FACULTAD DE CIENCIAS
INSTITUTO DE ESTADÍSTICA

MODELOS DE REGRESIÓN POISSON Y BINOMIAL NEGATIVA EN MODELOS LINEALES GENERALIZADOS APLICADOS A DATOS CORRESPONDIENTES A ACCIDENTES DE TRÁNSITO Y LESIONADOS ENTRE LOS AÑOS 2011 Y 2014 EN LA REGIÓN DE VALPARAÍSO.

Trabajo de titulación para optar al grado de Licenciado en Estadística y al título profesional de Ingeniero Estadístico

AUTOR
Paulina Alejandra López Vásquez

PROFESOR GUÍA
Claudia Navarro Villarroel, PhD

VALPARAÍSO, CHILE
2016

Agradecimientos

Primero que todo quiero agradecerle a mi mamá, la persona más importante en mi vida, sin ella yo no podría haber terminado la carrera. Gracias por siempre alentarme cuando más lo necesité, por darme besos y abrazos cuando estaba cansada de estudiar, porque a pesar de todo siempre estuviste ahí para hacerme sentir mejor.

Agradezco también a mi familia, ya que siempre confiaron en mí y me hicieron sentir que era capaz de todo con esfuerzo y dedicación.

Agradecerle a mi pololo Adrian, que siempre me apoyó en mi camino universitario desde que me cambié de carrera. Sé que estas orgulloso de mí como yo lo estoy de tí. Gracias por siempre pensar que era la mejor, ya que eso me dio fuerzas para seguir adelante. Te amo con el corazón.

No podía dejar de agradecerle también a un ex profesor que tuve en la universidad: profesor Freddy López Quintero. Siempre tuvo buena disposición para responderme cada vez que le preguntaba algo, lo que fuese, siempre se dio el tiempo de contestarme de la mejor manera. De verdad muchas gracias.

Gracias a todos mis profesores de la universidad por la paciencia que me tuvieron muchas veces, especialmente a mi profesora guía Claudia Navarro, al profesor Alberto Caro, profesor Germán Ibacache, profesor Carlos Felipe, profesor Víctor Leiva, profesor Harvey Rosas, profesora Mónica Catalán, etc. Muchas gracias por todos los conocimientos que me dieron estos 5 años que fui parte de la familia Estadística.

Para terminar quiero dedicarle todo esto a mi pequeña princesa, mi sobrina Emilia. Decirle que la amo con todo mi corazón y que todo lo que hago es para que ella se sienta orgullosa y feliz de mí.

Resumen

Los modelos lineales generalizados (MLG) planteados por Nelder y Wedderburn (McCullagh y Nelder, 1991) nacen a partir de la necesidad de formular de manera cuantitativa relaciones entre un grupo de variables, donde una de ellas es llamada variable respuesta y las demás son llamadas covariables. Los MLG admiten que la variable respuesta Y, provenga de la familia exponencial como por ejemplo: distribución binomial, normal, poisson, binomial negativa, entre otras, de manera de consolidar a los modelos con variables de respuesta categórica y numérica.

En estadística, el modelo lineal generalizado (MLG) es una generalización flexible de la regresión lineal ordinaria. Donde se relaciona a la distribución aleatoria de la variable dependiente (la "función de distribución") con la parte sistemática (no aleatoria) (o "predictor lineal") mediante una función llamada la "función de enlace".

El modelo de regresión poisson (MRP) es un modelo lineal generalizado, el cual se utiliza en estudios de variables de conteo. Este modelo se ha usado en distintas áreas de investigación y es apropiado para cuando los datos no presentan sobredispersión, es decir, cuando la varianza muestral es igual a la media (Contreras, 2012). Se aclara que existe sobredispersión en los datos cuando la varianza presentada es mucho mayor que la que presenta el modelo. Por el contrario, está el modelo de regresión binomial negativa (MRBN) el cual, en la mayoría de los casos, se piensa como un modelo alterno al modelo de regresión poisson ya que se utiliza cuando existe sobredispersión en los datos.

Abstract

The generalized linear models (GLM) proposed by Nelder and Wedderburn (McCullagh and Nelder, 1991) born from the need to formulate quantitatively relationships among a group of variables, where one of them is called response variable and the others are called covariates. The GLM admit that the variable response Y, it should come from the exponential family as for example: distribution binomial, normal, poisson, binomial negative, between others, of way of consolidating the models with variables of categorical and numerical response.

In statistics, generalized linear models (GLM) is a flexible generalization of the linear ordinary regression. Where it relates to the random distribution of the dependent variable (the "function of distribution") to the systematic part (not random) (or "predictor linear") by means of a function called the "function of link".

The Poisson regression model (PRM) it is a generalized linear model which is used in studies of count variables. This model has been used in different areas of investigation and is appropriate for when data is not present overdispersion, ie, when the sample variance is equal to the average (Contreras, 2012). It clarifies that the overdispersion on data submitted when the variance is much greater that presented by the model. on the contrary, it is the negative binomial regression model (NBRM) which in most cases, is intended as an alternative model to the Poisson regression model as it is used when there overdispersion on data.

Abreviaturas y Notación

A continuación se presentan algunas abreviaturas que se utilizan a lo largo de este trabajo.

MLG	Modelo Lineal Generalizado
MRP	Modelo de Regresión Poisson
MRBN	Modelo de Regresión Binomial Negativa
ANOVA	Análisis de Varianza (Analysis of variance)
MV	Máxima Verosimilitud
MCG	Mínimos Cuadrados Generalizados
MCP	Mínimos Cuadrados Ponderados
ELIO	Estimador Lineal Insesgado y Óptimo
EMV	Estimador Máximo Verosímil
EMC	Estimador Mínimos Cuadrados
EMCP	Estimador Mínimos Cuadrados Ponderados
AIC	Criterio de Información Akaike
BIC	Criterio de Información Bayesiano
CV	Coeficiente de Variación
RV	Razón de Verosimilitud
BN	Binomial Negativa
ACP	Análisis de Componentes Principales

Índice general

Objetivos

Los objetivos de este trabajo de titulación se presentan a continuación.

Objetivo general

El objetivo principal de este trabajo es estudiar y analizar los modelos de regresión Poisson y Binomial Negativa pertenecientes a los modelos lineales generalizados y aplicarlos a datos reales.

Objetivos específicos

(i) Estudiar las propiedades de los modelos de Poisson y Binomial Negativo (modelos de conteo).

(ii) Demostrar e implementar los métodos de estimación tradicional para los modelos Poisson y Binomial Negativo.

(iii) Aplicar los modelos de regresión de conteo para analizar resultados con datos reales.

Preliminares

A principios de los años 90 se confirma el aumento de los accidentes de tránsito en el país, así como también el número de muertos, lesionados y daños materiales que éstos conllevan. Desde el año 1987 al año 1992, los accidentes de tránsito aumentaron de 32.790 a 43.402 (32 % de aumento en 5 años) es más, en el mismo período aumentó el número de personas fallecidas por estos accidentes pasando de 1.198 a 1.700 (42 % aumento).

Frente a este escenario, surge la necesidad de confeccionar políticas públicas orientadas a solucionar los problemas de seguridad de tránsito. Para llevar a cabo lo anterior y debido que a la fecha no existía ningún organismo que se ocupara de esto directamente, es que como parte de la Política Nacional de Seguridad de Tránsito (PNST), se planteó crear un comité de ministros en esta materia (Comisión Nacional de Seguridad de Tránsito) y como cabecera designar una secretaría ejecutiva permanente para así apoyar el trabajo del comité de ministros (Programa CONASET, 2004).

La Comisión Nacional de Seguridad de Tránsito (CONASET) se formó el año 1993 como una comisión asesora del presidente de la república de Chile.

Los miembros de la CONASET son:

- Ministerio de Transportes y Telecomunicaciones
- Ministerio Interior y Seguridad Pública
- Ministerio Secretaría Gral. de Presidencia
- Ministerio Secretaría Gral. de Gobierno
- Ministerio de Educación
- Ministerio de Justicia

- Ministerio de Obras Públicas
- Ministerio de Salud
- Ministerio de Vivienda y Urbanismo
- Ministerio del Trabajo y Previsión Social
- Carabineros de Chile

Esta comisión tiene como propósito disminuir los accidentes de tránsito y sus consecuencias, controlando todos los posibles factores de riesgo mediante cambios en la normativa y mejorando la infraestructura. De esta manera, se busca promover prácticas eficientes de convivencia vial entre los usuarios y enfatizar la importancia de la seguridad vial, ampliando el conocimiento de los fenómenos del tránsito.

CONASET elabora y publica informes anuales a nivel nacional, por distrito y circunscripciones, por regiones y provincias. Además realiza focalización comunal estadística y gracias a la recopilación de datos obtiene informes temáticos, cuadros gráficos, registra y lleva un seguimiento de fallecidos y lesionados, y obtiene información sobre los accidentes que ocurren en las vías urbanas e inter urbanas de hasta 24 horas ocurrido el suceso.

Con respecto a la formación de los conductores, la comisión nacional de seguridad de tránsito ha evolucionado en colocar en todo el territorio nacional el examen teórico para así obtener licencia de conducir y encabeza el cambio que quieren lograr para el nuevo examen práctico que se tomará a los postulantes.

Conforme al decreto 223 del 27 de diciembre de 1993, a la comisión le corresponde asesorar a la presidencia de la república en cuanto a disminuir la gran cantidad de accidentes de tránsito que ocurren en el país. Para lo anterior, es que se deben considerar todas las materias relacionadas a la seguridad de tránsito.

Carabineros de Chile son los encargados de controlar y fiscalizar el tránsito en rutas, carreteras y caminos, tanto urbanos como interurbanos de todo el territorio nacional. Esta destacada institución entrega los datos estadísticos de todo el país, para su posterior elaboración a cargo de esta comisión.

Existen diversas causas que provocan un accidente. Dentro de las más frecuentes están:

1. Alcohol y Conducción.
2. Cinturón de Seguridad.
3. Exceso de Velocidad.

1. **Alcohol y Conducción, CONASET (2016).**

El consumo de alcohol afecta la capacidad que tienen las personas para poder efectuar una serie de acciones motoras. Al momento de conducir un vehículo se requiere de precisión, dependiendo en gran medida, de las habilidades de cada individuo, de los reflejos y de la capacidad de tomar decisiones rápidas. El tiempo de reacción de una persona que ha consumido alcohol se puede ver reducido en un 10 % a 30 % en comparación con una persona que no ha consumido. Es más, la visión se ve interrumpida volviéndose borrosa y las nociones de distancia, velocidad y peligro se estropean.

Lamentablemente en Chile, cercano al 20 % de los accidentes fatales de tránsito se debe al descarado consumo de alcohol de los conductores. En los últimos años, el ministerio de transportes y telecomunicaciones, mediante la CONASET, ha logrado avances significativos en cuanto a las leyes de alcohol y conducción. Por ejemplo, en marzo de 2012 se implementó la "Ley Tolerancia Cero" y en septiembre de 2014 entró en vigencia la "Ley Emilia".

A consecuencia de esto, en el transcurso del año 2014, se obtuvieron los mejores resultados de los últimos 13 años en lo que respecta a fallecidos en accidentes de tránsito que tengan relación con el alcohol. Los fallecidos por esta causa disminuyeron a nivel nacional en un 31 %, pasando de 205 víctimas en 2011 a 148 en los años 2012 y 2013, y a 142 fallecidos en el año 2014.

Figura 1: CONASET, extraído desde http://www.conaset.cl/.

- Ley tolerancia cero alcohol.

Esta ley entró en vigencia en marzo de 2012 como trasformación de la ley de tránsito. Esta iniciativa legal disminuyó los grados de alcohol permitidos en la sangre para poder conducir. Estableciendo nuevos rangos, es decir se considera "estado de ebriedad" cuando se tiene 0,8 gramos por litro de sangre y "bajo la influencia del alcohol" cuando se tiene 0,3 gramos por litro de sangre. De forma simultánea se aumentaron las sanciones referentes a la suspensión de la licencia de conducir, dependiendo de la infracción y de las consecuencias que ésta tenga, siendo mucho más rigurosas que en ley anterior.

Figura 2: CONASET, extraído desde http://www.conaset.cl/.

- Ley Emilia.

La ley Emilia es un complemento de la ley tolerancia cero. Esta ley nace en base a una petición ciudadana y lleva el nombre de Emilia por la menor fallecida en manos de un conductor irresponsable que llevaba alcohol en el cuerpo. El nombre de la pequeña era Emilia Silva Figueroa. Con esta ley, la cual está vigente desde el 16 de septiembre de 2014, se sanciona con cárcel efectiva de al menos un año a los conductores en estado de ebriedad que generen lesiones graves gravísimas o la muerte. Además, con esta nueva ley se decreta como un delito fugarse del lugar del accidente y negarse a realizar el alcohotest o la alcoholemia.

Penas de Cárcel Ley Emilia

Figura 3: CONASET, extraído desde http://www.conaset.cl/.

Sanciones de fuga

Figura 4: CONASET, extraído desde http://www.conaset.cl/.

¿Cuál es la diferencia entre la ley tolerancia cero y la ley Emilia?

- La ley tolerancia cero disminuyó los gramos de alcohol lícitos en la sangre para conducir y aumentó las sanciones de suspensión de la licencia de conductor.
- La Ley Emilia sanciona con cárcel efectiva de al menos 1 año a los conductores en estado de ebriedad que generan lesiones graves gravísimas o la muerte.

Figura 5: CONASET, extraído desde http://www.conaset.cl/.

2. **Cinturón de Seguridad, CONASET (2016).**

En Chile La ley de tránsito, entró en vigencia el 1 de enero de 1985 e instauró la obligatoriedad del uso del cinturón de seguridad en los asientos delanteros, por primera vez. Inicialmente, no existía precedente alguno que tratase de esto en el país. Hoy en día está más evidenciado que nunca: ¡El uso del cinturón te salva la vida!

Una de las mayores causas de riesgo para todo aquel que viaje en vehículos motorizados, es el viajar sin el cinturón de seguridad. Los traumatismos craneoencefálicos son las lesiones más usuales en los choques frontales. Es más, si una persona sale eyectada del vehículo tiene cinco veces más probabilidades de morir que aquélla que permanece en su interior.

El uso del cinturón de seguridad disminuye el riesgo de lesión mortal del conductor y de los pasajeros de los asientos delanteros en un 50 % y de los pasajeros de los asientos traseros en un 70 %. Todo lo anterior, según el informe sobre el estado de la seguridad vial en la región de las américas de la Organización Panamericana de la Salud (OPS).

3. **Exceso de Velocidad, CONASET (2016).**

Una de las faltas más irresponsables en el tránsito al momento de conducir es el exceso de velocidad. En Chile, alrededor del 30 % de las víctimas fatales se deben a manejar a gran velocidad o también a la pérdida de control del vehículo. En los últimos 10 años hubo más de 4.000 fallecidos debido a estas causas.

El Exceso de velocidad es un factor que por lo general no suele ser visto, por la mayoría, como un problema o un riesgo para la vida de las personas. Sin embargo, no deja de ser peligroso ya que actúa como un factor perjudicial en un accidente de tránsito, acrecentando sus consecuencias, convirtiéndose, por lo general, en accidentes fatales.

Finalmente, se puede observar que día a día son varios los ministerios encargados de sobrellevar el tema de la seguridad de tránsito del país. Pero cabe destacar que la mayor responsabilidad la tienen las personas que van al volante, como se mencionó anteriormente las causales que llevan a posibles accidentes de tránsito podrían evitarse si las personas toman conciencia de lo que pueden causar al manejar con alcohol en el cuerpo, al no respetar los límites de velocidad establecidos, al no utilizar el cinturón de seguridad etc. Es por eso el llamado a ser más responsables y consientes en la vía pública, porque no solamente peligra la vida de los demás sino que peligra en mayor magnitud su propia vida.

Introducción

En la mayoría de los casos cuando la variable respuesta sigue una distribución Normal se suele utilizar modelos de regresión lineal clásico. La ventaja de modelar datos a partir de los modelos lineales generalizados es que no necesariamente se debe partir con la restricción de una variable respuesta con distribución Normal, sino que pueden ampliarse las distribuciones a la variable respuesta a partir de distribuciones pertenecientes a la familia exponencial.

Los recuentos son definidos como la cantidad de sucesos/eventos los cuales, en una misma unidad de observación, ocurren en un determinado intervalo de tiempo específico. A partir de esta definición se desglosan dos fundamentales características de las variables de recuento las cuales son: naturaleza discreta y no negatividad, características que las diferencian de las variables cuantitativas continuas.

La distribución Poisson nace gracias al matemático francés Simeon Denis Poisson (1781-1840) el cual en el año 1837 mostró en un trabajo de investigación una distribución nueva para poder calcular probabilidades centradas en el ámbito penal. Esta distribución ya había sido presentada en 1718 por Abraham De Moivre (1667-1754) como una manera límite de la distribución Binomial la cual surge a partir de observar un evento extraño luego de una gran cantidad de repeticiones. La distribución binomial negativa se da a conocer en un estudio que realizó Pierre Rémond de Montmort (1678-1719) acerca de los juegos de azar en el año 1714, pero Blaise Pascal fue quien la definió (1623-1662). Posteriormente, la distribución Binomial Negativa fue presentada como una alternativa a la distribución Poisson para poder modelar el número de ocurrir un determinado suceso/evento cuando los datos presentan sobredispersión. (Epidat, 2014)

El Modelo de Regresión Poisson y el Modelo de Regresión Binomial Negativa son utilizados cuando se tienen variables de recuento, éstas están presentes en distintos ámbitos de la investigación como por ejemplo en el ámbito de las Ciencias Sociales o en las Ciencias de la Salud. Es por esta razón, que se encuentran investigaciones aplicadas con variables de recuento en áreas como Farmacología (Lindsey, Jones y Jarvis, 2001), Criminología (Osgoos, 2000), Relaciones Laborales (Sturman, 1999) entre muchos otros. Existen disciplinas en las cuales debiesen considerarse a parte, ya que no solamente describen una amplia aplicación de investigaciones con variable de recuento sino que han hecho excelentes aportes en el método estadístico en cuanto a este tipo de variables. Estas disciplinas que contienen variables de recuento son Medicina (Biggeri, Marchi, Lagazio, Martuzzi y Bohning, 2000; Navarro, Utzet, Puig, Caminal y Martín, 2001), Ciencias Políticas (King, 1998) y Ciencias Económicas (García-Crespo, 2001; Meliciani, 2000)

En cuanto a las variables de recuento, una estrategia frecuente es la transformación de los datos y el objetivo es generalmente la aplicación de los MLG. En este caso, Winkelmann (2000) hace mención que aunque la regresión lineal es usada generalmente como una herramienta de la regresión multipropósito, "para datos discretos en general, y para recuentos en particular, la regresión lineal normal presenta ciertos problemas que hace que su uso sea dudosa y lógicamente insatisfactorio".

Como hipótesis del trabajo se tiene que: El modelo Poisson y el modelo Binomial Negativa, son adecuados al momento de modelar datos de recuento de cualquier área.

Las primeras aplicaciones significativas de los modelos para este tipo de datos, son en el ámbito de la ciencia actuarial, bioestadística y demografía. Posteriormente, se aplicaron en campos tales como: economía, ciencias políticas y sociología, (Romero, Arcos, Cano y Sánchez, 2003)

Patil (1970) hace énfasis en diversas aplicaciones del análisis de datos de recuento; no obstante, "el enfoque metodológico adoptado en la mayoría de éstas queda fuera del contexto de la regresión" citado en (Romero, Arcos, Cano y Sánchez, 2003).

Capítulo 1

Modelos Lineales Generalizados

Como se mencionó anteriormente, los MLG planteados por Nelder y Wedderburn son una extensión de la teoría de los modelos lineales, agregando la posibilidad de modelar variables respuestas continuas o categóricas y donde no necesariamente se acepta normalidad en la variable respuesta, sino cualquier distribución que pertenezca a la familia exponencial.

1.1 Formalización del Modelo Lineal Generalizado (Contreras, 2012).

En un modelo lineal generalizado, se asume que la variable dependiente Y está generada a partir de una función de distribución de la familia exponencial. La media μ de la distribución depende de las variables independientes X mediante la fórmula:

$$E(Y) = \mu = g^{-1}(X\beta) \tag{1.1}$$

Donde

- $E(Y)$ corresponde al valor esperado de Y.
- $X\beta$ corresponde al predictor lineal, es decir, una combinación lineal de parámetros desconocidos β.
- g corresponde a la función de enlace.

Con esta notación, la varianza no es más que una función V de la media:

$$Var(Y) = V(\mu) = V(g^{-1}(X\beta)) \tag{1.2}$$

Es ventajoso si V proviene de una distribución que pertenezca a la familia exponencial, pero podría ser simplemente que la varianza sea una función del valor ajustado.

Los parámetros desconocidos β son, por lo general, estimados mediante máxima verosimilitud, máxima cuasi-verosimilitud, o técnicas de inferencia bayesiana.

A continuación se presentan los componentes de los MLG.

- **Componente Aleatoria**

  Dado Y_1,...,Y_n un conjunto de variables respuesta, que se identifica por los siguientes parámetros θ_i y ϕ, formará parte de la familia exponencial si sigue la siguiente forma:

$$f(y_i; \theta_i, \phi) = exp\left[\phi^{-1}\left\{y_i\theta_i - b(\theta_i)\right\} + c(y_i, \phi)\right] \tag{1.3}$$

- **Componente Sistemática**

  La componente sistemática de un modelo lineal generalizado describe a las variables explicativas que ingresan en forma de efectos fijos en un modelo lineal, la relación de las variables x_j se expresa de la siguiente manera.

$$\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad i = 1, \ldots, n \tag{1.4}$$

  Esta mezcla (lineal) formada por variables explicativas se conoce como *predictor lineal*. Dicho de otra forma, se puede generalizar como sigue a continuación

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{1.5}$$

  Donde

  - x_{ij} corresponde al valor del j-ésimo predictor en el i-ésimo individuo, con i=1,...,n y j=1,...,p.
  - β_j corresponde al j-ésimo coeficiente de regresión.

Además, los elementos η_i se pueden expresar de la siguiente manera (η_1,...,η_n) como un vector.

- **Función de enlace (Link).**

Los componentes anteriores se combinan en el MLG a través de la elección de un enlace expresado por $g(\cdot)$, de manera de relacionar μ_i con η_i (predictor lineal) mediante la función

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

De esta manera, para $i = 1,\ldots,n$ y para el valor esperado de la variable respuesta:

$$E(y_i|x_i) = \mu_i$$

Para cada distribución existe una función de enlace personal la cual se llama *enlace canónico*, para las cuales existe un estadístico suficiente y dará a lugar cuando $\eta_i = g(\mu_i) = \theta_i$.

Los enlaces más conocidos para los modelos lineales generalizados para $g(\mu_i)$ son:

Función	Enlace
Logaritmo	$\log \mu_i = \eta_i$
Identidad	$\mu_i = \eta_i$
Raíz Cuadrada	$\sqrt{\mu_i} = \eta_i$
Logit	$\log\left(\frac{\mu_i}{n-\mu_i}\right) = \eta_i$
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 de alguno de estos enlaces dependerá de: el tipo de respuesta, la distribución y la aplicación a la cual se recurrirá.

Los típicos modelos de regresión lineal para el caso de respuestas continuas son una excepción de los MLG. Este tipo de modelos, generalizan la regresión ordinaria de dos maneras: la primera es que permite que Y tenga distribuciones distintas a la distribución normal,

o sea, admite distribuciones las cuales pertenezcan a la familia exponencial, y la otra manera es que incluye distintas funciones de enlace de la media, lo que resulta bastante útil al momento de tener datos categóricos.

Los modelos lineales generalizados, admiten la combinación de una gran variedad de métodos estadísticos tales como: los modelos ANOVA, la regresión y los modelos categóricos.

1.2 Estimación por el método de Máxima Verosimilitud (Gonzalez, 2001).

Si la función de distribución de Y corresponde a una familia de distribuciones H conocida, el método de máxima verosimilitud es un método alterno para poder estimar el vector de parámetros desconocidos β.

Si se tiene un vector de observaciones dado por $y^{'} = (y_1, \ldots, y_n)$, la función de verosimilitud considera la posibilidad de que un vector $\beta \in R^p$ genere el vector de respuestas observado.

La función de verosimilitud expresada por la función de densidad conjunta de las variables aleatorias independientes $Y_1, \ldots, Y_n$ se reduce a la forma.

$$L(\beta) = h(y_1, \ldots, y_n) = h(y_1, \beta)h(y_2, \beta)\ldots h(y_k, \beta) = \prod_{i=1}^{n} h(y_i; \beta) \tag{1.6}$$

El estimador de máxima verosimilitud de β, corresponde al vector b el cual maximiza $L(\beta)$ en el espacio paramétrico $\Omega = \{(\beta, \sigma^2) : \beta \in R^p, \sigma^2 > 0\}$; esto es
$L(b) \geq L(\beta) \quad \forall\, \beta \in \Omega$

Para poder obtener el EMV se necesita maximizar $L(\beta)$ para $\beta \in R^p$.

Ya que la función logaritmo es monótona, si se aplica logaritmo a la expresión (1.6) se obtendrá lo siguiente

$$l(\beta) = logL(\beta) = \sum_{i=1}^{n} logh(y_i; \beta) \tag{1.7}$$

Por consiguiente, si la función $l(\beta)$ es derivable y continua, maximizar $L(\beta)$o $l(\beta)$ serán procesos equivalentes.

Como ejemplo, se obtendrá el EMV para el caso del modelo lineal general, considerando el supuesto que los errores se van a distribuir normal con vector de medias cero y matriz de covarianza V. De esta forma, la función de verosimilitud será

$$L = (2\pi)^{-n/2}|V|^{-1/2}exp\left\{-1/2(Y - X\beta)^{'}V^{-1}(Y - X\beta)\right\}$$

Y su función soporte será

$$l(\beta) = -\frac{n}{2}log\frac{2\pi}{2} - \frac{1}{2}log|V| - \frac{1}{2}(Y - X\beta)^{'}V^{-1}(Y - X\beta)$$

Como $l(\beta)$ corresponde a una función derivable y continua, si se deriva y luego se iguala a 0 se obtendrá la siguiente expresión

$$X^{'}V^{-1}Xb = X^{'}V^{-1}Y$$

La solución de este sistema de ecuaciones va a dar como resultado al estimador máximo verosímil de β.

$$b = (X^{'}V^{-1}X)^{-1}X^{'}V^{-1}Y \tag{1.8}$$

1.3 Igualdad en los estimadores Mínimos Cuadrados Ponderados con los estimadores de Máxima Verosimilitud en la familia exponencial (Gonzalez, 2001).

En la parte anterior se comprobó que, cuando la distribución que siguen los errores es normal, tanto los estimadores Máximo Verosímiles como los estimadores Mínimos Cuadrados Ponderados son semejantes. Es lo que reflejan los estudios realizados por Nelder y Wedderburn, quienes pudieron demostrar que la igualdad entre los EMV y los EMCP pueden ser desarrollados para algunos modelos lineales generalizados siempre y cuando la función de distribución de la variable respuesta pertenezca a la familia exponencial.

Luego, Bradley (1973) (citado en Gonzalez, 2001) aclaró que el EMCP del modelo de regresión múltiple es semejante al EMV siempre y cuando la variable dependiente siga una distribución que pertenece a la familia exponencial. Admitiendo que la variable respuesta Y, se describe por una función de densidad con la condición que pertenezca a la familia exponencial de la siguiente manera

$$h(Y, \mu) = exp\left\{p(\mu)y - q(\mu) + g(y)\right\}$$

Donde $p(\mu)$ y $q(\mu)$ son funciones, que por lo menos, son dos veces diferenciables. La esperanza y varianza del vector Y está dado de la siguiente forma

$$E(Y) = \frac{q^{'}(\mu)}{p^{'}(\mu)} = \mu \tag{1.9}$$

$$V(Y) = \left[p^{'}(\mu)\right]^{-1} \tag{1.10}$$

Considere lo siguiente

Sea $\{Y_1, Y_2, \ldots, Y_n\}$ una muestra aleatoria de tamaño n, la cual proviene desde una distribución de probabilidad perteneciente a la familia exponencial con $E(Y_i/X_i) = X_i\beta$. Por lo que, el EMV de β es exactamente igual al EMCP, es decir, satisface la expresión (1.10)

A continuación se demostrará lo anterior

El logaritmo de la función de verosimilitud $l(\beta)$ es:

$$l(\beta) = logL(\beta) = \sum \left\{p\left(x_i\beta\right) y_i - q(x_i\beta) + g(y)\right\}$$

Resolviendo el sistema de ecuaciones, y haciendo $\beta_j = b_j$

$$\frac{\partial logL(\beta)}{\partial \beta_j} = 0$$

Se tiene lo siguiente

$$\sum_{i=1}^{n} p^{'}(x_i\beta)\left(\frac{\partial x_i\beta}{\partial \beta_j}\right) y_i - \sum_{i=1}^{n} q^{'}(x_i\beta)\left(\frac{\partial x_i\beta}{\partial \beta_j}\right) = 0 \qquad j = 0, 1, \ldots, k \tag{1.11}$$

Como $\frac{\partial x_i\beta}{\partial \beta_j} = x_{ij}$ se obtiene

$$\sum_{i=1}^{n}\left[p^{'}(x_i\beta)y_i - q^{'}(x_i\beta)\right] x_{ij} = 0 \qquad j = 0, 1, \ldots, k$$

$$\sum_{i=1}^{n} p^{'}(x_i\beta)\left[y_i - \frac{q^{'}(x_i\beta)}{p^{'}(x_i\beta)}\right] x_{ij} = 0 \qquad j = 0, 1, \ldots, k$$

De las expresiones (1.18) y (1.19) se tiene lo siguiente

$$\frac{\partial logL(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} V(y_i)^{-1}\left[y_i - x_i\beta\right] x_{ij} \qquad j = 0, 1, \ldots, k \tag{1.12}$$

Finalmente, (1.21) es igual a (1.10). Por lo que, se comprueba la demostración. Charles, Frome y Yu (1976) (citado en Gonzalez, 2001) ampliaron lo presentado por Bradley para el caso de funciones enlace que no son necesariamente lineales, es decir cuando se tiene

$$E(Y/X) = f(x_i, \beta)$$

Donde $f(x_i, \beta)$ corresponde a una función no necesariamente lineal.

1.4 Selección del Modelo (Contreras, 2012) y (Posada y Rosero, 2011).

Es de conocimiento, dentro de los modelos, que existen diversos criterios para compararlos y así seleccionar el mejor bajo algún criterio. En este caso, la máxima verosimilitud es lo que permitirá obtener el mejor modelo el cual deberá ajustar de mejor manera a los datos, no obstante no sanciona su complejidad, lo que si sucede cuando se utilizan medidas de contraste como lo son el criterio AIC y BIC. Estos criterios utilizan verosimilitud *(log Lik)* que corresponde al logaritmo de la máxima verosimilitud y donde obtienen un término que resulta ser proporcional al número de parámetros (K) en el modelo, de tal manera que: log Lik - αK, donde α es igual a 2 para el caso de AIC y, para el caso de BIC igual a $\log(N)$.

1.4.1 Criterio de Información AIC

Este fue planteado por Akaike (1974) y hace referencia a la mezcla entre la teoría de máxima verosimilitud, con la información teórica y la entropía de información, este criterio considera los cambios en la bondad de ajuste y los distintos números de parámetros tomando en cuenta dos modelos. Está definido de la siguiente manera:

$$AIC = -2 \cdot logLik + 2K$$

1.4.2 Criterio de Información BIC

Este criterio se calcula sobre los distintos modelos como una función de la bondad de ajuste del $logLik$ con el número de parámetros ajustados (K) y finalmente con N que corresponde al número total de datos. Está definido de la siguiente manera:

$$BIC = -2 \cdot logLik + log(N) \cdot K$$

En ambos criterios, el modelo que dé como resultado el valor más bajo de AIC y BIC será considerado el más adecuado para explicar los datos utilizando la menor cantidad posible de parámetros.

1.5 Modelos Lineales Generalizados para recuentos (Figueroa, 2005)

Son muchos los casos en donde las variables respuestas son recuentos. Las variables de conteo o también conocidas como variables de recuento son definidas como la cantidad de sucesos/eventos que, en una misma unidad de observación, ocurren en un determinado intervalo de tiempo específico (Lindsey, 1995b) citado en (Figueroa, 2005).

Desde esta definición expuesta por Lindsey, surgen dos características fundamentales proveniente de una variable de recuento, lo que la hace diferenciarse de una variable cuantitativa de tipo continua, estas diferencias son su naturaleza discreta y no negatividad. En reiteradas situaciones, los recuentos aparecen al momento de resumir tablas de contingencia.

Es el modelo más simple en donde se asume que el componente aleatorio Y sigue una distribución de Poisson. Se sabe que esta distribución es unimodal, es decir, que tiene una sola media y su propiedad más importante es que la media y la varianza son iguales.

$$E(Y) = V(Y) = \mu$$

De manera que, cuando se tenga un número de recuentos mayor en media, también tenderá a tener mayor variabilidad.

En el caso de los modelos lineales generalizados, usualmente se utiliza el logaritmo de la media para la función de enlace, de modo que el modelo con una variable explicativa X pueda expresarse de la siguiente forma

$$log(\mu) = \mu = \beta x$$

De modo que

$$\mu = exp\left[\mu + \beta x\right] = e^{\alpha}(e^{\beta})^{x}$$

En el siguiente capítulo se podrán conocer los modelos Poisson y Binomial Negativa en cuanto a componentes, propiedades, estimaciones entre otros.

Capítulo 2

Modelos de Regresión Poisson y Binomial Negativa

2.1 Aplicaciones de variable de Poisson (Figueroa, 2005)

Ya es de conocimiento que, un conteo corresponde al número en que un determinado evento/suceso acontece en una unidad de observación durante un específico periodo de tiempo o espacio. Algunos ejemplos pueden ser los siguientes:

Conteos en el espacio

- Número de accidentes de tráfico que se originan en el cruce de 2 carreteras.
- Número de árboles infectados por hectárea en un bosque.
- Número de organismos infecciosos propagados en una placa.
- Entre otros.

Conteos en el tiempo

- Número de mutaciones en una población de animales durante 5 años.
- Número del registro de partículas de una desintegración radioactiva por segundo.
- Número de accidentes de tráfico en un tramo de cierta carretera en un mes.
- Entre otros.

2.2 La Distribución de probabilidad Poisson

La función de probabilidad de la distribución de Poisson es:

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k \in 0,1,2,3,\ldots \qquad \lambda \in (0,\infty) \tag{2.1}$$

Donde k es el número de ocurrencias del evento o fenómeno, λ corresponde al parámetro positivo que representa al número de veces que se espera que ocurra el fenómeno durante un intervalo específico de tiempo o espacio y e corresponde a la constante exponencial.

Una distribución Binomial con una probabilidad de éxito p muy pequeña y n grande se aproxima a una distribución de Poisson con $\lambda = n \cdot p$.

Algunas referencias utilizan esta aproximación cuando $n > 30$ y $p > 0{,}1$ y/o $np < 5$.

2.2.1 Propiedades de la Distribución de Poisson (Contreras, 2012).

1. Si μ aumenta, la forma de la distribución se traslada hacia la derecha. Por lo que $E(y_i) = \mu$. El parámetro μ se conoce como *tasa* dado que es el número esperado de veces que ocurre un determinado evento por unidad de tiempo.

2. La varianza y la esperanza, en la distribución Poisson, son iguales. A esta propiedad se le llama equidispersión, es decir

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

3. Mientras que μ aumenta, $P(Y_i = 0)$ disminuye.

4. Mientras que μ aumenta, la distribución de Poisson se acerca a la distribución Normal.

2.3 La Distribución Poisson como Familia Exponencial (Contreras, 2012).

Sea $Y_1, \ldots, Y_n$ variables aleatorias independientes e idénticamente distribuidos. La función de probabilidad de este vector se puede escribir de la siguiente manera, ya que pertenece a la familia exponencial y haciendo la equivalencia con la familia exponencial, se tiene:

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

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 demuestra que la distribución Poisson, pertenece a la familia exponencial.

Además, en el modelo Poisson la media y la varianza son iguales entre sí y a su vez iguales a μ_i.

2.4 Función de enlace (Link)

Debido a que el parámetro canónico de la distribución de Poisson está dado por $\theta = \log\mu$, la función de enlace canónico para la distribución de Poisson es $\eta = \theta = \log\mu$, donde μ simboliza el valor medio de la distribución Poisson. Mediante este enlace, las covariables en vez de tener un efecto aditivo sobre la media tienen un efecto multiplicativo sobre ésta.

Existen funciones enlace alternativas, si es que llegase a fallar el enlace canónico, éstas se presentan a continuación:

- Enlace Identidad: $g(\mu) = \mu$.
- Enlace Raíz cuadrada: $g(\mu) = \sqrt{\mu}$

No obstante, estas funciones podrían resultar problemáticas para el caso de las predicciones de μ_i, ya que la siguiente expresión podría ser negativa.

$$g(\hat{\mu}_i) = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

2.5 Modelo de Regresión Poisson (MRP)

El Modelo de Regresión Poisson (MRP) es un modelo el cual se distingue por ser utilizado en estudios de variable de recuento. Es un modelo ideal para poder modelar valores enteros no negativos, principalmente cuando la frecuencia que ocurra un determinado suceso es baja.

Se dice que una variable Y_i sigue el MRP si se cumple lo siguiente:

$$Y \sim P(\mu_i), \qquad i = 1, 2, 3, \ldots, n$$
$$g(\mu_i) = \eta_i = x_i^t \beta$$

Donde

- $x_i = (x_{i1}, \ldots, x_{ip})^t$ corresponde al vector de covariables explicativas.
- $\beta = (\beta_0, \ldots, \beta_p)^t$ corresponde al vector de parámetros desconocidos.

Los elementos que conforman este modelo son:

- **Componente Aleatoria**

Dado $Y_1, \ldots, Y_n$ un vector de variable respuesta positiva y sea $x_i = (x_{i1}, \ldots, x_{ip})^T$ un vector de covariables explicativas con parámetro μ_i define lo siguiente

$$Y \sim P(\mu_i), \qquad i = 1, 2, 3, \ldots, n$$

- **Componente Sistemática**

Dado μ_i, y el predictor lineal representado por

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

- **Función de enlace (Link)**

Mediante la siguiente elección de una función enlace, ambos componentes presentados anteriormente se combinan en el modelo.

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

Las funciones de enlace más utilizadas para este tipo de modelo de regresión son

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

Esto es, para cuando el enlace logaritmo $\hat{\mu}_i = exp(x_i^T \hat{\beta})$ es positivo.

2.6 Similitudes y diferencias entre el MRP con otros Modelos de Regresión

Se puede apreciar que la variable dependiente o variable respuesta, en todo lo expuesto anteriormente, corresponde a un número entero no negativo. Se puede expresar esta variable en término de un conjunto de covariables.

Existen estudios que utilizan esta distribución, como por ejemplo los estudios de demanda de salud los cuales modelan datos del número de individuos que utilizan un servicio de salud, como es visita al doctor en un determinado año. Son en estas circunstancias llevadas a casos prácticos en donde la variable respuesta de un estudio experimental es un conteo.

(a) Variable respuesta

Esta variable asume una distribución de probabilidad Poisson, donde la variable aleatoria se detalla como la cantidad de eventos/sucesos que acontecen en un determinado intervalo de tiempo o espacio donde su ocurrencia es aleatoria, independiente en el transcurso del tiempo y donde posee una tasa de ocurrencia constante. Esta distribución se utiliza para poder modelar eventos por unidad tanto espacial como de tiempo. Por el contrario a lo que sucede en un modelo de regresión tradicional, la variable respuesta en el MRP es discreta, con valores enteros positivos y con una distribución de probabilidad Poisson.

(b) Distribución de probabilidad

Para el caso de los datos continuos se tiene la distribución Normal, así como se tiene la distribución de Poisson para datos de conteo. En esta distribución se tiene un solo parámetro que es la media, la cual tiene la condición de ser siempre positivo. Así, esta distribución en su totalidad es determinada por ese único parámetro.

El MRP tiene un rol fundamental para analizar datos de conteo. Sus principales características son:

- Facilita una descripción satisfactoria de datos en donde la varianza es conforme a su media.
- Sin muchas restricciones.
- Los eventos o conteos ocurren independientemente y aleatoriamente en el tiempo, con una tasa de ocurrencia constante, el modelo determina el número de eventos dentro de un intervalo específico.

2.7 Estimación Máxima Verosimilitud del Modelo de Regresión Poisson

Sea $Y_1, \ldots, Y_n$ un conjunto con n observaciones aleatorias e independientes donde el predictor es x, entonces se define la función de verosimilitud de la siguiente manera:

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

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

Ahora se podrá maximizar, luego de haber sido elegida la función de enlace. El logaritmo de la función de verosimilitud está dado por:

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

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

2.8 Sobredispersión (Figueroa, 2005).

El modelo Poisson es presentado como un modelo capaz de mejorar diversas cosas para poder representar datos de conteos, no obstante este modelo podría resultar inadecuado debido a la violación de algunos supuestos, cuyo origen es disparejo Winkelmann (2000) citado

en (Figueroa, 2005). Frecuentemente, puede presentarse subdispersión o sobredispersión, pero la que surge con mayor costumbre es ésta última. Es más, las pruebas que se utilizan para poder apreciar equidispersión generalmente son denominadas pruebas de sobredispersión.

La equidispersión compone un supuesto básico, es decir se toma que $Var(Y) = \sigma^2 E(Y)$, donde el parámetro de dispersión $\sigma^2 = 1$. La sobredispersión acontece cuando $Var(Y) > E(Y)$, es decir $\sigma^2 > 1$. Si se está en el caso de un exceso de variación en los datos, podrían resultar sesgadas las estimaciones de los errores estándar, lo que podría causar errores como por ejemplo en las inferencias comenzando en los parámetros del modelo de regresión.

Existen diferentes causas para la sobredispersión, de las cuales se pueden mencionar las siguientes, (Contreras, 2012).

- Error al momento de optar por la función de enlace, por ejemplo no fue conveniente elegir el enlace log-lineal.
- Falta de estabilidad, es decir, la probabilidad de ocurrencia de un evento puede ser independiente de la ocurrencia de un evento previo pero no es constante.
- Alta inestabilidad en los datos.
- Los eventos no ocurren independientemente a través del tiempo.
- Los datos no provienen de una distribución Poisson.
- Errores de especificación de la media m (Winkelmann, 2000) citado en (Figueroa, 2005), como por ejemplo omitir variables explicativas o que entran al modelo a través de alguna transformación en lugar de linealmente.

Hay diferentes fórmulas para descubrir una sobredispersión, por ejemplo Lindsey (1995b) citado en (Figueroa, 2005), plantea que se aplique un coeficiente de variación (CV).

$$CV = \frac{Var(\mu_i)}{\mu_i}$$

Este CV, hipotéticamente, debiese ser igual a 1, en el caso que se cumpliera la equidispersión.

Por lo general, se calcula la sobredispersión apreciando la relación entre la estadística de Pearson χ^2 o la función de desvío D y sus pertenecientes grados de libertad (gl), es decir calcular

$$\frac{\chi^2}{gl} \qquad y \qquad \frac{D}{gl}$$

Si al calcular estos valores se obtiene un resultado mayor a 1, se estará frente a una sobredispersión.

Por otro lado, está el análisis fundamentado en una prueba de Razón de Verosimilitud (RV) apoyada en las distribuciones Poisson y Binomial Negativa.

- En el caso de la distribución Poisson $Var(Y) = \mu$.
- En el caso de la distribución Binomial Negativa $Var(Y) = \mu + k\mu^2$.

Si $k = 0$, desde una distribución Binomial Negativa se reducirá a una distribución Poisson.

Por lo tanto, las hipótesis que se plantean son:

$$H_0 : k = 0 \qquad v/s \qquad H_1 : k > 0$$

Para realizar esta prueba, se ajustarán los 2 modelos: Poisson y Binomial Negativa (BN). Para ambos modelos se deberá obtener su respectiva función de log verosimilitud, denotada por l.

La estadística de prueba es la siguiente

$$RV = -2(l(Poisson) - l(BN))$$

Cameron y Trivedi (1998) citado en (Figueroa, 2005), dicen que esta prueba tiene una distribución asintótica $\chi^2_{(1-2\alpha,1)}$. Por consiguiente, se rechazará H_0 si es que la estadística logra ser mayor que $\chi^2_{(1-2\alpha,1)}$. De darse esa situación, será mucho más apropiado ajustar el número de ocurrencias mediante una distribución Binomial Negativa, pero la interpretación será igual que en el caso de la Regresión Poisson.

2.9 Distribución Binomial Negativa

Esta distribución se puede considerar como una extensión de la distribución Geométrica. Es un modelo conveniente para tratar algunos procesos en los cuales se repite un explícito ensayo o prueba hasta obtener un número determinado de resultados favorables (por primera vez). Es por esta razón que es de gran utilidad para muestreos que provengan de esta manera. En el caso que el número de resultados favorables fuese 1, se estará en presencia de una distribución Geométrica.

La densidad de la distribución Binomial Negativa está dada por (Contreras, 2012)

$$f(y_i) = \frac{\Gamma(\phi + y_i)}{\Gamma(y_i + 1)\Gamma(\phi)}(1 - \mu_i)^{y_i}{\mu_i}^{\phi} \qquad con \quad y_i = 0, 1, 2, 3, \ldots$$

Propiedades de la distribución Binomial Negativa.

Las propiedades que tiene la distribución Binomial Negativa son las siguientes.

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

La función de probabilidad puede tomar numerosas formas y valores de los parámetros que identifica a la distribución Binomial Negativa.

Esta distribución o modelo puede derivarse a partir de un proceso Bernoulli en el cual se presentan las siguientes condiciones:

- El proceso constituye un número no definido de pruebas separadas o separables. El proceso de una distribución Binomial Negativa finalizará cuando se obtenga un determinado números de resultados favorables llamado k.
- Cada prueba puede dar origen a resultados mutuamente excluyentes A y no A.
- La probabilidad de obtener un resultado A en cada una de las pruebas es p, donde q es la probabilidad de no A, lo que da a lugar a $p + q = 1$.
- Las probabilidades p y q son constantes en todas las pruebas. Todas las pruebas son independientes (si se trata de un experimento de extracción, éste se realizará con devolución del individuo extraído, a no ser que se trate de una población en la que el número de individuos tenga de carácter infinito).
- Derivación de la distribución. En estas condiciones, la variable aleatoria x corresponde al número de experimentos que se necesitan para poder alcanzar K éxitos (resultados A) ; entonces la variable aleatoria X seguirá una distribución binomial negativa con parámetros p y k.

Entonces

$$X \Rightarrow BN(p, k)$$

2.10 La Distribución Binomial Negativa como Familia Exponencial (Contreras, 2012)

Otra forma de escribir la función de probabilidad de la distribución binomial Negativa es la siguiente

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

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

Si $1/\phi\to 0$, entonces $Var(Y_i)\to\mu_i$ lo que indica que la distribución Binomial Negativa converge a una distribución de Poisson.

Cuando ϕ es fijo, esta densidad pertenece a la familia exponencial y se podría hablar de un modelo lineal generalizado Binomial Negativa.

Entonces, si se indica $Y|z\sim P(z)$ y $Z\sim G(\mu,\phi)$ donde ϕ no depende de μ (Paula, 2010) citado en (Contreras, 2012).

En este caso

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

Se tiene lo siguiente

$$f(y|z)=\frac{e^{-z}z^y}{y!}$$
$$g(z;\mu,\phi)=\frac{1}{\Gamma(\phi)}\left(\frac{z\phi}{\mu}\right)^{\phi}e\frac{-\phi z}{\mu}\frac{1}{z}$$

La función de probabilidad Y viene dada por

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

Transformando la variable

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

Tenemos

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

Luego se deduce que

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

En el que $\pi = \frac{\mu}{\mu+\phi}$.

Por lo tanto Y sigue una distribución Binomial Negativa con

- Media μ.
- Parámetro de dispersión ϕ.

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

$$logf(y) = exp\left\{ylog\left(\frac{\mu}{\mu+\phi}\right) - \phi log\left(\frac{\mu+\phi}{\phi}\right) + log\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\right\} \tag{2.2}$$

Además, se deduce de lo anterior lo siguiente:

$$E(Y_i) = \mu_i$$

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

2.11 Modelo de Regresión Binomial Negativa (MRBN)

La más frecuente de las razones por las que el modelo Poisson falla es la heterogeneidad no observada. Es decir, existen factores no observados, que son característicos de los individuos, que realizan ciertas influencias sobre la variabilidad que se relaciona con la variable de respuesta.

La dificultad que hay, es que la heterogeneidad no observada podría tener ciertas consecuencias para los procesos de inferencia estadística. En primer lugar, puede haber sobredispersión y, en segundo lugar, un número enorme de ceros. La heterogeneidad, que no toma en cuenta el modelo Poisson, puede modelarse de manera evidente mediante el uso de la regresión binomial negativa.

Existen, al menos, dos formas las cuales permiten a la distribución binomial negativa poder descender: la más recurrente asume que se está en presencia de una mezcla de distribuciones, en donde las observaciones se distribuyen como una distribución Poisson, pero se presupone un elemento de heterogeneidad individual no observado (la cual sigue una distribución gamma en su fórmula tradicional) la cual manifiesta el hecho de que la verdadera media no ha sido medida correctamente. La segunda forma, asume que existe una manera específica de dependencia entre sucesos, de forma que la ocurrencia de un evento aumenta la probabilidad de ocurrencia de eventos posteriores, aunque a pesar del último comentario esto podría solo ocurrir en estudios longitudinales.

2.12 Formulación del Modelo de Regresión Binomial Negativa (Contreras, 2012).

Se dice que una variable Y_i sigue el MRBN, si se cumple lo siguiente

$$Y_i \sim BN(\mu_i, \phi), \qquad i = 0, 1, 2, 3, \ldots$$
$$g(\mu_i) = x_i^t \beta$$

Donde

- $x_i = (x_{i1}, \ldots, x_{ip})^t$ corresponde al vector de covariables explicativas.
- $\beta = (\beta_0, \ldots, \beta_p)^t$ corresponde al vector de parámetros desconocidos.

Los elementos que conforman este modelo son:

- **Componente Aleatoria**

Sea $Y_1, \ldots, Y_n$ una variable aleatoria independiente la cual hace referencia al número de sucesos necesarios para poder obtener r-éxitos. Es decir, el número de éxito está establecido y la aleatoriedad es el número de suceso, de modo que

$$Y_i \sim BN(\mu_i, \phi), \qquad i = 0, 1, 2, 3, \ldots$$

- **Componente Sistemática**

Dado μ_i, y el predictor lineal representado por

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

- **Función de enlace (Link)**

Ambos componentes presentados anteriormente se combinan en el modelo, a través de la elección de la función de enlace

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

Donde la $g(\cdot)$ es una función de enlace.

Las funciones de enlace más utilizadas para este tipo de Modelo de Regresión son

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

2.13 Estimación Máxima Verosimilitud del Modelo de Regresión Binomial Negativa (Contreras, 2012).

Se considera la siguiente partición $\theta = (\beta^T, \phi)^T$, la cual denota el logaritmo de la función de verosimilitud por Paula (2010).

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

Donde $\mu_i = g^{-1}(x_i^T \beta)$, es una función score para β.

Se calcula primeramente las derivadas para la función score de β.

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

Donde

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

Luego, se puede expresar la función score en forma matricial para β.

$$U_\beta(\theta) = X^T W F^{-1}(y - \mu) \tag{2.3}$$

Donde

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

Lo mismo se puede expresar para la función score de ϕ, dada de la siguiente manera

$$U_\phi(\theta) = \sum_{i=1}^{n} \left[\psi(\phi + y_i) - \psi(\phi) - \frac{(y_i + \phi)}{(\phi + \mu_i)} + log\left\{ \frac{\phi}{(\phi + \mu_i)} \right\} + 1 \right] \tag{2.4}$$

Donde $\psi(\cdot)$ es una función digama (derivada logarítmica de la función gama).

Para poder obtener la matriz de información de Fisher se deben calcular las derivadas.

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

Donde los valores esperados están dados por

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

Luego, se puede expresar la información de Fisher para β, en forma matricial de la siguiente manera

$$K_{\beta\beta}(\theta) = E\left\{\frac{\partial^2 L(\theta)}{\partial\beta\partial\beta^T}\right\} = X^T W X$$

Lawlees (1982) señala que la información de Fisher para ϕ se puede expresar como sigue a continuación

$$K_{\beta\beta}(\theta) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{\infty}(\phi+j)^2 Pr(Y \geq j) - \phi^{-1}\mu_i/(\mu_i+\phi)\right\}$$

Donde β y ϕ corresponden a dos parámetros ortogonales. Por lo tanto, la matriz de información de Fisher para θ asume una forma de bloque diagonal.

$$K_{\theta\theta} = \begin{bmatrix} K_{\beta\beta} & 0 \\ 0 & K_{\phi\phi} \end{bmatrix}$$

La estimación de máxima verosimilitud para θ y ϕ puede ser conseguida mediante un algoritmo de mínimos cuadrados ponderados para obtener $\hat{\theta}$ desarrollado a partir del punto (2.3) y el Método de Newton-Raphson para obtener $\hat{\phi}$ desarrollado a partir del punto (2.4) el cual se muestra a continuación

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

$$\phi^{m+1} = \phi^m - \left\{\frac{U_\phi{}^m}{\ddot{L}^m_{\phi\phi}}\right\}$$

Para $m = 0, 1, 2, 3, \ldots$, en la que

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

2.14 Método Newton Raphson (Verdin, 2005)

Este método resuelve ecuaciones que siguen la siguiente forma $f(x) = 0$ de esta manera:

Sea $\bar{x} = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n]^T$ lo cual corresponde a una raíz del sistema no lineal de $n \times n$, de la ecuación $f(x) = 0$, cuya i-ésima ecuación está dada por:

$$f_i(x) = f_i(x_1, \ldots, x_n) = 0, \qquad i = 1, \ldots, n \tag{2.5}$$

Se supone que x_k corresponde a una aproximación presente de $\bar{x}$. La forma en que se puede obtener una aproximación corregida x_{k+1} será resolviendo un sistema lineal el cual acerque al sistema (2.5) para que x esté cerca de x_k. Puntualmente, si $x = x_k + dx$, donde $dx = [dx_1 \quad dx_2 \dots dx_n]^T$ se podrá conseguir una aproximación a la ecuación $f_i(x_k + dx) = 0$ en (2.5) utilizando la diferencia completa.

De esta manera, se obtiene

$$f_i(x_k) + \frac{\partial f_i(x_k)}{\partial x_1}dx_1 + \frac{\partial f_i(x_k)}{\partial x_2}dx_2 + \ldots + \frac{\partial f_i(x_k)}{\partial x_n}dx_n = 0, \qquad i = 1, \ldots, n \tag{2.6}$$

El sistema anterior es lineal en $dx_1, dx_2, \ldots, dx_n$ y su forma matricial está dada de la siguiente manera:

$$\begin{bmatrix} \frac{\partial f_1(x_k)}{\partial x_1} & \frac{\partial f_1(x_k)}{\partial x_1} & \ldots & \frac{\partial f_1(x_k)}{\partial x_1} \\ \frac{\partial f_2(x_k)}{\partial x_1} & \frac{\partial f_2(x_k)}{\partial x_1} & \ldots & \frac{\partial f_2(x_k)}{\partial x_1} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n(x_k)}{\partial x_1} & \frac{\partial f_n(x_k)}{\partial x_1} & \ldots & \frac{\partial f_n(x_k)}{\partial x_1} \end{bmatrix}^{-1} \begin{bmatrix} dx_1 \\ dx_2 \\ \vdots \\ dx_n \end{bmatrix} = - \begin{bmatrix} f_1(x_k) \\ f_2(x_k) \\ \vdots \\ f_n(x_k) \end{bmatrix} = J^{-1}dx = -f(x_k) \tag{2.7}$$

Donde $J = f^{'}(x_k)$

Por consiguiente, se puede obtener x_{k+1} a partir de x_k como $x_{k+1} = x_k + dx_k$ donde dx_k corresponde a la solución de $f^{'}(x_k)^{-1}dx = -f(x_k)$.

La matriz $J = f^{'}(x_k)$ en (2.7) corresponde a la *matriz jacobiana no lineal* que se relaciona con la ecuación $f(x) = 0$ en x_k. Se observa que en la fila i de J posee todas las derivadas parciales de $f_i(x)$ (i-ésima ecuación) por otro lado, la columna j de J posee todas las derivadas parciales con respecto a x_j (j-ésima variable).

De esta forma

$$f^{'}(x_k) = \left[\frac{\partial f_i(x_k)}{\partial x_j}\right]_{n \times n}$$

Por el contrario, si se desea estimar un vector de parámetros θ a través de máxima verosimilitud, se deberán solucionar las ecuaciones de verosimilitud, en las cuales

$$\theta_{(n+1)} = \theta_n - F^{'}(\theta_n)^{-1} f(\theta_n)$$

Obteniendo el logaritmo de la función de verosimilitud, entonces

$$f(\theta_n) = \frac{dln(L)}{d\theta} = S(\theta) \qquad y$$

$$F^{'}(\theta_n) = -I(\theta) = \begin{bmatrix} \frac{\partial^2 ln(L)}{\partial\theta_1\partial\theta_1} & \frac{\partial^2 ln(L)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 ln(L)}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2 ln(L)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 ln(L)}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2 ln(L)}{\partial\theta_2\partial\theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 ln(L)}{\partial\theta_k\partial\theta_1} & \frac{\partial^2 ln(L)}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2 ln(L)}{\partial\theta_k\partial\theta_k} \end{bmatrix}$$

A la matriz anterior se le llamará *matriz de información de Fisher*. Por consiguiente, la estimación mediante el método de Newton Raphson será:

$$\theta_{(n+1)} = \theta_n - F^{'}(\theta_n)^{-1} f(\theta_n)$$

El método de Newton Raphson posee una ventaja en la resolución de sistemas de ecuaciones no lineales, la cual corresponde a su rapidez de convergencia, al momento que se tiene una aproximación lo bastantemente precisa. Por el contrario, una de sus desventajas es que necesita una aproximación inicial exacta de la solución para así lograr certificar la convergencia.

Se aclara que no siempre será sencillo poder establecer valores iniciales que logren dar lugar a una solución. Por lo general, los estimadores de momentos son considerados como valores iniciales para poder aplicar el método, no es sencillo obtener los estimadores mediante el Método de Momentos ya que existen cálculos de por medio que dificultan encontrarlos.

2.15 Comparación entre modelos (Huachel, Boggio y Harvey, 2010)

Es de suma importancia observar que la regresión Binomial Negativa es una extensión de la regresión Poisson bajo el supuesto de varianza más liberal la cual tendería a una Regresión Poisson cuando el parámetro α es igual a 0. WenSui Liu y Jimmy Cela (2008) y Hilbe (2007) citado en (Huachel, Boggio y Harvey, 2010), se fundamentan en este hecho para así poder plantear y comparar ambos modelos a través de sus log-verosimilitudes mediante un test de razón de verosimilitud.

No obstante, cuando se habla de MLG, se considera un impedimento el hecho que se diferencien en el componente aleatorio, para la comparación de dos modelos. Es por eso que se plantea considerar el criterio antes mencionado, es decir el criterio de información de AKAIKE (AIC), u otra opción es el criterio de información Bayesiano (BIC), para poder basarse y así seleccionar el modelo más conveniente.

Capítulo 3

Aplicación

3.1 Introducción.

En este capítulo se llevará a cabo la aplicación de los modelos Poisson y Binomial Negativo en los modelos lineales generalizados. Previo a esto, se presentarán los datos de conteo a utilizar y su correspondiente análisis exploratorio. Todo lo anterior se realizará utilizando el Software R-Project (versión 3.1.3).

Los datos fueron obtenidos desde CONASET el 2015.11.02. Excel "Accidentes de transito 2011-2014 OK", pestaña "Consolidación 2011-2014 comuna".

3.2 Variables.

A continuación se presentan todas las variables de interés de la base de datos:

- X_1: Número de accidentes en el año 2011.
- X_2: Número de lesionados en el año 2011.
- X_3: Número de accidentes en el año 2012.
- X_4: Número de lesionados en el año 2012.
- X_5: Número de accidentes en el año 2013.

- X_6: Número de lesionados en el año 2013.
- X_7: Número de accidentes en el año 2014.
- X_8: Número de lesionados en el año 2014.

Cabe destacar que para la aplicación se considerarán 2 modelos: un modelo para accidentes y otro modelo para lesionados, con el objetivo de concluir cual de los dos modelos de regresión es mejor, para estos datos, y así saber si es que existe relación entre un año y otro. Las variables son las siguientes:

- Modelo 1.

  Se considerará la variable respuesta Y como el número de accidentes del año 2014 en la región de Valparaíso y el número de accidentes del año 2011, 2012 y 2013 como sus respectivas covariables.

- Modelo 2.

  Se considerará la variable respuesta Y como el número de lesionados del año 2014 en la región de Valparaíso y el número de lesionados del año 2011, 2012 y 2013 como sus respectivas covariables.

3.3 Análisis Exploratorio.

Luego de plantear los modelos con los cuales se trabajará y previo a la implementación de los modelos de regresión a cada uno de éstos, es que se llevó a cabo el análisis exploratorio de los datos.

Tabla 3.1: Estadística descriptiva de las variables a utilizar.

Variables	**Promedio**	**Desv. Estándar**	**Varianza**	**Mínimo**	**Máximo**
Accidentes2011	203,26	332,89	110819	0	1657
Lesionados2011	176,42	240,08	57637,39	0	1106
Accidentes2012	201,74	330,56	109270,8	0	1565
Lesionados2012	173,08	236,16	55770,94	0	1057
Accidentes2013	227,92	380,33	144651,2	0	1765
Lesionados2013	182,84	240,42	57801,87	0	1026
Accidentes2014	214,21	366,49	134313,6	0	1696
Lesionados2014	146,45	200,20	40078,79	0	862

De la Tabla 3.1, se puede observar que en promedio hay 146,45 lesionados en el año 2014 de un promedio de 214,21 accidentes. Se concluye también que el año 2012 fue el año que en promedio tuvo menos accidentes, por el contrario el que tuvo más accidentes fue el año 2013.

Se observa también que en cuanto a los lesionados, el año que más los tuvo fue el año 2013 con un valor promedio igual a 182,84 personas, mientras que en el año 2014 hubo en promedio 146,45 personas lesionadas, siendo el año con menor frecuencia.

Por otra parte se puede observar que existen comunas en las cuales no hay ningún accidente, ni lesionados como lo es la Isla de Juan Fernández.

3.4 Comportamiento de los datos.

Antes de aplicar cualquier modelo lineal a unos datos es necesario verificar el comportamiento de la variable respuesta con sus respectivas covariables para que, de esta manera, se verifique linealidad.

Figura 3.1: Gráfico de dispersión entre los accidentes del año 2014 y los accidentes del año 2011 con un coeficiente de correlación igual a 0,953498

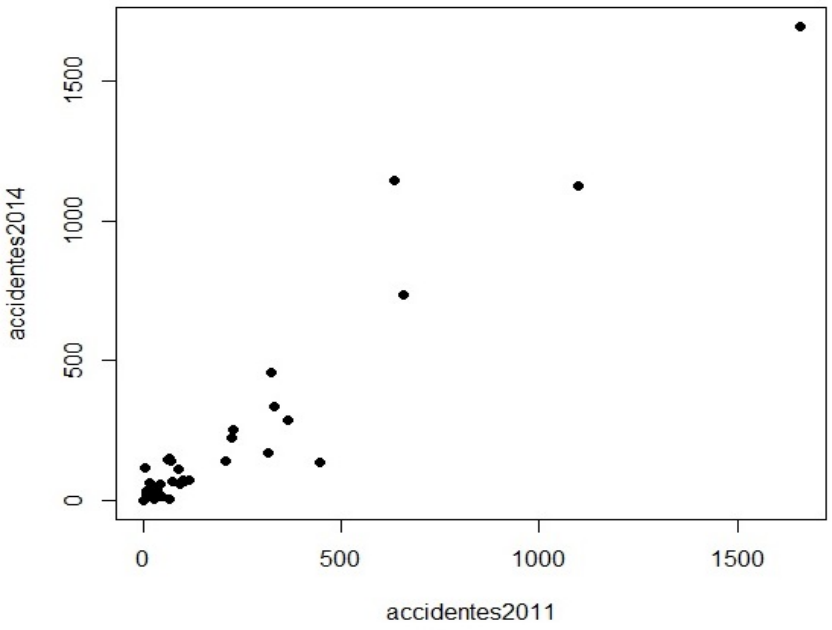


Figura 3.2: Gráfico de dispersión entre los accidentes del año 2014 y los accidentes del año 2012 con un coeficiente de correlación igual a 0,9719625

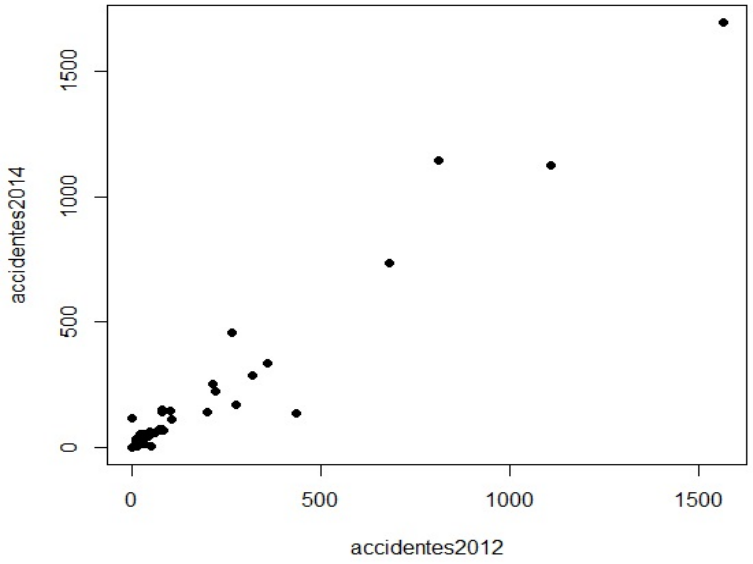


Figura 3.3: Gráfico de dispersión entre los accidentes del año 2014 y los accidentes del año 2013 con un coeficiente de correlación igual a 0,9810542

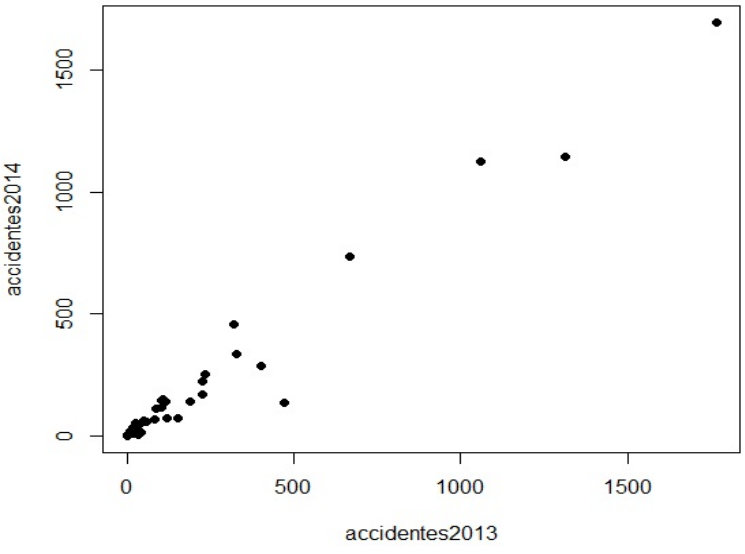


Figura 3.4: Gráfico de dispersión entre los lesionados del año 2014 y los lesionados del año 2011 con un coeficiente de correlación igual a 0,9246874

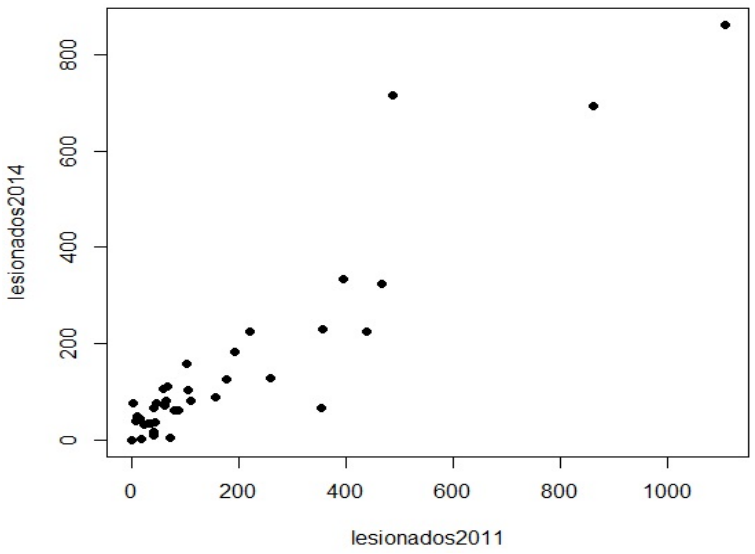


Figura 3.5: Gráfico de dispersión entre los lesionados del año 2014 y los lesionados del año 2012 con un coeficiente de correlación igual a 0,9532273

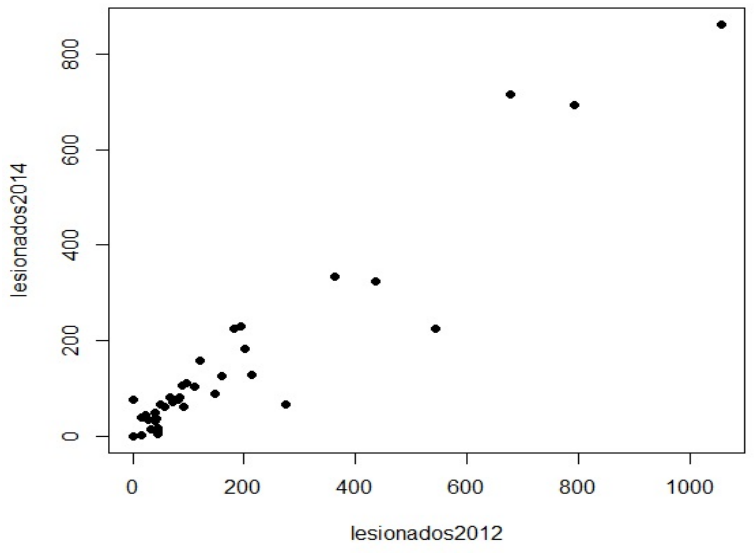
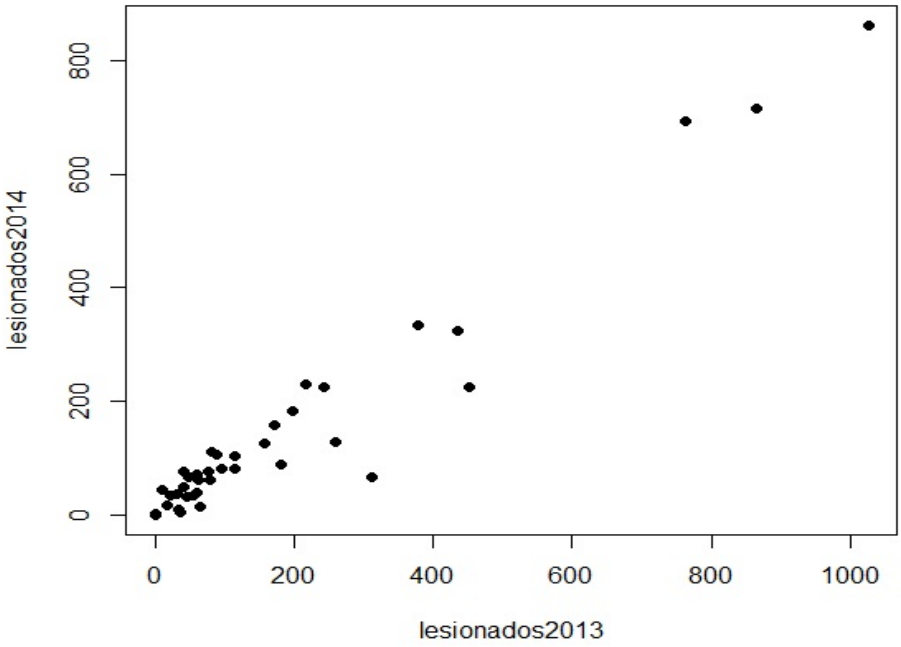


Figura 3.6: Gráfico de dispersión entre los lesionados del año 2014 y los lesionados del año 2013 con un coeficiente de correlación igual a 0,9694394



De los gráficos anteriores se puede observar el comportamiento de cada variable respuesta con sus respectivas covariables, de los cuales se nota la linealidad que existe.

Resumiendo lo anterior se tiene que:

Tabla 3.2: Coeficientes de correlacion (r).

Variables	**Correlación**
accidentes2014 - accidentes2011	0,953498
accidentes2014 - accidentes2012	0,971962
accidentes2014 - accidentes2013	0,981054
lesionados2014 - lesionados2011	0,924687
lesionados2014 - lesionados2012	0,953227
lesionados2014 - lesionados2013	0,969439

Finalmente, de la Tabla 3.2 se observa que existe relación lineal fuerte entre las variables descritas anteriormente.

3.5 Aplicación del Modelo de Regresión Poisson considerando el Modelo 1.

El supuesto fundamental para la aplicación del modelo Poisson es que exista equidispersión, el cual fue descrito anteriormente en el Capítulo 2. Para determinar si existe o no sobredispersión es que utilizará el siguiente test de la librería *AER*.

`dispersiontest`

Este modelo habla que $E[y] = \mu$, que supone que es igual a la varianza Var$[y] = \mu$. "dispersiontest" evalúa la hipótesis de que esta suposición es válida (equidispersión) contra la alternativa de que la varianza es de la forma:

$$Var[y] = \mu + \alpha * trafo(\mu)$$

Se concluirá que existirá sobredispersión si $\alpha > 0$, por el contrario si $\alpha < 0$ no habrá sobredispersión.

El coeficiente α puede estimarse a través de una regresión MCO auxiliar (mínimos cuadrados ordinarios) y probado con la estadística correspondiente, que es asintóticamente normal estándar bajo la hipótesis nula.

Especificaciones comunes de la función de transformación $trafo$ son $trafo(\mu) = \mu^2$ o $trafo(\mu) = \mu$. La primera corresponde a un modelo binomial negativo (BN) con función cuadrática varianza (denominada BN2 por (Cameron y Trivedi, 2005), y la segunda corresponde a un modelo (BN) con la función lineal de la varianza (denominado BN1 por (Cameron y Trivedi, 2005)) o cuasi-Poisson, modelo con los parámetros de dispersión, es decir:

$$V[y] = (1 + \alpha) * \mu = dispersion * \mu$$

A continuación se observa el valor que arroja el test "dispersiontest" bajo el criterio de α:

$$\alpha = 67,45195$$

Entonces, con un valor de $alpha = 67{,}45$ y considerado lo que dice el test, se concluye que existe Sobredispersión. Lo que indica que, para estos datos de conteo, debiese ser mejor el modelo Binomial Negativa por sobre el modelo Poisson.
Se observa también con este test, sin especificar $trafo$, el siguiente valor:

$$dispersion = 68,45195$$

Si el valor *dispersion* es mayor que 1, se estará en presencia de una Sobredispersión. Por el contrario, si el valor *dispersion* es menor que 1, se estará en ausencia de una Sobredispersión. Por lo tanto, se concluye que: con un valor de $dispersion = 2,324$, debiese ser mejor el modelo Binomial Negativa por sobre el modelo Poisson, confirmando lo que dice el criterio basado en α.

Por lo tanto, se concluye que: con un valor de $dispersion = 68,45$, debiese ser mejor el modelo Binomial Negativa por sobre el modelo Poisson, confirmando lo que dice el criterio basado en α.

Se supone que el número de accidentes del año 2014 en la Región de Valparaíso son independiente al de otros para un Modelo de Regresión Poisson con parámetro μ_i. Sea Y el número de accidentes del año 2014, entonces:

$$Y \sim Poisson(\mu_i)$$

En la tabla 3,3 se presentan los estimadores de los coeficientes de regresión para el modelo 1:

Tabla 3.3: Estimación de los coeficientes de regresión para el modelo 1, mediante el Modelo Poisson con enlace Log.

Coeficiente	**Estimación**	**Error Estándar**	**z value**	**Pr**$(>\|z\|)$	
(Intercept)	4,362e+00	1,876e-02	232,495	$<$ 2e-16	***
accidentes2011	-5,386e-03	1,997e-04	-26,978	$<$ 2e-16	***
accidentes2012	8,246e-03	2,898e-04	28,457	$<$ 2e-16	***
accidentes2013	-4,553e-04	9,206e-05	-4,945	7,6e-07	***

Se puede observar de la Tabla 3.3 que para el modelo Poisson, todas las variables resultan ser significativas. Además se nota que, con respecto a la primera variable ésta disminuye en 5,286e-03 (0,000528) accidentes con respecto al año 2014, contrario a lo que sucede con los accidentes del 2012 ya que éstos aumentan en 8,246e-03 (0,0008246) accidentes con respecto al 2014.

De esta manera el modelo queda de la siguiente forma:

$$\text{Y} = 4,362\text{e+}00 - 5,386\text{e-}03X_1 + 8,246\text{e-}03X_2 - 4,553\text{e-}04X_3$$

Para saber que modelo se ajusta mejor con respecto a otro es que, como se mencionó anteriormente, se utilizará el criterio de AKAIKE (AIC).

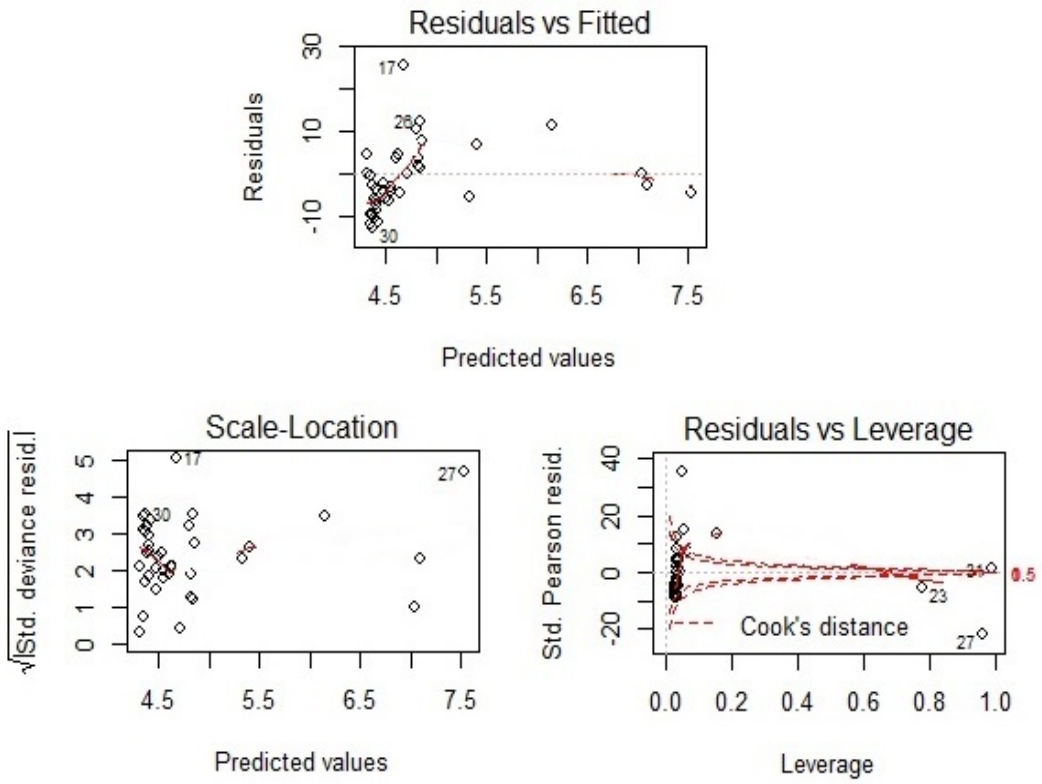
Para el modelo Poisson, considerando el modelo 1, el valor AIC es el siguiente:

$$AIC = 2562,264$$

Este valor se comparará posteriormente con el valor AIC entregado por el modelo Binomial Negativa, que según los test de Sobredispersión, debiese ser menor al valor entregado por el modelo Poisson, ya que para estos datos se observa que existe sobredispersión.

3.6 Diagnóstico del modelo 1 mediante el MRP con enlace Log lineal.

Figura 3.7: Diagnóstico del modelo 1



De lo anterior se observa que en el gráfico de residuos frente a pronósticos los residuos manifiestan una tendencia a la media. También, la condición de independencia de los errores parece cumplirse. Además, la dispersión vertical de los residuos es razonablemente pequeña.

3.7 Aplicación del Modelo de Regresión Binomial Negativa considerando el Modelo 1.

Se supone que el número de accidentes del año 2014 en la Región de Valparaíso son independiente al de otros para un modelo Binomial Negativa. Sea Y el número de accidentes del año 2014, entonces:

$$Y \sim BN(\mu_i, \phi)$$

En la Tabla 3.4 se presentan los estimadores de los coeficientes de regresión para el modelo 1:

Tabla 3.4: Estimación de los coeficientes de regresión para el modelo 1, mediante el modelo Binomial Negativa con enlace Log.

Coeficiente	**Estimación**	**Error Estándar**	**z value**	**Pr**$(>\|z\|)$	
(Intercept)	4,016377	0,167591	23,965	$<$ 2e-16	***
accidentes2011	0,000106	0,004290	0,025	0,980	
accidentes2012	0,001735	0,005882	0,295	0,768	
accidentes2013	0,001383	0,002081	0,665	0,506	

Se puede observar de la Tabla 3.4, que ninguna variable resulta ser significativa en el modelo (sólo el intercepto β_0), es decir, que para este modelo Binomial Negativa no existe relación entre los accidentes de 2014 con respecto a los accidentes del 2011, 2012 y 2013.

De esta manera el modelo queda de la siguiente forma:

$$\text{Y} = 4{,}016377 + 0{,}000106X_1 + 0{,}001735X_2 + 0{,}001383X_3$$

Para el Modelo de Regresión Binomial Negativa, considerando el modelo 1, el valor AIC es el siguiente:

$$AIC = 442,3643$$

Resumiendo, luego de aplicar el criterio AIC a ambos modelos, se obtuvieron los siguientes valores:

Tabla 3.5: Comparación final entre ambos Modelos de Regresión, considerando de modelo 1

	Modelo de Regresión Poisson	**Modelo de Regresión Binomial Negativa**
AIC	2562,264	442,3643

Con la Tabla 3.5 se corrobora lo que dice el test "dispersiontest" que, al haber Sobredispersión, el modelo que debería ajustarse mejor a estos datos de conteo debiese ser el MRBN por sobre el MRP, lo que claramente señala el criterio AKAIKE, ya que el valor AIC Poisson es mayor que el valor AIC Binomial Negativa.

Cabe destacar que no solamente existe el criterio AKAIKE (AIC) para saber que modelo se ajusta mejor a un cierto grupo de datos. Es por eso, que también se implementó un test llamado $odTest$ de la librería $pscl$.

Este test compara las log-verosimilitudes de un modelo Binomial Negativa y un modelo de regresión de Poisson.

El modelo Binomial Negativo suaviza el supuesto del modelo Poisson mediante la estimación de un parámetro adicional. La razón de verosimilitud (RV) se puede utilizar para probar la hipótesis nula de que la restricción implícita en el modelo Poisson es cierto.

La prueba estadística de la razón de verosimilitud (RV) tiene una distribución no estándar, incluso asintóticamente , ya que el exceso de dispersión de parámetros en la binomial negativa (denominada θ en glm.nb) se limita a ser positivo. Esto significa que en la prueba en el nivel $\alpha = 0,05$, no se debe rechazar la hipótesis nula a menos que la estadística de prueba de la razón de verosimilitud exceda el valor crítico asociado a $2\alpha = 0,10$. Esta prueba de la razón de verosimilitud implica sólo una restricción de parámetros, por lo que el valor crítico de la prueba estadística al nivel de $p = 0,05$ es $2,7$, en lugar del habitual $3,8$.

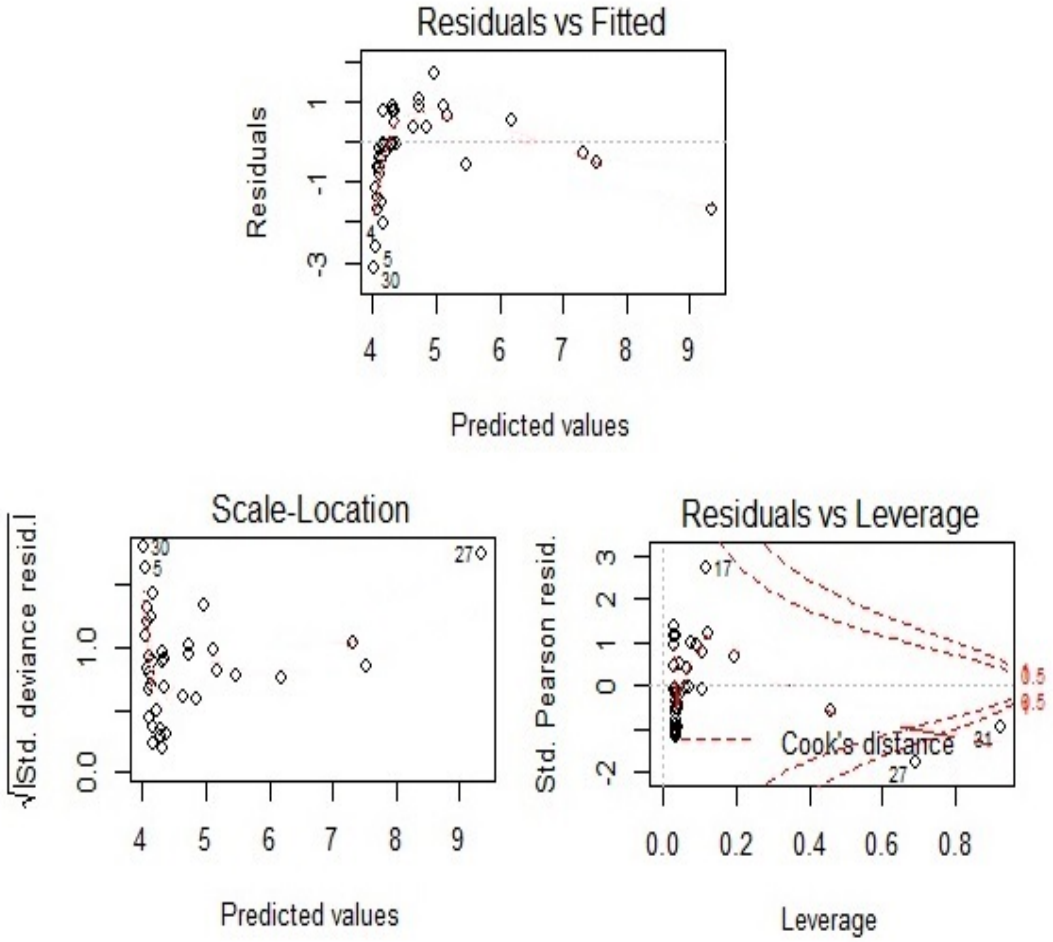
Este test se aplica sobre el modelo Binomial Negativa el cual arroja los siguientes valores, considerando el modelo 1:

level: 2,7055
Chi-Square Test Statistic = 2121,8995
p-value < 2,2e-16

Por lo anterior, se rechaza la hipótesis en que el modelo Poisson es mejor por sobre el modelo Binomial Negativa, debido a que la estadística de prueba tiene un valor = $2121,8995$ y excede al nivel $2,7055$ con un $p - valor <$ 2,2e-16.

3.8 Diagnóstico del modelo 1 mediante el MRBN con enlace Log lineal.

Figura 3.8: Diagnóstico del modelo 1.



De lo anterior se observa que en el gráfico de residuos frente a pronósticos los residuos manifiestan una leve tendencia a la media, esta tendencia no es tan fuerte como en el MRP pero se observa un comportamiento similar. También, la condición de independencia de los errores parece cumplirse. Además, la dispersión vertical de los residuos es razonablemente pequeña. Se puede sugerir un modelo cuadrático para estos casos.

3.9 Aplicación del Modelo de Regresión Poisson considerando el Modelo 2.

Para determinar si existe o no sobredispersión es que utilizará nuevamente el test.

Ahora, se observa el valor que arroja el test "dispersiontest" bajo el criterio de α:

$$\alpha = 26,42223$$

Entonces, con un valor de α = 26,42 y considerado lo que dice el test, se concluye que existe sobredispersión. Lo que indica que, para estos datos de conteo, debiese ser mejor el modelo Binomial Negativa por sobre el modelo Poisson, al igual que en el modelo 1.

Se observa también con este test, sin especificar $trafo$, el siguiente valor:

$$dispersion = 27,42223$$

Por lo tanto, se concluye que: con un valor de $dispersion = 27,42$, debiese ser mejor el modelo Binomial Negativa por sobre el modelo Poisson, confirmando lo que dice el criterio basado en α.

Se supone que el número de lesionados del año 2014 en la Región de Valparaíso son independiente al de otros para un Modelo de Regresión Poisson con parámetro μ_i. Sea Y el número de lesionados del año 2014, entonces:

$$Y \sim Poisson(\mu_i)$$

En la Tabla 3.6 se presentan los estimadores de los coeficientes de regresión para el modelo 2:

Tabla 3.6: Estimación de los coeficientes de regresión para el modelo 2, mediante el modelo Poisson con enlace Log.

Coeficiente	**Estimación**	**Error Estándar**	**z value**	**Pr**$(>\|z\|)$	
(Intercept)	4,0535907	0,0223920	181,029	< 2e-16	***
lesionados2011	0,0016066	0,0002401	6,690	2,23e-11	***
lesionados2012	-0,0034902	0,0004905	-7,116	1,11e-12	***
lesionados2013	0,0047614	0,0002865	16,620	< 2e-16	***

De esta manera el modelo queda de la siguiente forma:

$$\text{Y} = 4,0535907 + 0,0016066X_1 \text{ - } 0,0034902X_2 + 0,0047614X_3$$

Se puede observar de la Tabla 3.6 que para el modelo Poisson, todas las variables resultan ser significativas. Con respecto a la primera variable ésta aumenta en 0,0016066 lesionados con respecto al año 2014, contrario a lo que sucede con los lesionados del 2012 ya que éstos disminuyen en 0,0034902 con respecto al 2014.

Para saber que modelo se ajusta mejor con respecto a otro es que, como se mencionó anteriormente, se utilizará el criterio de AKAIKE (AIC).

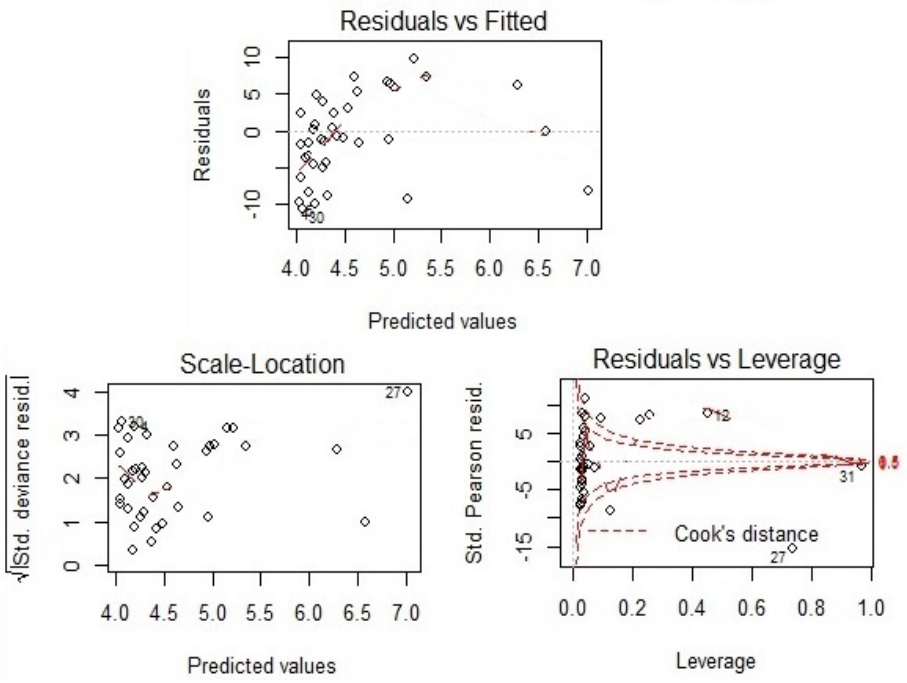
Para el modelo Poisson, considerando el modelo 2, el valor AIC es el siguiente:

$$AIC = 1445,02$$

Este valor se comparará posteriormente con el valor AIC entregado por el MRBN.

3.10 Diagnóstico del modelo 2 mediante el MRP con enlace Log lineal.

Figura 3.9: Diagnóstico del modelo 2



El gráfico de residuos frente a valores pronosticados señalan cómo los residuos no presentan una tendencia a la media, lo que implica el incumplimiento de la condición de independencia de los errores.

3.11 Aplicación del Modelo de Regresión Binomial Negativa considerando el Modelo 2.

Se supone que el número de lesionados del año 2014 en la Región de Valparaíso son independiente al de otros para un Modelo de Regresión Binomial Negativa. Sea Y el número de lesionados del año 2014, entonces:

$$Y \sim BN(\mu_i, \phi)$$

En la Tabla 3.7 se presentan los estimadores de los coeficientes de regresión para el modelo 2:

Tabla 3.7: Estimación de los coeficientes de regresión para el modelo 2, mediante el Modelo Binomial Negativa con enlace Log.

Coeficiente	**Estimación**	**Error Estándar**	**z value**	**Pr**$(>\|z\|)$	
(Intercept)	3,800393	0,151607	25,067	$<$ 2e-16	***
lesionados2011	0,001672	0,002349	0,712	0,4765	
lesionados2012	-0,003039	0,004062	-0,748	0,4545	
lesionados2013	0,005150	0,002843	1,811	0,0701	

Se puede observar de la Tabla 3.7, que ninguna variable resulta ser significativa en el modelo (sólo el intercepto β_0), es decir, que para este modelo Binomial Negativa no existe relación entre los lesionados de 2014 con respecto a los lesionados del 2011, 2012 y 2013.

De esta manera el modelo queda de la siguiente forma:

$$\text{Y} = 3{,}800393 + 0{,}001672X_1 \text{ - } 0{,}003039X_2 + 0{,}005150X_3$$

Si se analiza un poco más la última covariable del modelo, es decir, accidentes 2013 se puede observar que con un valor p = a 0,0701 podría ser significativa. O sea, esta variable podría tener una relación con respecto a los accidentes del 2014.

Para el modelo Binomial Negativa, considerando el modelo 2, el valor AIC es el siguiente:

$$AIC = 422,6567$$

Resumiendo, luego de aplicar el criterio AIC a ambos modelos, se obtuvieron los siguientes valores:

Tabla 3.8: Comparación final entre ambos Modelos de Regresión, considerando de modelo 2

	Modelo de Regresión Poisson	**Modelo de Regresión Binomial Negativa**
AIC	1445,02	422,6567

Con la Tabla 3.8 se corrobora lo que dice el test "dispersiontest" que, al haber Sobredispersión, el modelo que debería ajustarse mejor a estos datos de conteo debiese ser el MRBN por sobre el MRP, lo que claramente señala el criterio AKAIKE, ya que el valor AIC Poisson es mayor que el valor AIC Binomial Negativa.

Como se mencionó anteriormente, no solamente existe el criterio AKAIKE (AIC) para saber que modelo se ajusta mejor a un cierto grupo de datos. Es por eso, que también se implementó un test llamado $odTest$ de la librería $pscl$.

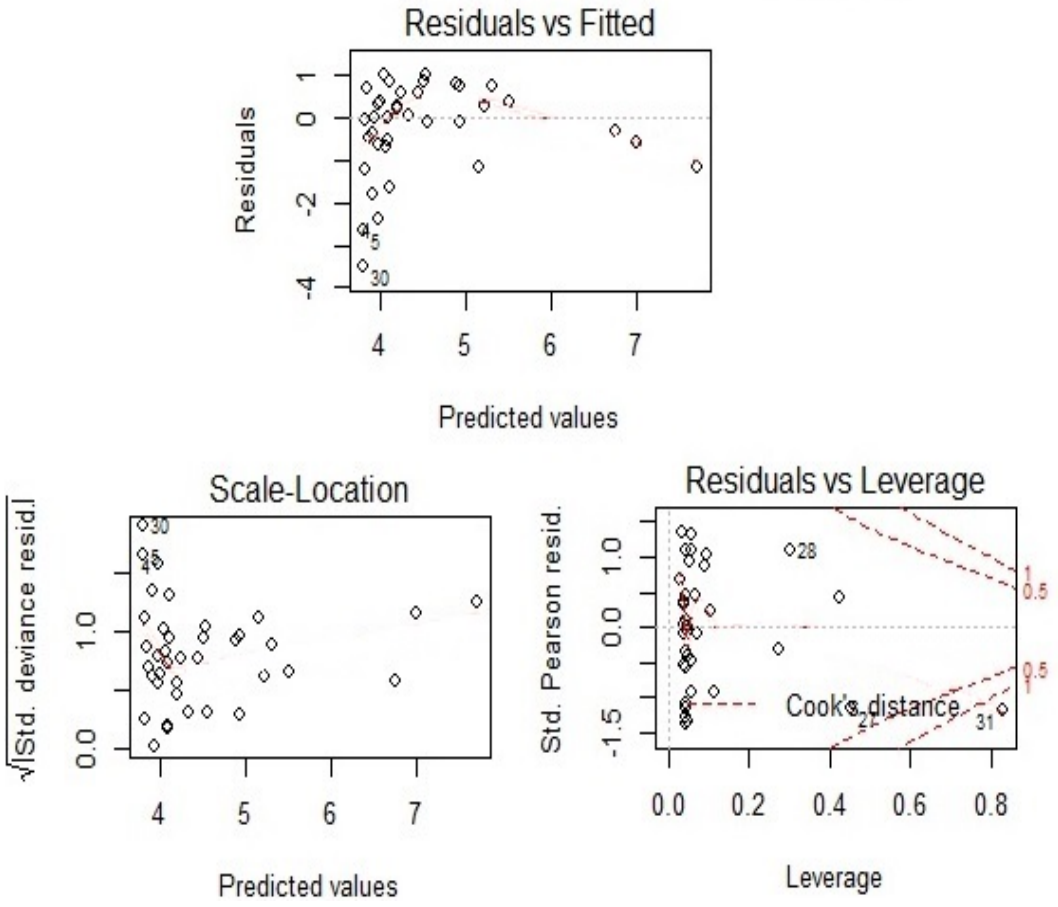
Este test se aplica sobre el modelo de Regresión Binomial Negativa el cual arroja los siguientes valores, considerando el modelo 2:

level: 2,7055
Chi-Square Test Statistic = 1024,3635
p-value < 2,2e-16

Por lo anterior, se rechaza la hipótesis en que el modelo Poisson es mejor por sobre el modelo Binomial Negativa, debido a que la estadística de prueba tiene un valor $= 1024,3635$ y excede al nivel $2,7055$ con un $p - valor <$ 2,2e-16.

3.12 Diagnóstico del modelo 2 mediante el MRBN con enlace Log lineal.

Figura 3.10: Diagnóstico del modelo 2



De lo anterior se observa que en el gráfico de residuos frente a pronósticos los residuos señalan una tendencia a la media. También, la condición de independencia de los errores parece ser cierta y cumplirse. Además, la dispersión vertical de los residuos es considerablemente pequeña.

Conclusión y Recomendación

3.13 Conclusiones

- Considerando estos datos de conteo, número de accidentes y lesionados de los años 2011 al 2014 de la Región de Valparaíso, es que se llegó a la conclusión que el mejor modelo era el MRBN por sobre el MRP, ya que se estaba en presencia de una Sobredispersión.
- Lo anterior se pudo concluir mediante el criterio AKAIKE (AIC) y el test $odTest$, ya que éstos ayudaron a escoger cual era el mejor para estos datos considerando los dos modelos.
- Se puede concluir también que no existe relación entre los accidentes del año 2014 con los accidentes de años anteriores (2011, 2012 y 2013) ya que al ser mejor el MRBN, éste no arroja ninguna variable significativa.
- En cuanto a los lesionados del año 2014 se puede concluir que tampoco existe relación con los lesionados de años anteriores (2011, 2012 y 2013) ya que al ser mejor el MRBN, éste no arroja ninguna variable significativa. A pesar de lo anterior es que se hace notar que lesionados 2013 podría ser significativa en los próximos años o si se estudia detalladamente podría serlo en este caso ya que su valor p es igual a 0,0701.

3.14 Recomendaciones

- Para el análisis de datos sobre el número de accidentes en la Región de Valparaíso o cualesquiera sean éstos, es conveniente tener ciertas consideraciones sobre datos de conteo. Conocer sus características y particularidades, para realizar una correcta aplicación.
- Es importante determinar si existe sobredispersión o no en los datos, es decir que la varianza sea mayor a la media, ya que sabiendo esto es que se decidirá, convenientemente, por un modelo adecuado.
- Existen diversas técnicas para medir la sobredispersión, eso quedará al criterio de cada investigador y de cual sea la más adecuada para los datos con los que se están trabajando.
- Se puede extender este estudio para el análisis de accidentes a nivel nacional, considerando otros años o agregando más variables de estudio para así determinar cuales podrían ser las causas de los accidentes a nivel país.
- En el caso que no se cumplan los supuestos en el diagnóstico se recomienda utilizar un modelo cuadrático.

Apéndice

3.15 Apéndice A - Código en R

```
### Librerías utilizadas ###

install.packages("MASS")
install.packages("glm2")
install.packages("modEvA", repos = "http://R-Forge.R-project.org")
install.packages("AER")
install.packages("pscl")

library(MASS)
library(glm2)
library(modEvA)
library(AER)
library(pscl)

### Lectura BD ###

Accidentes <- read.csv("DatosAccidentes.csv", header = T,sep=";")
attach(Accidentes)

### Análisis Exploratorio ###

summary(Accidentes)
```

```
### Comportamiento datos ###

plot(density(fallecidos2014),main = "", xlab = "Número de Fallecidos año 2014
, Región de Valparaíso", ylab = "Frecuencia",lwd=2)
hist(fallecidos2014,probability=T,main="Histograma",ylab="Densidad"
,xlab="Número de Fallecidos año 2014, Región de Valparaíso",col="pink")
lines(density(fallecidos2014))

### Modelo de Regresión Poisson ###

##### Modelo 1 #####

ModReg_Poisson1<-glm(formula = accidentes2014 ~ accidentes2011 + accidentes2012
 + accidentes2013, family=poisson(link = "log"),data=Accidentes)

summary(ModReg_Poisson1)

fit.modelpo = ModReg_Poisson1
stepAIC(fit.modelpo)

### Test de Sobredispersión ###

dispersiontest(ModReg_Poisson1)
dispersiontest(ModReg_Poisson1,trafo=1)

### Gráfico de los Residuos ###

par(mfrow=c(2,2))
plot(ModReg_Poisson1)

##### Modelo 2 #####

ModReg_Poisson2<-glm(formula = lesionados2014 ~ lesionados2011 + lesionados2012
 + lesionados2013, family=poisson(link = "log"),data=Accidentes)

summary(ModReg_Poisson2)

fit.modelpo = ModReg_Poisson2
stepAIC(fit.modelpo)
```

```
### Test de Sobredispersión ###

dispersiontest(ModReg_Poisson2)
dispersiontest(ModReg_Poisson2,trafo=1)

### Gráfico de los Residuos ###

par(mfrow=c(2,2))
plot(ModReg_Poisson2)

### Modelo de Regresión Binomial Negativa ###

##### Modelo 1 #####

ModReg_BinNeg1<-glm.nb(formula = accidentes2014 ~ accidentes2011 + accidentes2012
 + accidentes2013, link ="log",data=Accidentes)

summary(ModReg_BinNeg1)

fit.modelbn = ModReg_BinNeg1
stepAIC(fit.modelbn)
#### Selección mejor modelo ####

odTest(ModReg_BinNeg1)

### Gráfico de los Residuos ###

par(mfrow=c(2,2))
plot(ModReg_BinNeg1)

## Test Shapiro-Wilks ###

shapiro.test(residuals(ModReg_BinNeg1))

##### Modelo 2 #####

ModReg_BinNeg2<-glm.nb(formula = lesionados2014 ~ lesionados2011 + lesionados2012
 + lesionados2013, link ="log",data=Accidentes)

summary(ModReg_BinNeg2)
```

```
fit.modelbn = ModReg_BinNeg2
stepAIC(fit.modelbn)

#### Selección mejor modelo ####

odTest(ModReg_BinNeg2)

### Gráfico de los Residuos ###

par(mfrow=c(2,2))
plot(ModReg_BinNeg2)

### Gráfico de dispersión y coeficiente de correlación###

plot(accidentes2011,accidentes2014,pch=19, main="Gráfico de dispersión")
cor(accidentes2014,accidentes2011)

plot(accidentes2012,accidentes2014,pch=19, main="Gráfico de dispersión")
cor(accidentes2014,accidentes2012)

plot(accidentes2013,accidentes2014,pch=19, main="Gráfico de dispersión")
cor(accidentes2014,accidentes2013)

plot(lesionados2011,lesionados2014,pch=19, main="Gráfico de dispersión")
cor(lesionados2014,lesionados2011)

plot(lesionados2012,lesionados2014,pch=19, main="Gráfico de dispersión")
cor(lesionados2014,lesionados2012)

plot(lesionados2013,lesionados2014,pch=19, main="Gráfico de dispersión")
cor(lesionados2014,lesionados2013)
```

3.16 Apéndice B - Salidas computacionales.

- Test de sobredispersión, bajo el criterio α, para el Modelo 1 aplicando MRP.

```
        Overdispersion test

data:  ModReg_Poisson2
z = 2.2101, p-value = 0.01355
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
67.45195
```

- Test de sobredispersión, bajo el criterio *dispersion*, para el Modelo 1 aplicando MRP.

```
        Overdispersion test

data:  ModReg_Poisson2
z = 2.2101, p-value = 0.01355
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  68.45195
```

- odTest para el Modelo 1 aplicado al MRBN.

```
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  2121.8995 p-value = < 2.2e-16
```

- Test de sobredispersión, bajo el criterio α, para el Modelo 2 aplicando MRP.

```
        Overdispersion test

data:  ModReg_Poisson3
z = 5.8504, p-value = 2.452e-09
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
26.42223
```

- Test de sobredispersión, bajo el criterio *dispersion*, para el Modelo 2 aplicando MRP.

```
        Overdispersion test

data:  ModReg_Poisson3
z = 5.8504, p-value = 2.452e-09
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  27.42223
```

- odTest para el Modelo 2 aplicado al MRBN.

```
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  1024.3635 p-value = < 2.2e-16
```

Referencias Bibliográficas

- Comisión Nacional de Seguridad de Tránsito. Extraído el Sábado 02 de Julio de 2016 a las 15:46 desde http://www.conaset.cl/ a las 18:54 horas.
- Cameron, CA y Trivedi, PK (2005) Microeconometría: Métodos y Aplicaciones. Cambridge: Cambridge University Press.
- Contreras Vilca, N. (2012). Análisis de votos electorales usando modelos de regresión para datos de conteo.
- De la Fuente Fernández, S. (2011). Componentes Principales (ACP).
- Espinoza Morriberón, D. (2010). Estandarización de la CPUE de la flota industrial de cerco del stock norte-centro de anchoveta peruana (Engraulis ringens Jenyns 1842) entre 1996 y el 2008.
- Epidat 4: Ayuda de distribuciones de probabilidad. Extraído el Martes 16 de Septiembre de 2014 a las 10:19.
  desde http://www.sergas.es/gal/documentacionTecnica/docs/SaudePublica/Apli/Epidat4/Ayuda/Distribuciones%20de%20probabilidad.pdf
- Figueroa Arboccó, G. (2005). La fecundidad y su relación con variables socioeconómicas, demográficas y educativas aplicando el Modelo de Regresión Poisson.
- Gonzales King-keé, K. (2001). Método de mínimos cuadrados ponderados para la Estimación de los modelos lineales generalizados.
- Hachuel, L., Boggio, G., & Harvey, G. (2010). Modelos alternativos para el análisis de datos de conteo con exceso de ceros.
- McCullagh, P., & Nelder, J. A. (1991). Generalized linear models.

- Posada, S. L., & Rosero Noguera, R. (2009). Comparación de modelos matemáticos: una aplicación en la evaluación de alimentos para animales. *Revista Colombiana de Ciencias Pecuarias*, 20(2), 141-148.
- Cartes, F., Tudela, A., Vergara, L. (2004). Programa Comisión Nacional de Tránsito (CONASET).
- Romero Rodríguez, M. E., Arcos, E. L., Cano Fernández, V., & Sánchez Padrón, M. (2003). Análisis de citas de patentes a través de modelos de regresión para datos de recuento. *Estadística española*, 45(154), 455-478.
- Salinas-Rodríguez, A., Manrique-Espinoza, B., & Sosa-Rubí, S. G. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. *Salud pública de méxico*, 51(5), 397-406.
- Salmerón Gómez, R. El modelo Lineal General. Universidad de Granada. Extraído el Martes 16 de Septiembre de 2014 a las 23:12.
  desde http://www.ugr.es/ñomansg/material/WebEco/temas2y3.pdf
- Terrádez Gurrea, M. Análisis de Componentes Principales.
- Verdin Medina, L. A. (2005) Estimación de máxima verosimilitud en la distribución Weibull para muestras completas, censuradas y su aplicación en el análisis de tiempos de vida.