Negativa con enlace logaritmo para los datos de conteo The Aircraft Damage específicamente el modelo de la ecuación (4.1) mejora notable presentando un AIC=77.93 y explicando el modelo un 55 %, aproximadamente, verificandose en la figura N°4.5. Los códigos utilizados para esta aplicación se presentan en el Apéndice B en el punto B.1.

Finalmente el modelo estimado de la ecuación (4.1) eliminando el punto 25 mediante la Regresión Binomial Negativa con enlace logaritmo es:

$$log(damage_i) = -2{,}03263 + 0{,}25408 bombload \quad i = 1, 2, ...,30$$

De la ecuación, se puede interpretar que, por cada aumento en una unidad de la variable carga de bombas en aviones, el número de daños ubicados en las aeronaves aumenta en 0.25408 unidades.

4.2. Aplicación en Resultados Electorales

Para el análisis de resultados electorales, se utilizan los datos del número de votos alcanzado por un determinado candidato en una circunscripción electoral y se opta por el modelo de Regresión Poisson, teniendo en cuenta además los factores de las variables que influyen en la determinación del candidato electo. En ésta oportunidad se hará uso de la base de datos de los resultados electoral de las Elecciones Generales Presidenciales del 2011 en el Perú.

El estudio se centrará en la influencia del contexto social en la elección de un candidato, entendiéndose como covariables el ingreso promedio percapita, porcentaje de mujeres analfabetas, Indice de desarrollo humano, etc. Los datos corresponden a las 25 regiones del país.

4.2.1. Definición y descripción de las variables

En el cuadro 4.6 define las variables con las que se desarrollará el modelo alcanzado por un determinado candidato en una circunscripción electoral y sus 16 variables explicativas.

Cuadro 4.6: Variables de Datos Electorales Peruanos considerados en la aplicación a nivel de Regiones

Código	Nombre de Variable	Descripción
y	Voto Hum	Votos obtenido por Ollanta Humala Tasso
x_1	Pob 11	Total de Población Estimada a Junio de 2011
x_2	P11 65	Población Estimada a Junio de 2011 mayores de 65 años
x_3	Ele Hab	Números de electores hábiles
x_4	Ele 65	Número de electores mayores de 65 años
x_5	PobRura	Población en el área rural
x_6	Quint	Índice de carencias - Quintil
x_7	SinAgua	Población sin agua
x_8	SinDesa	Población sin desagüe
x_9	SinElec	Población sin electricidad
x_{10}	TasaAnaf	Mujeres analfabetas
x_{11}	Nino0 12	Niño entre 0 a 12 años
x_{12}	TasaDes	Tasa de desnutrición. Niños de 6-9 años
x_{13}	IndDesHu	Índice de Desarrollo Humano (IDH) 2007
x_{14}	Ing Per	Ingreso Promedio Percapital Mensual (Nuevos Soles)
x_{15}	Sever	Severidad (FGT2)
x_{16}	GiniDes	Coeficiente Gini

Descripción de las variables:

- **Voto Hum**: «Votos obtenidos por el candidato Ollanta Humala» en la primera vuelta de las elecciones Generales y Parlamento Andino 2011, realizado el 10 de abril de 2011, la cual fue convocada por la Presidencia del Consejo de Ministro - PCM con DS Nº 105-2010-PCM de fecha 05 de diciembre de 2010, para elegir Presidente de la República, Vicepresidentes, Congresistas y representantes peruanos ante el Parlamento Andino

para el periodo 2011-2016, elaborado por Bazán, J. and Sulmont, D. and Calderón, A. and Millones, O. (2010).

El número obtenido en cada Región del Perú se procedió a dividir entre 10,000, una vez dividido se adecuó a números enteros para ser usado en datos de conteo.

- **Pob 11**: Total de Población Estimada a Junio de 2011. «Las proyecciones de población por provincias y distritos del país son derivadas de las proyecciones de población por departamento, obtenidas previamente. Uno de los modelos matemáticos empleados en demografía para analizar las tendencias del crecimiento de una población y de diversos indicadores demográficos es la función logística, la que fueron utilizadas para la estimación de la población» (Boletín Especial N°18 - INEI - 2009).
- **P11 65**: Población Estimada a Junio de 2011 de personas mayores de 65 años, proyección extraída del Total de población estimada a junio de 2011 realizada por el Instituto Nacional de Estadística e Informática - INEI.
- **Ele Hab.**: Número de personas mayores de 18 años, sin impedimento de votar para el proceso de las Elecciones Generales y Parlamento Andino 2011.
- **Ele 65**: Número de electores hábiles mayores de 65 años, sin impedimento de votar.

  Para el estudio de las covariables antes mencionadas se procedió igual que la variable en estudio, dividir entre 10,000, una vez dividido se adecuó a números enteros.
- **Quint**: Quintil. Representan a los más pobres por carencias (1= más pobres y 5=menos pobre). El primer quintil se llama «Más pobre», el segundo quintil «Quintil 2», el tercer quintil «Quintil 3», el cuarto quintil «Quintil 4» y el quintil 5 «Menos pobre».
- **SinAgua**: Porcentaje de población viviendas que carecen de agua potable.
- **SinDesa**: Porcentaje de la población que carecen de desagüe o letrinas.
- **SinElec**: Porcentaje de la población que carecen de electricidad.
- **TasaAnaf**: Porcentaje de mujeres analfabetas de 15 años y más.
- **Nino0 12**: Porcentaje de niños de 0 a 12 años de edad.
- **TasaDes**: Porcentaje de niños desnutridos de 6 a 9 años.
- **IndDesHu**: Índice de Desarrollo Humano (IDH) es un indicador del desarrollo humano por país, elaborado por el Programa de las Naciones Unidas para el Desarrollo (PNUD). Se basa en un indicador social estadístico compuesto por la esperanza de vida al nacer, el logro educativo y los ingresos, cada uno de los cuales está influenciado directa o indirectamente por los servicios provistos por el Estado.
- **Ing Per**: Ingreso Promedio Percapital Mensual - Nuevos Soles.

- **Sever**: Severidad. Es una medida de distribución del gasto en consumo entre los pobres respecto a la línea de pobreza. La estimación da una mayor ponderación a las distancias relativas de los más pobres, siendo que a mayor distancia mayor sea la severidad.
- **GiniDes**: Índice de Desigualdad Estimada - Coeficiente Gini. Ésta medida es estimada con los gastos deflactados; es decir con los gastos a precios de Lima Metropolitana (utilizando la relación del valor de la línea de pobreza total del área urbano y rural de cada departamento respecto al valor de la línea de Lima Metropolitana). Es igual a cero cuando el gasto total se distribuye por igual entre toda la población (plenamente equitativa) y es uno cuando una sola concentra dicho gasto (plenamente inequitativa).

4.2.2. Fuente de Información

Para el desarrollo del presente estudio, fue necesario crear una base de datos que contenga la información con las diferentes fuentes, que se describe a continuación:

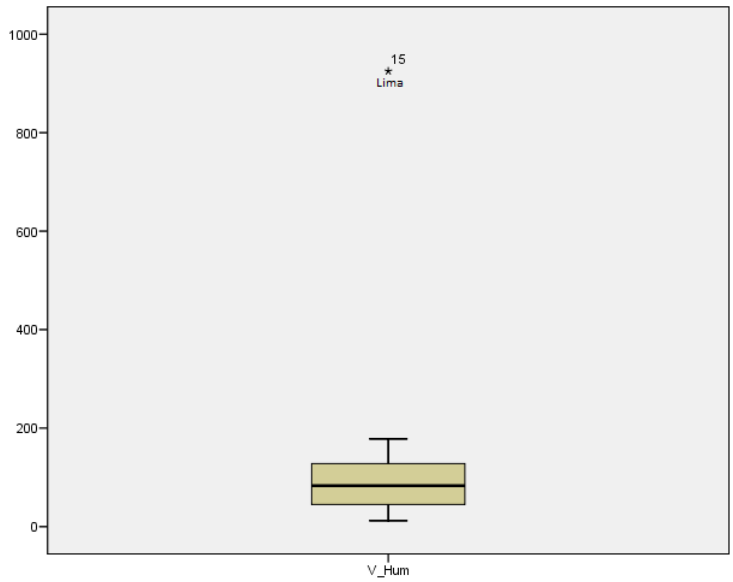
Los datos de la variable en estudio «Votos obtenidos por el candidato Ollanta Humala» **(Voto Hum**) corresponde a los resultados publicado por la Oficina Nacional de Procesos Electorales - ONPE, realizado el 10 de abril de 2011 para «Elecciones Generales y Parlamento Andino 2011», elaborado por (Bazán, J. and Sulmont, D. and Calderón, A. and Millones, O. (2010)).

Las siguientes covariables, como:

- Los datos de las variables **Pob 11 y Pob11 65** corresponde a los resultados del Censo 2007 y que es proyectada a junio de 2011 por el Instituto Nacional de Estadística e Informática - INEI.(Robles (2009))
- Los números obtenidos de las variables **Ele Hab y Ele 65** compete al padrón electoral de la Oficina Nacional de Procesos Electorales - ONPE, del proceso realizado el 10 de abril de 2011 para las elecciones Generales y Parlamento Andino 2011. (ONPE (2011))
- Los datos de las variables **PobRura, Quint, SinAgua, SinDesa, SinElec, TasaAnaf, Nino0 12, TasaDes y IndDesHu** pertenece a la publicación realizada por Foncodes con datos del Censo 2007.(Robles (2009))
- El Ingreso Promedio Percapita Mensual del Censo 2007, de la variable (**Ing Per**) es publicado por el Instituto Nacional de Estadística e Informática - INEI. (INEI (2007))
- La información de la variables **Sever y GiniDes** es analizada en la publicación sobre el enfoque de la pobreza monetaria divulgada por el Instituto Nacional de Estadística e Informática.(Díaz (2006))

En Apéndice A, se muestra la base de datos elaborado con los resultados obtenidos por el Candidato Ollanta Humala Tasso en las diferentes regiones del país en relación a las variables del contexto social, así como las covariables mencionadas.

Figura 4.6: Box Plot



4.2.3. Análisis Descriptivo preliminar

Previo al análisis de los Modelos Lineales Generalizados, se llevo a cabo un análisis exploratorio de los datos en estudio.

En figura 4.6 se puede apreciar la dispersión de los datos «Votos obtenidos por el candidato Ollanta Humala» en la diferentes regiones del Perú.

Además, la figura 4.6 muestra valores outlier, este dato atípico pertenece al Departamento de Lima. Removiendo el dato outlier, se realiza la prueba de ajuste del modelo para contrastar la hipótesis sobre la distribución de la variable «Votos emitido a favor del candidato Ollanta Humala», para las 24 regiones del país sin Lima mediante la prueba de Kolmogorov.

Esta prueba, sirve para contrastar la hipótesis de que la distribución de una variable se ajusta a una determinada distribución teórica de probabilidad, en nuestro caso se compara con la distribución Poisson, el estadístico de prueba es la máxima diferencia de:

$$D = max|F_n(x) - F_o(x)|$$

Donde $F_n(x)$ es la función de distribución muestral y $F_o(x)$ la función de distribución teórica.

Se desea comprobar si los «Votos obtenidos por el candidato Ollanta Humala» sigue una distribución Poisson, sobre la base de la Prueba de Kolmogorov.

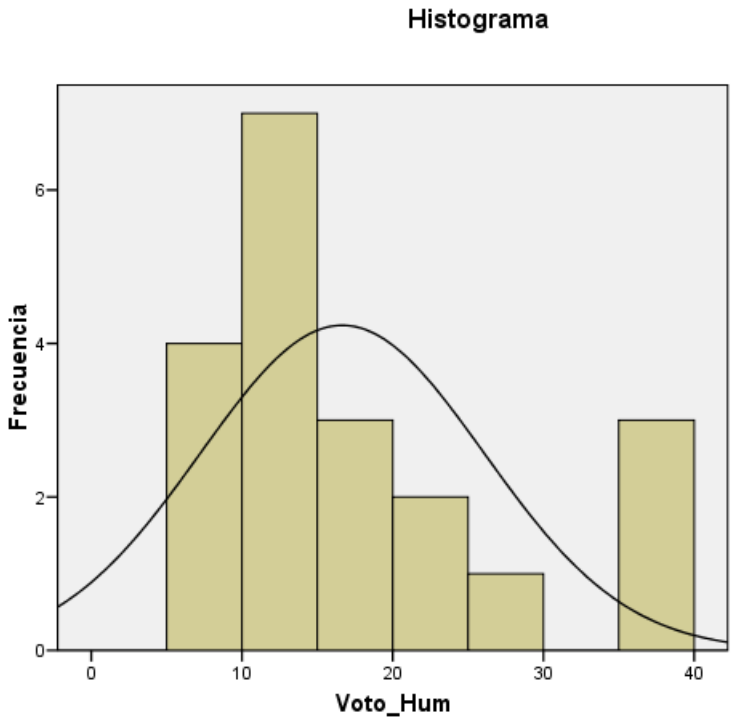
En el cuadro N° 4.7 se muestra los resultados de la prueba de Kolmogorov para las regiones del Perú sin considerar las regiones de Madre de Dios, Moquegua, Pasco y Tumbes,

Cuadro 4.7: Prueba de Kolmogorov-Smirnov para datos «Votos obtenido por el candidato Ollanta Humala»

		Votos Humala
Parámetro de Poisson(a,b)	Media	16.650
Diferencias más extremas	Absoluta	.302
	Positiva	.302
	Negativa	-.167
Z de Kolmogorov-Smirnov		1.351
Sig. asintót. (bilateral)		.052

a La distribución de contraste es la de Poisson.
b Se han calculado a partir de los datos.

Figura 4.7: Histograma - Números de votos obtenidos en las regiones del Perú



se verifica que no hay discrepancia entre los datos y la distribución Poisson, estas regiones fueron eliminadas sobre la base de un análisis del aporte de cada región, la distribución de los votos obtenidos por el candidato Humala en las regiones consideradas en el estudio, se muestra en el gráfico N° 4.7.

Análisis descriptivo para los «Votos obtenidos por el candidato Ollanta Humala» para las 20 regiones, la figura 4.7 muestra una leve asimetría a la derecha, debido a la concentración de votos en las regiones donde obtuvo mayor preferencia el candidato Ollanta Humala.

En el cuadro 4.8 se observan las estadísticas descriptivas de las variables relacionadas con «Votos obtenidos por el candidato Ollanta Humala» y que fueron descritas en el cuadro N° 4.6.

Cuadro 4.8: Estadística Descriptiva Preliminar para las variables relacionadas con los Votos obtenidos por el candidato Ollanta Humala

Var.	**Rango**	**Mín.**	**Máx.**	**Media**	**Desv. Típ.**	**Varianza**	**Asimetría**	**Error Típ.**
y	30	6	36	16.65	9.42	88.66	1.12	0.51
x_1	146	32	178	98.70	44.61	1,990.22	0.20	0.51
x_2	9	2	11	5.65	3.07	9.40	0.19	0.51
x_3	90	22	112	60.45	29.00	841.21	0.19	0.51
x_4	10	2	12	5.95	3.10	9.63	0.29	0.51
x_5	0.68	0	0.68	0.35	0.20	0.04	0.00	0.51
x_6	4	1	5	2.25	1.21	1.46	0.66	0.51
x_7	0.51	0.09	0.6	0.30	0.14	0.02	0.46	0.51
x_8	0.55	0.03	0.58	0.22	0.12	0.02	1.10	0.51
x_9	0.54	0.05	0.59	0.33	0.15	0.02	-0.06	0.51
x_{10}	0.3	0.02	0.32	0.15	0.09	0.01	0.39	0.51
x_{11}	0.11	0.23	0.34	0.29	0.04	0.00	-0.21	0.51
x_{12}	0.5	0	0.5	0.26	0.13	0.02	-0.16	0.51
x_{13}	0.2	0.5	0.7	0.58	0.06	0.00	0.25	0.51
x_{14}	470.87	144.74	615.61	336.07	128.92	16,621.36	0.72	0.51
x_{15}	29.1	0.7	29.8	7.90	6.61	43.72	1.93	0.51
x_{16}	0.14	0.26	0.4	0.34	0.04	0.00	-0.38	0.51

Donde se puede apreciar que el promedio de «Votos obtenido por el candidato Ollanta Humala» es 17*10,000 hab. aproximadamente, con una tendencia a variar de 9*10,000 hab. Asimismo se muestra que el valor mínimo obtenido es 6*10,000 hab. y como valor máximo 36*10,000 hab. Presenta asimetría positiva y variabilidad de 9.42. Además se aprecia que las variables x_9, x_{11}, x_{12} y x_{16} presentan asimetría negativa. (Variables descrito en el punto 4.2.1, del presente capítulo).

Como se ha mencionado, el supuesto fundamental para la aplicación correcta del modelo de regresión Poisson es que exista equidispersión. Para determinar la condición de equidispersión de la variable respuesta, se presenta a continuación el Indice de dispersión:

$$\begin{aligned} I_n &= \frac{S_y^2}{\overline{y}} \\ &= \frac{88{,}66}{16{,}65} \\ &= 5{,}32 \end{aligned}$$

De la ecuación anterior se puede observar que existe sobredispersión debido a que el coeficiente de variación es mayor a 1, violando uno de los supuestos fundamentales para el Modelo de Regresión Poisson, el cual asume que los valores de la media son iguales a los de la varianza. En consecuencia se espera que el modelo de Regresión Binomial Negativa ajuste mejor los datos que el modelo de Regresión Poisson.

En la sección 4.2.4 y 4.2.5 se presenta el análisis de los datos electorales considerando ambos modelos de regresión de conteo. En la sección 4.2.6 se realiza un resumen de la comparación de los Modelos de Regresión Poisson y el Modelo de Regresión Binomial Negativa.

4.2.4. Modelo de Regresión Poisson para los Votos obtenidos por el candidato Ollanta Humala

Como hemos indicado previamente, los datos presentan sobredispersión, no obstante para mostrar la metodológía del Modelo de Regresión Poisson, ignoraremos la sobredispersión y se estimará los parámetros mediante el método de Máxima Verosimilitud descrito en el Capítulo 3, para los datos de la aplicación.

Se denota Y_i como el números de «Votos obtenidos por el candidato Ollanta Humala» en la i-ésima región del Perú para las Elecciones Generales realizadas el 9 de abril del 2011.

Determinamos: $Y_i \sim P(\mu_i)$ como parte aleatoria.

Como parte sistemática:

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 + x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$$

donde $i = 1, ..., 20$

En este caso las variables x_1 al x_{16} son las descritas en el cuadro N° 4.5.

Variable Offset: Electores Hábiles

Los datos no son homogéneos entre los valores de las variables explicativas, por lo que se incluirá en el modelo una variable «offset» Electores hábiles.

La variable Electores hábiles actúa como una variable offset, esto es debido a que influye en la respuesta directamente, ya que es lógico asumir que a más electores, puede existir mayor cantidad de votos a favor del candidato Ollanta Humala. Los resultados son mostrados en el cuadro N° 4.9.

El cuadro 4.9 muestra que la única variable que no es sígnificativa es x_4, «Número de electores mayores de 65 años», esto significa que no tiene efecto sobre la variable en estudio, mediante la estimación con variable offset.

Cuadro 4.9: Estimación de los coeficientes para los «Votos obtenidos por el candidato Ollanta Humala» con variable offset, considerando un Modelo de Regresión Poisson

Coeficiente	**Estimación**	**Error Estándar**	**z value**	$Pr(>\|z\|)$
(Intercept)	-66.242253	8.463775	-7.827	5.01e-15
x_1	-0.493527	0.009553	-51.663	2e-16
x_2	-0.105106	0.238677	-0.440	0.65967
x_4	-0.892230	0.222422	-4.011	6.04e-05
x_5	-31.445007	3.696310	-8.507	2e-16
x_6	10.172386	0.990575	10.269	2e-16
x_7	-12.754609	1.561775	-8.167	3.17e-16
x_8	14.512762	2.356889	6.158	7.39e-10
x_9	59.380382	5.489817	10.816	2e-16
x_{10}	25.576785	5.526580	4.628	3.69e-06
x_{11}	292.525191	28.416331	10.294	2e-16
x_{12}	-12.336520	3.915368	-3.151	0.00163
x_{13}	-28.058842	3.025210	-9.275	2e-16
x_{14}	-0.025743	0.002637	-9.763	2e-16
x_{15}	0.319479	0.039471	8.094	5.77e-16
x_{16}	-91.215010	4.802671	-18.993	2e-16

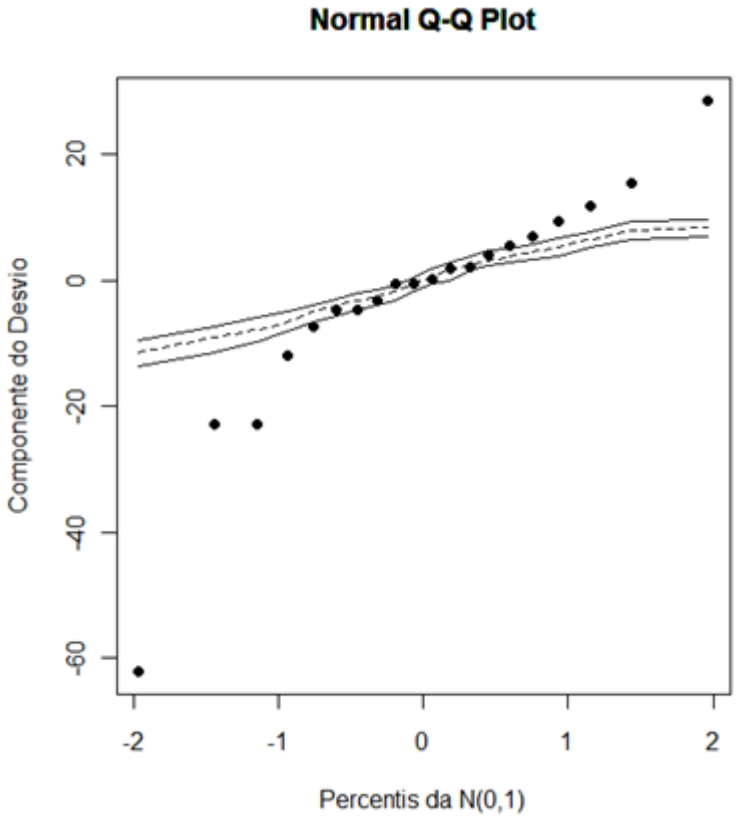
De la figura N°4.8 de probabilidad normal para los datos en estudio mediante la estimación con variable offset, se observa que agregando una variable offset no beneficia la predicción del modelo. Procediendo a evaluar los datos en estudio, utilizando el criterio AIC para la selección del mejor modelo.

Para evaluar posibles modelos alternativos se considera la diferentes covariables y se realizó un análisis mediante la función StepAIC, descrito en la sección 3.5, esta función, selecciona el modelo más apropiado para los «Votos obtenidos por el candidato Ollanta Humala», presentado en el cuadro N° 4.10.

La función StepAIC busca el modelo que describa adecuadamente los datos y tenga el mínimo AIC de variables regresoras, presentados en el cuadro 4.10.

Usando el criterio AIC que se utiliza para la selección del modelo más apropiado para los «Votos obtenidos por el candidato Ollanta Humala» de acuerdo al criterio de información de Akaike, se opta por aquel modelo con menor AIC= 112.99 entre los demás modelos presentado en el cuadro N° 4.10 para la regresión Poisson es el modelo N° X, que denominaremos a partir de ahora el modelo seleccionado, y cuyo valor AIC es el más pequeño comparando con los otros modelos como se puede observar en el cuadro N° 4.11.

Figura 4.8: Probabilidad Normal para residuos del Modelo Poisson para los Votos obtenidos por el Candidato Ollanta Humala con variable offset



Cuadro 4.10: Modelos encontrados para Votos obtenidos por el candidato Ollanta Humala

N°	Modelos
I	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}$ $+\beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
II	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} +$ $\beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
III	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} +$ $\beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
IV	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} +$ $\beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
V	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} +$ $\beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
VI	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} +$ $\beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
VII	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} +$ $\beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
VIII	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} +$ $\beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
IX	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_8 x_{i8} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{15} x_{i15} + \beta_{16} x_{i16}$
X	$\beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_8 x_{i8} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{16} x_{i16}$

Cuadro 4.11: Valores AIC de los modelos para los «Votos obtenidos por el candidato Ollanta Humala»

Modelo	Null Deviance	Residual Deviance	Función Desvío	AIC
I	92.234	3.279	96.445	127.79
II	92.234	3.280	96.444	125.79
III	92.234	3.287	96.436	123.80
IV	92.235	3.343	96.376	121.85
V	92.234	3.372	96.344	119.88
VI	92.234	3.688	96.001	118.20
VII	92.234	4.456	95.169	116.97
VIII	92.234	5.199	94.363	115.71
IX	92.234	5.559	93.973	114.07
X	92.234	6.483	92.971	112.99

Este modelo seleccionado se puede escribir como:

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_8 x_{i8} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{13} x_{i13} + \beta_{16} x_{i16} \tag{4.2}$$

Donde:
x_1: Total de Población estimada a junio de 2011
x_4: Número de electores mayores de 65 años
x_8: Población sin Desagüe
x_{10}: Mujeres analfabetas
x_{11}: Niños entre 0 y 12 años
x_{13}: Indice de Desarrollo Humano
x_{16}: Coeficiente de Gini

Cuadro 4.12: Estimación de los coeficientes para el modelo de la ecuación (4.2) mediante el Modelo de Regresión Poisson con enlace log lineal

Coeficiente	Estimación	Error Estándar	z value	$Pr(>\|z\|)$
(Intercept)	11.974152	1.956621	6.120	9.37e-10
x_1	0.010834	0.004674	2.318	0.0205
x_4	-0.126919	0.079018	-1.606	0.1082
x_8	1.157691	0.721404	1.605	0.1085
x_{10}	-3.220388	1.633975	-1.971	0.0487
x_{11}	-18.855574	4.538705	-4.154	3.26e-05
x_{13}	-9.587168	2.155329	-4.448	8.66e-06
x_{16}	4.506500	1.854887	2.430	0.0151

En el cuadro N° 4.12, se observa los parámetros estimados para el modelo seleccionado, con enlace logaritmo.

Se define desvianza nula como la desviación para el modelo que tiene solo la constante, la desvianza residual es la desviación del modelo que tiene la constante mas las covariables de la ecuación (4.2) con valores 92,2348 y 6,4834 respectivamente. La diferencia entre los dos valores tiene una distribución chi-cuadrado con 12 grado de libertad. Sobre la variabilidad, el modelo explica:

$$\begin{aligned} D &= \frac{92{,}2348 - 6{,}4834}{92{,}2348} \times 100 \\ &= 92{,}97 \end{aligned}$$

El modelo dado en la eduación (4.2) para la Regresión Poisson con enlace logaritmo explica aproximadamente un 93 % los resultados de los «Votos obtenidos por el candidato Ollanta Humala» en relación a sus covariables determinado.

Del cuadro N° 4.12 se observa que las variables total de Población estimada a junio de 2011 y Coeficiente de Gini, son significativas y positias, las variables Mujeres analfabetas, Niños entre 0 y 12 años e Indice de Desarrollo Humano, son significativas pero negativas, siendo las variables número de electores mayores de 65 años y Población sin desagüe no significativa para el coeficiente estimado, lo que indica un efecto nulo sobre la variable en investigación.

Diagnóstico para el modelo de la ecuación (4.2) mediante el Modelo de Regresión Poisson:

Para validar el modelo seleccionado, se realiza gráfico de diagnóstico para el modelo que explica el comportamiento de los «Votos obtenidos por el Candidato Ollanta Humala» variable: Voto Hum, en función a sus covariables determinadas en la ecuación (4.2) que se encuentra representada en la figura N°4.9.

En la figura N°4.9 a) se observa que el gráfico de punto leverage, representa más o menos una nube de puntos, lo cual demuestra normalidad. La figura N°4.9 b) de influencia presenta un dato aberrante, la figura N°4.9 c) de residuos se observa la etiqueta 4, siendo la Región de Arequipa, como dato que influye en el ajuste del modelo.

Realizado el ajuste para el modelo el modelo de la ecuación (4.2) sin Arequipa, se presenta en la figura N°4.10 una comparación de los gráficos probabilistico de normalidad que permite contrastar la normalidad de la distribución de los residuos del modelo de la ecuación (4.2) para la regresión Poisson, donde se aprecia que no mejora la predicción de los datos de la figura a) en relación a la figura b). Además la figura b) vemos que existe grandes desvíos con respecto a la diagonal q-q plot, por lo que no existe linealidad de los datos, determinando que no existe una mejora importante cuando se elimina Arequipa, cuando se ajusta con el modelo de Regreisón Poisson.

Los códigos correspondiente son mostrados en el Apéndice B en el punto B.2.

Figura 4.9: Diagnóstico para el modelo de la ecuación (4.2) mediante el Modelo de Regresión Poisson con enlace Log lineal

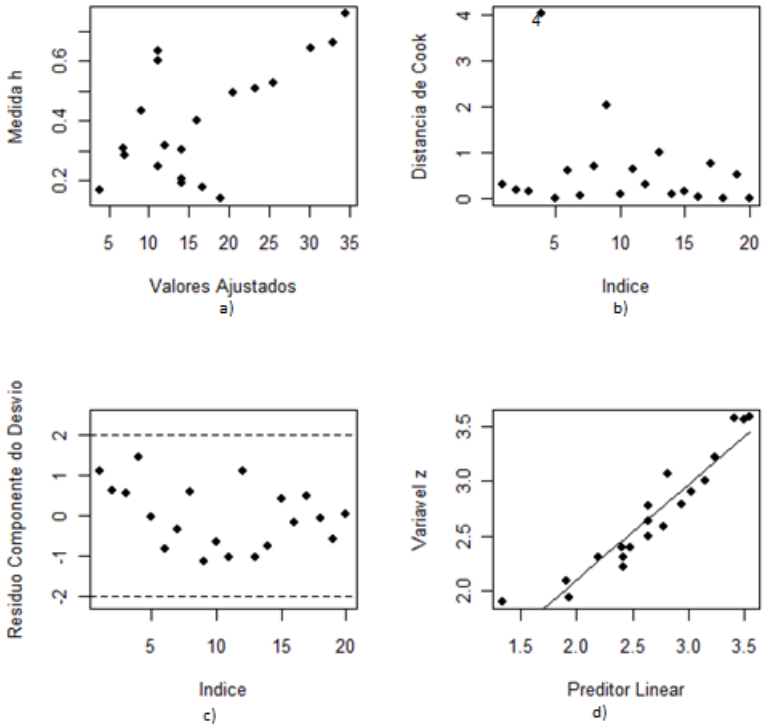
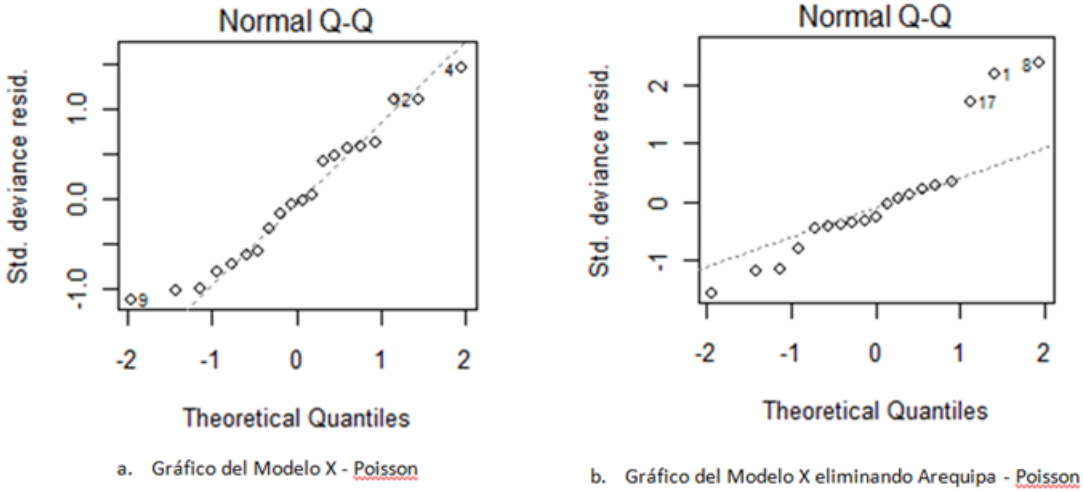


Figura 4.10: Comparación con Q-Q Normal del modelo de la ecuación (4.2) sin Arequipa mediante el Modelo de Regresión Poisson con enlace Log lineal



a. Gráfico del Modelo X - Poisson

b. Gráfico del Modelo X eliminando Arequipa - Poisson

4.2.5. Modelo de Regresión Binomial Negativa para los Votos obtenidos por el candidato Ollanta Humala

Como se ha indicado el Modelo de Regresión Binomial Negativa es adecuado cuando los datos cumplen todos los requisitos del modelo de Poisson pero adicionalmente muestran sobredispersión.

En esta sección esperamos verificar este resultado con los datos de la aplicación, para el modelo determinado en la ecuación (4.2)

La elección de la función enlace no siempre resulta fácil. En tal sentido existen diferentes funciones enlace aplicable para la Regresión Binomial Negativa. Para los datos en estudio se comparará dos funciones enlace para determinar cuál es el mejor enlace para la variable en estudio número de votos obtenidos por el Candidato Ollanta Humala.

Los enlaces que se utilizará para linealizar la relación entre la variable respuesta y sus covariables mediante la transformación de la variable respuesta para el Modelo de Regresión Binomial Negativa son:

- Identidad $\mu_i = \eta_i$
- Logaritmo $\log\mu_i = \eta_i$

Modelo de Regresión Binomial Negativa usando enlace identidad para el modelo de la ecuación (4.2)

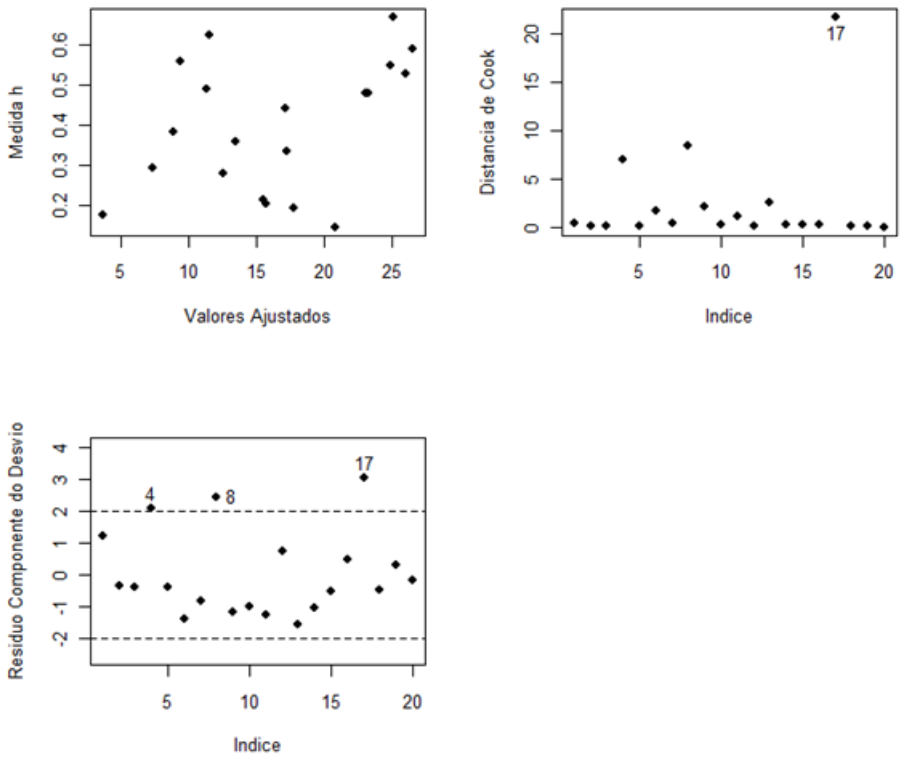
Ajustando un Modelo Lineal Generalizado (MLG) como la Binomial Negativa con la variable en estudio se tiene:

Cuadro 4.13: Estimación de los coeficientes mediante el Modelo Regresión Binomial Negativa con enlace Identidad

Coeficiente	**Estimación**	**Error Estándar**	**z value**	$Pr(>\|z\|)$
(Intercept)	87.42719	30.88033	2.831	0.00464
x_1	0.06065	0.08935	0.679	0.49731
x_4	0.21428	1.39170	0.154	0.87763
x_8	0.25174	12.07245	0.021	0.98336
x_{10}	-30.93932	24.33727	-1.271	0.20363
x_{11}	-124.35508	50.43381	-2.466	0.01367
x_{13}	-93.27954	39.42063	-2.366	0.01797
x_{16}	46.39993	37.50093	1.237	0.21598

En el cuadro 4.13 se muestra el criterio de evaluación AIC=128.06 para el modelo de la ecuación (4.2). Para determinar la variabilidad del modelo se tiene que la desvianza nulo

Figura 4.11: Diagnóstico del modelo de la ecuación (4.2) mediante el Modelo de Regresión Binomial Negativa con enlace Identidad



es la desviación del modelo que tiene la constante. Desvianza residual es la desviación del modelo que tiene la constante y las variables población mayores de 65 años estimada a junio de 2011, Número de electores mayoras de 65 años, Población sin desague, Mujeres Analfabetas, Niños entre 0 a 12 años, Índice de Desarrollo Humano (IDH) e Índice de Desigualdad. La diferencia entre los valores tiene una distribución chi-cuadrado de 7 grado de libertad y permite contrastar si el coeficiente de las variables puede considerarse nulo. El modelo de la ecuación (4.2) con enlace Identidad explicaria aproximadamente 79 % los «Votos obtenidos por el candidato Ollanta Humala» en relación a sus covariables.

Diagnóstico del modelo mediante el Modelo Regresión Binomial Negativa con enlace identidad para el modelo de la ecuación (4.2):

Se procede a realizar los gráficos de diagnóstico para el modelo de la ecuación (4.2), que se encuentra representada en la figura N° 4.11.

En la figura N°4.11 vemos que el gráfico de punto leverage, representa más o menos una nube de puntos, lo cual demuestra normalidad (Arriba izquierda), el gráfico de influencia presenta un dato aberrante (Arriba derecha), el gráfico residuos se observa las etiquetas 17, 8 y 4 como datos que influye en el ajuste. Por lo tanto existe evidencia de que el modelo no describe bien a los datos. Determinándose que el enlace identidad no ayuda a la predicción del modelo de la ecuación (4.2) para los «Votos obtenidos por el candidato Ollanta Humala».

Modelo Binomial Negativa usando enlace log lineal para el modelo de la ecuación (4.2)

Los Votos obtenidos por el candidato Ollanta Humala se caracterizan por los valores enteros positivos, lo que implica la incorporación de efecto multiplicativo y esto es expresado mediante la función de enlace logaritmo.

En el cuadro 4.14 se puede observar que el modelo de la ecuación (4.2) para la regresión Binomial Negativa con enlace log, explica aproximadamente 92,97 % los «Votos obtenidos por el candidato Ollanta Humala» en relación a las demás covariables, con un AIC=114.99.

Cuadro 4.14: Estimación de los coeficientes del Modelo de la ecuación (4.2) mediante el Modelo de Regresión Binomial Negativa con enlace log lineal

Coeficiente	**Estimación**	**Error Estándar**	**z value**	$Pr(>\|z\|)$
(Intercept)	11.974106	1.956656	6.120	9.38e-10
x_1	0.010834	0.004674	2.318	0.0205
x_4	-0.126918	0.079019	-1.606	0.1082
x_8	1.157684	0.721415	1.605	0.1086
x_{10}	-3.220368	1.633995	-1.971	0.0487
x_{11}	-18.855478	4.538764	-4.154	3.26e-05
x_{13}	-9.587125	2.155367	-4.448	8.67e-06
x_{16}	4.506471	1.854924	2.429	0.0151

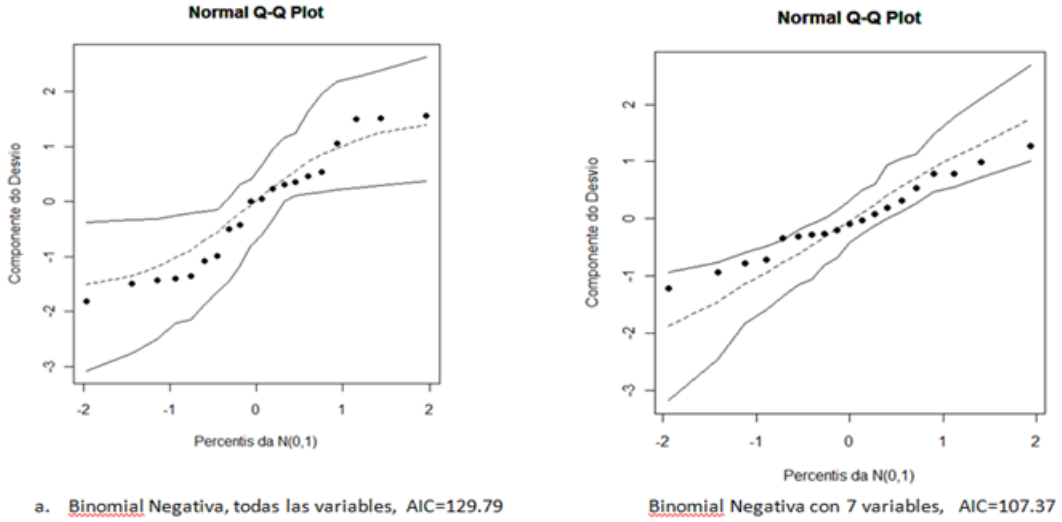
Diagnóstico del modelo de la ecuación (4.2) mediante el Modelo de Regresión Binomial Negativa con enlace log Lineal

Del modelo de la ecuación (4.1), se realiza los gráficos de diagnóstico, el modelo que explica los «votos obtenidos por el Candidato Ollanta Humala» en función a la población estimada para el año 2011, Número de electores mayoras de 65 años, Población sin desagüe, Mujeres Analfabetas, Niños entre 0 a 12 años, Indice de Desarrollo Humano (IDH) e Indice de Desigualdad, se encuentra representada en las figuras N°4.12 y N°4.13.

Se presenta una comparación de gráficos probabilístico de normalidad (Envelopes), observándose que el modelo final del grafico b) del modelo de la ecuación (4.2), mejora notablemente en relación al gráfico a). Asimismo, en la b) de la figura N°4.12 de probabilidad normal para el modelo de la ecuación (4.2) nos confirma que el modelo determinado ajusta mejor a los «Votos obtenidos por el candidato Ollanta Humala».

Sin embargo, realizando un diagnóstico de residuos se detectó un puntos leverage en el gráfico influencia, en la figura N°4.13 donde se aprecia la detección de un punto leverage que puede influir con en el ajuste del Modelo Lineal Generalizado Binomial Negativa con enlace log. Se muestra el punto 4 «Arequipa» como un posible atípico.

Figura 4.12: Probabilidad normal del modelo de la ecuación (4.2) mediante el Modelo de Regresión Binomial Negativa con enlace log lineal



a. Binomial Negativa, todas las variables, AIC=129.79

Binomial Negativa con 7 variables, AIC=107.37

Figura 4.13: Diagnóstico para el modelo de la ecuación (4.2) mediante el Modelo de Regresión Binomial Negativa con enlace Log lineal

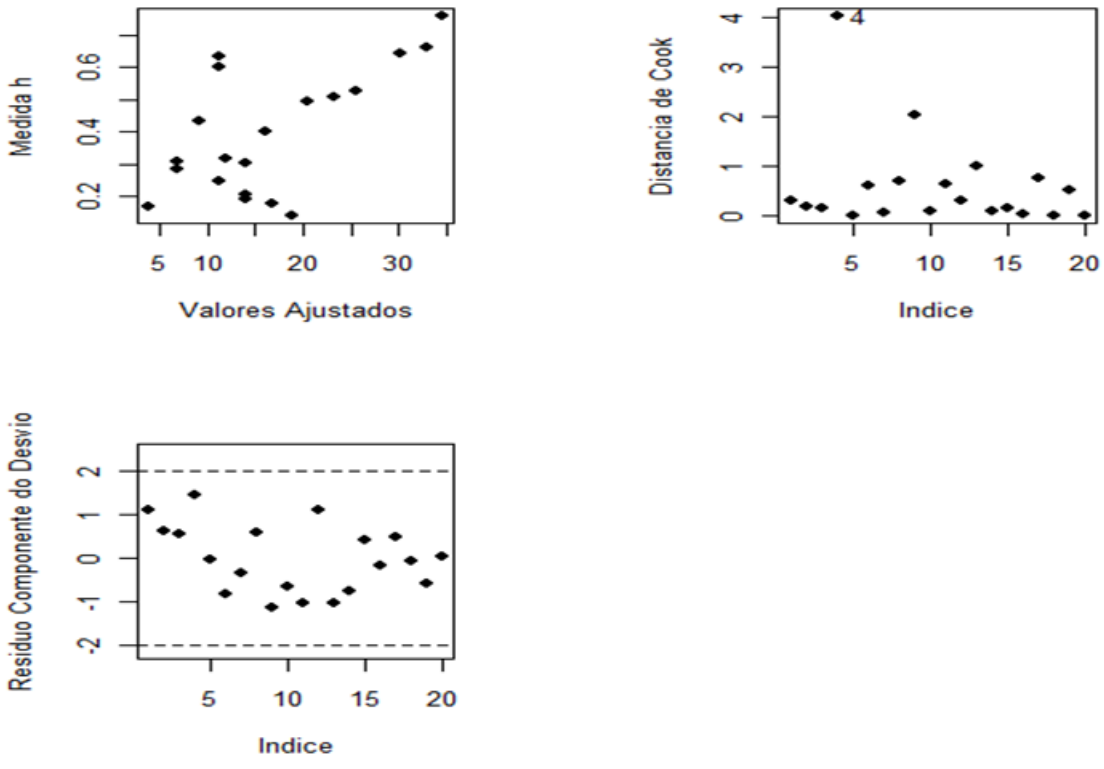
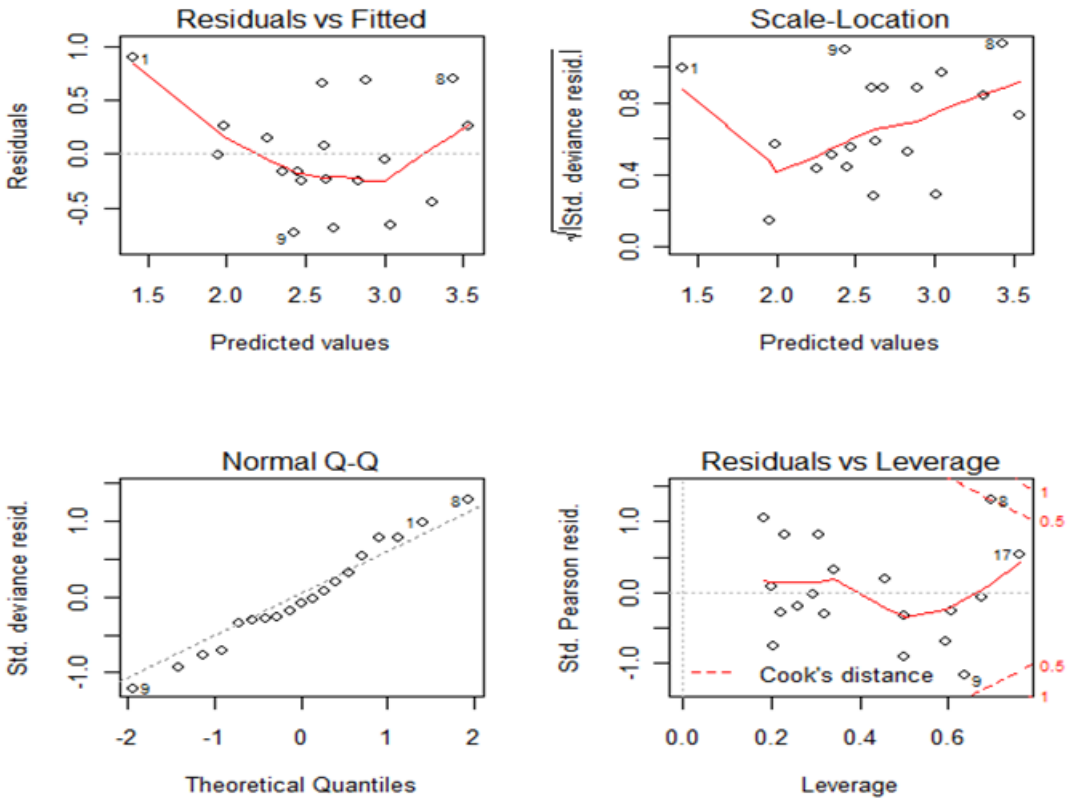


Figura 4.14: Análisis de Residuos del modelo de la ecuación (4.2) eliminando Arequipa mediante el Modelo de Regresión Binomial Negativa con enlace Log lineal



Retirando la etiqueta 4 «Arequipa» se observa que el modelo de la ecuación (4.2), estima mejor los «Votos obtenidos por el candidato Ollanta Humala», como se puede apreciar en figura N°4.14.

La figura N°4.14 izquierda, parte superior, vemos que los residuos estandarizados frente a los valores predichos representa una nube de puntos, lo cual indica normalidad.

Asimismo, la figura N°4.14 izquierda, parte inferior, vemos que no hay desvío muy grande respecto a la diagonal en el Q-Q plot, el gráfico probabilístico de normalidad nos permite contrastar la normalidad de la distribución de los residuo y nos confirmar la linealidad de los Votos del Candidato Ollanta Humala.

La figura N°4.14 derecha, parte inferior, vemos que no hay datos atípicos ni sobre-influyentes.

4.2.6. Resumen de la comparación del modelo de Regresión Poisson y Binomial Negativa para los Votos obtenidos por el candidato Ollanta Humala

El cuadro N° 15 muestra la comparación final entre ambos modelos.

Cuadro 4.15: Comparación final entre ambos modelos de regresión para el modelo de la ecuación (4.2), sin Arequipa

	Regresión de Poisson			**Regresión Binomial Negativa**		
Variable	**Estimación**	**Error Estándar**	$Pr(>\|z\|)$	**Estimación**	**Error Estándar**	$Pr(>\|z\|)$
(Intercept)	132.0689	38.4607	0.00558	11.048353	2.079881	1.08e-07
x_1	0.2506	0.1085	0.04130	0.015454	0.005646	0.006195
x_4	-2.8637	1.6937	0.11898	-0.187898	0.090696	0.038290
x_8	23.4318	14.0956	0.12464	1.380982	0.732420	0.059361
x_{10}	-23.4912	31.3431	0.46929	-2.035610	1.824025	0.264422
x_{11}	-235.4111	73.0171	0.00810	-18.043970	4.622857	9.49e-05
x_{13}	-109.5384	46.6349	0.03857	-8.045477	2.410364	0.000844
x_{16}	14.7726	43.4793	0.74043	2.930108	2.151347	0.173202
AIC		120.29			107.37	
Desvío Exp.		81.78			94.39	

La estimación de los parámetros del modelo de Regresión Poisson y el Modelo de Regresión Binomial Negativa tienen similares valores de predicción e igual porcentaje de desvío explicada tal como se observa en los cuadros N° 4.11 y 4.14.

Sin embargo, se puede apreciar en el cuadro N° 4.15, eliminando Arequipa, resulta para el Modelo de Regresión Poisson que indica un AIC= 112.99 y para el Modelo de Regresión Binomial Negativa que indica un AIC= 107.37, se muestra que el modelo más adecuado entre los dos antes mencionados, es el Modelo de Regresión Binomial negativa eliminando Arequipa con enlace logaritmo, para la variable «votos obtenido por el candidato Ollanta Humala». Además se puede observar que el Modelo de Regresión Binomial Negativa con enlace logaritmo sin Arequipa explica mejor con un 94 % los «Votos obtenidos por el candidato Ollanta Humala» en relación a las demás covariables.

El modelo final obteniendo por máxima verosimilitud los parámetros alternativo para el modelo de la ecuación (4.2) sin Arequipa, se muestra:

$$Log(VotoHum_i) = 11{,}04 + 0{,}02x_{i1} - 0{,}19x_{i4} + 1{,}39x_{i8} - 2{,}04x_{i10} - 18{,}04x_{i11} - 8{,}05x_{i13} + 2{,}93x_{i16}$$

Del modelo de la ecuación (4.2) sin Arequipa, ajustando para el modelo de Regresión

Binomial Negativa de enlace logaritmo se aprecia para la variable en estudio número de «votos obtenido por el Candidato Ollanta Humala», respecto a las variables Población estimadas a junio de 2011, así también Número de electores mayores de 65 años, Niños entre $0-12$ años e Índice de Desarrollo Humano son significativos $Pr(>|z|)$. El Intercepto y la variable Población estimadas a junio de 2011 son significativo y positivo. Esto nos indica que estas variables incrementan la posibilidad de votar por el candidato Humala. Las variables Número de electores mayores de 65 años, Niños entre $0-12$ e Índice de desarrollo Humano también son significativas pero negativas, es decir disminuyen con el aumento de votos para el candidato Ollanta Humala. Sin embargo las variables Mujeres analfabetas, coeficidente de Gini y Población sin desague no son significativas, lo que indica un efecto nulo sobre la variable «votos obtenido por el candidato Ollanta Humala».

Capítulo 5

Conclusiones y Recomendaciones

5.1. Conclusiones

- En la aplicación de los datos «The Aircraft Damage», donde se desea predecir el número de daños encontrados en las aeronaves durante la guerra de Vietnam, se pudo determinar que el mejor modelo es aquel que considera solamente la variable «Bombload» y que este modelo explica alrededor del 55.42 % de la variabilidad dentro de un Modelo Binomial Negativa con enlace logaritmo.

- Para el análisis de datos sobre resultados electorales se deben tener varias consideraciones sobre datos de conteo. Adicionalmente es importante determinar si existe sobredispersión o no (varianza mayor que la media) a fin de decidir convenientemente por un modelo adecuado.

- Para este estudio se elaboró una base de datos propia acerca de resultados electorales peruanos del 2011 a partir de diferentes fuentes de información, los cuales se presentan en el Apéndice A.

- Para los datos analizados, donde se intenta modelar el número de votos del candidato Ollanta Humala en cada una de las regiones del país, en función de un conjunto de predictores se encontró que el mejor modelo es aquel que presenta las covariables Población estimadas a junio de 2011, así también Mujeres Analfabetas, Niños entre $0-12$ años, Índice de Desarrollo Humano e Índice de Desigualdad explican el 94 % de la varianza, dentro de un modelo Binomial Negativa. Entre los factores identificados positivo son el Intercepto, la variable Población estimadas a junio de 2011 e Índice de Desigualdad y los factores o covariables identificado como negativos o de efecto inverso, identificamos a Mujeres analfabetas, Niños entre $0-12$ e Índice de desarrollo Humano.

- El modelo de Regresión Poisson resulta adecuado cuando no hay evidencia de sobredispersión. Si existe sobredispersión y se usa, es posible que se eliminen covariables que realmente si son significativas, como se puede observar en las aplicaciones analizadas.

- El Modelo de Regresión Binomial Negativa resulta ser más adecuado para datos que presentan sobredispersión, de acuerdo a las aplicaciones descritas.

- La librería **glm2** y **MASS** del paquete **R** implementan el método de Máxima Verosimilitud convenientemente tanto para la Regresión Poisson como para la Regresión Binomial Negativa.

5.2. Recomendaciones

- Presentar y desarrollar la Inferencia Bayesiana de los Modelos presentados.
- Extender el estudio para el análisis de la votación de otros candidatos del proceso electoral analizado.
- Realizar un modelo para otro tipo de circunscripción electoral por ejemplo provincias, distritos, o al interior de un departamento.
- Análizar otros procesos electorales y eventualmente medir modelos de Regresión Poisson y Binomial Negativa de efecto mixto ó de multinivel.

Apéndice A

Datos Electorales

Cuadro A.1: Datos Electorales Parte I: Votación de Ollanta Humala en la Elección Presidencial de 2011 de la Primera Vuelta a Nivel Regional y Covariables Asociadas

REGIÓN	Voto Hum	Pob 11	P11 65	Ele Hab	Ele 65	Pob Rura	Quint	Sin Agua
AMAZONAS	6	42	2	23	2	0.6	1.0	0.5
ÁNCASH	16	112	8	74	8	0.4	3.0	0.2
APURÍMAC	8	45	3	24	3	0.5	1.0	0.4
AREQUIPA	35	123	9	89	10	0.1	4.0	0.1
AYACUCHO	14	66	3	37	4	0.4	1.0	0.4
CAJAMARCA	18	151	8	89	9	0.7	1.0	0.3
CALLAO	11	96	6	65	6	0.0	5.0	0.2
CUSCO	35	128	8	78	7	0.4	2.0	0.3
HUANCAVELICA	9	48	2	25	3	0.7	1.0	0.6
HUÁNUCO	12	83	4	45	4	0.6	1.0	0.5
ICA	13	76	5	52	6	0.1	3.0	0.1
JUNÍN	21	131	7	79	7	0.3	3.0	0.3
LA LIBERTAD	20	177	11	112	12	0.2	3.0	0.2
LAMBAYEQUE	16	122	8	78	8	0.2	3.0	0.1
LORETO	10	100	3	54	4	0.3	1.0	0.4
PIURA	25	178	10	111	10	0.3	2.0	0.3
PUNO	36	137	9	78	9	0.5	2.0	0.3
SAN MARTÍN	11	80	3	47	3	0.4	2.0	0.4
TACNA	10	32	2	22	2	0.1	4.0	0.1
TUMBES	7	47	2	27	2	0.2	2.0	0.3

Descripción de las variables y de su unidad de medida:

- **Voto Hum**: Votos obtenido por Ollanta Humala Tasso. Número de personas * 10,000.
- Pob 11: Total de Población Estimada a Junio de 2011. Número de personas * 10,000.
- P11 65: Población Estimada a Junio de 2011 mayores de 65 años. Número de personas * 10,000.
- Ele Hab: Números de electores hábiles. Número de personas * 10,000.
- Ele 65: Número de electores mayores de 65 años. Número de personas * 10,000.

- PobRura: Población en el área rural. Porcentaje.
- Quint: Índice de carencias - Quintil
- SinAgua: Población sin agua. Porcentaje.

Cuadro A.2: Datos Electorales Parte II: Votación de Ollanta Humala en la Elección Presidencial de 2011 de la Primera Vuelta a Nivel Regional y Covariables Asociadas

REGIÓN	Sin Desa	Sin Elec	Tasa Anaf	Nino 0 12	Tasa Des	Ind DesHu	Ing Per	Sever	Gini Des
AMAZONAS	0.2	0.5	0.2	0.3	0.3	0.6	236.7	8.4	0.34
ÁNCASH	0.3	0.2	0.2	0.3	0.3	0.6	350.3	6.1	0.36
APURÍMAC	0.2	0.4	0.3	0.3	0.4	0.5	199.1	11.3	0.31
AREQUIPA	0.1	0.1	0.1	0.2	0.1	0.6	494.7	2.1	0.35
AYACUCHO	0.3	0.4	0.3	0.3	0.4	0.5	224.1	13.8	0.36
CAJAMARCA	0.2	0.6	0.3	0.3	0.4	0.5	218.4	10.7	0.36
CALLAO	0.0	0.0	0.0	0.2	0.1	0.7	615.6	1.1	0.29
CUSCO	0.3	0.3	0.2	0.3	0.3	0.5	270.4	10.8	0.40
HUANCAVELICA	0.6	0.4	0.3	0.3	0.5	0.5	144.7	29.8	0.40
HUÁNUCO	0.3	0.6	0.2	0.3	0.4	0.5	236.2	11.5	0.35
ICA	0.1	0.2	0.0	0.2	0.1	0.6	418.3	0.7	0.26
JUNÍN	0.2	0.3	0.1	0.3	0.3	0.6	318.2	5.0	0.33
LA LIBERTAD	0.2	0.3	0.1	0.3	0.2	0.6	414.2	5.5	0.40
LAMBAYEQUE	0.1	0.2	0.1	0.3	0.2	0.6	350.2	3.8	0.33
LORETO	0.3	0.4	0.1	0.3	0.3	0.6	308.0	7.8	0.38
PIURA	0.3	0.3	0.1	0.3	0.2	0.6	335.2	6.4	0.37
PUNO	0.4	0.4	0.2	0.3	0.3	0.5	226.5	13.3	0.29
SAN MARTÍN	0.1	0.4	0.1	0.3	0.2	0.6	293.9	6.3	0.35
TACNA	0.1	0.1	0.1	0.2	0.0	0.7	551.4	2.5	0.33
TUMBES	0.2	0.3	0.1	0.3	0.2	0.6	515.2	1.1	0.28

Descripción de las variables y de su unidad de medida

- SinDesa: Población sin desagüe. Porcentaje.
- SinElec: Población sin electricidad. Porcentaje.
- TasaAnaf: Mujeres analfabetas. Tasa.
- Nino0 12: Niño entre 0 a 12 años. Porcentaje.
- TasaDes: Tasa de desnutrición. Niños de 6-9 años. Tasa.
- IndDesHu: Índice de Desarrollo Humano (IDH) 2007. Indice.
- Ing Per: Ingreso Promedio Percapital Mensual (Nuevos Soles). Promedio.
- Sever: Severidad (FGT2). Porcentaje.
- GiniDes: Coeficiente Gini. Indice.

Apéndice B

Programa en R

B.1. Programa para los datos The Aircraft Damage

a. Modelo de Regresiòn Poisson:

```
require (glm2)

gPD<-glm2(formula = damage ~ type + bombload + airexp,
family=poisson(link = "log"))
fit.model = gPD
stepAIC(fit.model)

gPDD<-glm2(formula = damage ~ type, family=poisson(link = "log"))
fit.model = gPDD
source("diag_pois.txt")

summary(gPDD)

gPDD25<-glm2(formula = damage ~ bombload, subset=-25)
fit.model = gPDD25
source("envel_pois.txt")
```

b. Modelo de Regresiòn Binomial Negativa:

```
require (MASS)

gBND<-glm.nb(formula = damage ~ bombload, link = log)
fit.model = gBND
source("diag_pois.txt")

summary(gBND)

gBND25<-glm.nb(formula = damage ~ bombload, subset=-25)
fit.model = gBN25
source("diag_nbin.txt")
 identify(fitted(fit,model),h,n=3)
```

B.2. Programa para la aplicación de datos Electorales

a. Modelo de Regresiòn Poisson:

```
gPO<-glm2(formula = Voto_Hum ~ Pob_11 + P11_65 + offset(Ele_Hab)
+ Ele_65 + PobRura + Quint + SinAgua + SinDesa + SinElec + TasaAnaf
+ Nino0_12 + TasaDes + IndDesHu + Ing_Per + Sever + GiniDes, family=poisson)

summary(gPO)

gP<-glm2(formula = Voto_Hum ~ Pob_11 + P11_65 + Ele_Hab + Ele_65 + PobRura
+ Quint + SinAgua + SinDesa + SinElec + TasaAnaf + Nino0_12 + TasaDes
+ IndDesHu + Ing_Per + Sever + GiniDes, family=poisson(link = "log"))
fit.model = gP
stepAIC(fit.model)

gPX<-glm2(formula = Voto_Hum ~ Pob_11 + Ele_65 + SinDesa + TasaAnaf
+ Nino0_12 + IndDesHu + GiniDes, family=poisson(link = "log"))
fit.model = gPX
source("envel_pois.txt")

gPX<-glm2(formula = Voto_Hum ~ Pob_11 + + Ele_65 + SinDesa + TasaAnaf
+ Nino0_12 + IndDesHu + GiniDes, subset=-4)
fit.model = gPX
source("diag_pois.txt")
```

b. Modelo de Regresiòn Binomial Negativa:

Enlace Identidad

```
gBNi<-glm.nb(formula = Voto_Hum ~ Pob_11 + Ele_65 + SinDesa + TasaAnaf
+ Nino0_12 + IndDesHu + GiniDes, link =identity)
fit.model = gBNi
source("diag_nbin.txt")
```

Enlace Log

```
gBNX<-glm.nb(formula = Voto_Hum ~ Pob_11 + Ele_65 + SinDesa + TasaAnaf
+ Nino0_12 + IndDesHu + GiniDes, link=log)
fit.model = gBNX
source("diag_nbin.txt")
gBNX4<-glm.nb(formula = Voto_Hum ~ Pob_11 + Ele_65 + SinDesa + TasaAnaf
+ Nino0_12 + IndDesHu + GiniDes,subset=-4)
fit.model = gBNX4
source("diag_nbin.txt")

par(mfcol=c(2,2))
plot(gBNX4)
```

Bibliografía

Akaike, H. (1974). *A new look at statistical model identification.*

Atkinson, A. C. (1985). *Plots, Transformations and Regressions*, Oxford Statistical Science Series, Oxford.

Bazán, J. and Sulmont, D. and Calderón, A. and Millones, O. (2010). *Modelos de Regresión en el Intervalo Unitario con aplicaciones en el análisis de resultados electorales*, Lima, Perú. Proyecto DGI 20100173,.

Cameron, A. y Trivedi, P. (1986). *Econometric models based on count data:comparisons and applications of estimators and tests*, Journal of Applied Econometrics.

Cameron, A. y Trivedi, P. (1998). *Regression Analysis of Count Data*, Cambridge University Press.

Cayuela, L. (2011). *Modelos lineales generalizados (MLG)*, Universidad Rey Juan Carlos, Madrid.

Cook, R. D. y Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

Díaz, J. (2006). *Nuevo Mapa de Pobreza*, Fondo de Cooperación para el Desarrollo Social - FONCODES.

INEI (2007). *Censo 2007*, http://www.inei.gob.pe.

Jong, P. y Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*, Cambridge.

Krzanowki, W. (1998). *An introduction to Statistical Modelling*, Arnold.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley, New York.

Lindsey, J. (1995B). *Modelling Frequency and Count Data*, Clarendon Press.

McCullagh, P. (1987). *Tensor Methods in Statistics*, Chapman and Hall, London.

McCullagh, P. y Nelder, J. A. (1991). *Generalized Linear Models*, Chapman & Hall.

Montgomery, D. (2006). *Desing and Analysis of Experiments*, Wiley, Hoboken, NJ.

Nelder, J. A. y Wedderburn, R. W. (1972). *Generalized Linear Models*, Journal of The Royal Statistical Association.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*, John Wiley & Sons.

ONPE (2011). *Padrón Electoral*, http://www.onpe.gob.pe.

Paula, G. A. (2010). *Modelos de Regressao*, Universidade de Sao Paulo.

Pregibon, D. (1981). *Logistic regression diagnostics*, Annals of Statistics 9,705-724.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robles, M. (2009). *Mapa de Pobreza Provincial y Distrital 2007, El enfoque de la pobreza monetaria*, Instituto Nacional de Estadística e Informática.

Winkelmann, R. (2000). *Econometric Analysis of Count Data*, Springer-Verlag.