Universidad de Valparaíso
CHILE
Facultad de Ciencias
Instituto de Estadística

Aplicación del modelo Tweedie a datos de conteos sobre consultas médicas del sector privado de Chile realizadas en el año 2015

Presentado por :
Franco Daniel Jaime Bernales

Profesora Guía:
Claudia Navarro Villarroel, Ph.D

Valparaíso Chile
Diciembre, 2018

Agradecimientos

Principalmente quiero dar gracias a todo mi círculo familiar, en especial a mis padres Elisabeth Bernales y Luciano Jaime, que fueron mi principal apoyo durante todos estos años de estudio universitario. También, agradezco todos mis amigos y compañeros que me dieron muchos ánimos y consejos en este proceso.

Agradezco a mi profesora guía Claudia Navarro V., quien me ha dado mucho apoyo y sabiduría en el transcurso del año académico, siendo un pilar fundamental en la realización del trabajo de titulación. Agradezco, también, a mis otros profesores de carrera quienes me han dado muchos consejos y educación estadística en todo el ciclo correspondiente al pregrado universitario.

En especial quiero agradecer a mi padre Luciano Jaime, quien no me ha podido seguir en este hermoso momento (QEPD). A ti, que has sido mi gran motivación y también, a mi familia, les dedico este trabajo que ha sido fruto de mucho esfuerzo y dedicación.

Abreviaturas

En este apartado se presentan algunas notaciones o abreviaturas que se utilizaron a lo largo de este trabajo de titulación.

AIC	Criterio de Información de Akaike
CASEN	Encuesta de Caracterización Socioeconómica Nacional
DEIS	Departamento de Estadística e Información de Salud
EMV	Estimación Máxima Verosimilitud
MAE	Muestreo Aleatorio Estratificado
MD	Modelo de Dispersión
MDE	Modelo de Dispersión Exponencial
MLG	Modelo Lineal Generalizado
SERPLAC	Secretaría Regional de Planificación y Coordinación

Tabla de contenidos

Lista de figuras

Lista de tablas

Resumen

En este trabajo se estudia el modelo de distribución Tweedie, el cual es un subconjunto de los modelos de dispersión exponencial que se caracteriza por modelar datos discretos, continuos y mixtos. En la familia Tweedie existen distribuciones muy conocidas como la distribución Normal, Poisson, Gamma e Inversa Gaussiana, las cuales están determinadas por el parámetro de potencia p, que se encarga de dar forma a la distribución. Una de las características principales de la distribución es la de poder modelar datos que contengan ceros exactos, es por esto que, el objetivo principal del presente trabajo es aplicar y ajustar el modelo de distribución Tweedie a un conjunto de datos de conteos, provenientes del departamento de estadística e información de salud, del Ministerio de Salud de Chile.

Abstract

In this work we studied the Tweedie distribution model, which is a subset of the exponential dispersion models that is characterized by modeling discrete, continuous and mixed data. In the Tweedie family there are well-known distributions such as the Normal distribution, Poison, Gamma and Inverse Gaussian, which are determined by the power parameter p, that is responsible for shaping the distribution. One of the main characteristics of the distribution is that of being able to model data that contain exact zeros, that is why, the main objective of this work is to apply and adjust the Tweedie distribution model to a set of data about medical consultation counts, from the department of statistics and health information, for the Ministry of Health of Chile.

Objetivos

Las metas y objetivos planteados para realizar el trabajo de titulación se muestran a continuación:

Objetivo general

El objetivo principal para el presente trabajo de titulación es estudiar y aplicar el modelo de distribución Tweedie a un conjunto de datos reales sobre consultas y atenciones médicas del sector privado de Chile.

Objetivos específicos

- Demostrar empíricamente que la distribución Tweedie ajusta adecuadamente a los datos de consultas médicas.
- Determinar y comparar las diferencias de ajuste de los distintos grupos etarios incluidos en el conjunto de datos estudiado.

Hipótesis

Mediante el estudio de los modelos lineales generalizados y la distribución Tweedie, es posible determinar que el modelo es apropiado para ser utilizado en datos del área administrativa de la salud de Chile.

Capítulo 1

Introducción

Muchos autores se refieren a los modelos como un conjunto de herramientas que ayudan a estudiar las relaciones existentes entre las variables aleatorias. La clase de modelos más conocido es denominado modelos de regresión lineal clásicos, que estudia cómo se vincula linealmente las variables regresoras con la variable de interés (Mood y Graybill, 1978). Para que este tipo de modelos funcione correctamente y brinde el tipo de conclusiones deseadas, es necesario que cumplan con los supuestos de normalidad, linealidad entre las variables y varianza constante. Sin embargo, la gran diversidad de datos que existen actualmente en las distintas áreas de estudio, no siempre cumplen con los supuestos del modelo de regresión clásico, ya que la variable de interés pertenece a otra familia de distribuciones distinta a la Normal.

Para tratar con este tipo de problema, McCullagh y Nelder (1989) proponen una alternativa a los modelos clásicos, denominados modelos lineales generalizados (MLG). Montgomery, Peck y Vining (2002) afirman que el modelo lineal generalizado es una unificación de los modelos de regresión lineal y no lineal. En este tipo de modelos, la distribución de la variable de respuesta solo necesita pertenecer a la familia exponencial, que comprende las distribuciones Normal, Poisson, Binomial, Exponencial, Gamma, entre otras. Además, Montgomery *et al.* (2002), explican que el modelo lineal con error normal, no es más que un caso especial del MLG, por lo que en muchos casos se puede considerar que el modelo lineal generalizado unifica muchos aspectos del modelado y el análisis empírico de datos.

La teoría de los MLG se basa en los modelos de dispersión exponencial (MDE). Estos modelos son muy versátiles; ya que permiten datos de tipo discretos, continuos y de composición mixta (Jiang, 2007). Algunas de las distribuciones más conocidas forman parte de los MDE, éstas son; la Normal, Binomial, Poisson, Inversa Gaussiana, Exponencial y Gamma. Sin embargo, en el año 1987, el estadístico danés Jorgensen solidificó el concepto de los modelos de dispersión exponencial y nombró la clase de distribución Tweedie en honor al científico Maurice Tweedie.

La distribución Tweedie, introducida por el físico y estadístico Maurice Tweedie (1984), se caracteriza por modelar datos de tipo discretos y continuos. Esta distribución, según

Glen (2018), puede tener un grupo de elementos de datos ceros exactos, que lo hacen un candidato perfecto para modelar reclamos en la industria de seguros, consultas médicas, riesgos en estudios actuariales, entre otros.

Uno de los objetivos del presente trabajo, es aplicar el modelo de distribución Tweedie a un caso real. Para concretar esto, se trabajó con un conjunto de datos provenientes del área de la salud, relacionadas con conteos de consultas médicas, obtenidas directamente del departamento de estadística e información de salud (DEIS), del Ministerio de Salud de Chile.

Departamento de estadística e información de salud

El DEIS es el encargado de administrar, mantener y difundir información estadística pertinente, confiable y oportuna, dentro del marco definido por la autoridad sanitaria, participando en el diseño y la implantación de mecanismos de control y evaluación, que apoyen la formación de políticas, la planificación y la ejecución de las diversas acciones de salud, contribuyendo de esta manera al mejoramiento del nivel de salud de la población (Ministerio de Salud, 2015).

Conjunto de datos

Los datos fueron obtenidos directamente desde la página principal del Departamento de Estadísticas e Información de Salud, bajo la debida autorización de la organización. El conjunto de datos consiste en un total de 28.280 observaciones de consultas y atenciones médicas asociadas al sector privado de Chile en el año 2015. Cada unidad de observación, del conjunto de datos, es un conteo de consultas médicas que se realizaron bajo alguna especialidad médica, establecimiento de salud y período de mes en específico. El registro continuo fue efectuado en todos los establecimientos del sector privado, con ocasión de la realización de actividades de promoción, protección, recuperación y rehabilitación de la salud desde la región de Arica hasta la región de Magallanes. La cantidad de establecimientos del sector privado que reportaron el número de consultas médicas realizadas por cada mes del año y cada especialización médica, fue de 278 establecimientos, con un total de 34 especialidades médicas para las 73 comunas a lo largo del país. Los tipos de especialidades médicas asociados al sector privado de Chile se pueden ver en la Tabla 1.

Especialidades médicas	
• Anestesiología	• Rehabilitación
• Broncopulmonar	• Medicina interna
• Cardiocirugía	• Nefrología
• Cardiología	• Neonatología
• Cirugía infantil	• Neurocirugía
• Cirugía vascular	• Neurología
• Cirugía adulto	• Nutrilogía
• Cirugía plástica	• Obstetricia
• Cirugía máxilo facial	• Oftalmología
• Dermatología	• Otorrinolaringología
• Endocrinología	• Pediatría
• Geriatría	• Psicología
• Genética	• Reumatología
• Ginecología	• Salud ocupacional
• Gostroenterología	• Traumatología
• Hematología	• Urología
• Infectología	• Otras especialidades

Tabla 1: Tipos de especialidades médicas del sector sector privado

Cada unidad de observación (de un total de 28.280), es un conteo de consultas médicas asociadas a una especialidad de un establecimiento privado en específico, para cada uno de los meses del año. Al tratar con datos relacionados con conteos, es muy común encontrarse con situaciones en donde no hubo realizaciones de consultas médicas durante un mes y especialidad médica en específico. Dado este tipo de sucesos, el conteo se vuelve cero en determinado caso, habiendo en general un gran conjunto de datos equivalentes a ceros exactos.

Agrupación de variables

Los conteos generales de consultas médicas, para cada unidad de observación, se encuentran registrados por rangos etarios. Los rangos de edades considerados por el DEIS fueron los siguientes: menores de 10 años, 10 a 14 años, 15 a 19 años, 20 a 24 años, 25 a 39 años, 40 a 54 años, 55 a 64 años y 65 años o superior. En donde los conteos de atenciones médicas se resumen en los anteriores 8 vectores de datos. Para facilitar la manipulación e interpretación de estos datos, en el presente trabajo se propone agrupar estos subconjuntos de edades en grupos etarios más específicos, como lo son los grupos etarios establecidos por la Encuesta de Caracterización Socioeconómica Nacional (CASEN, 2015).

Grupos etarios

La agrupación o segmentación de la población en cuanto a grupos etarios específicos es muy importante. Es un paso muy necesario para comprender y conocer la realidad nacional y regional del país. Según los artículos publicados por la Secretaría Regional de Planificación y Coordinación (SERPLAC), la idea de segmentar la población en grupos etarios no solamente permite apreciar fenómenos demográficos, sino que ayuda a comparar las características que tienen los grupos de personas según la diferencia de edad, los problemas y las diferentes necesidades que presentan. Existen muchas diferencias entre los grupos de edades, ya que un adolescente no se enfrenta a los mismos desafíos que una persona adulta. Tampoco requieren de los mismos tipos de ayuda, pues sus necesidades evidentemente son distintas. Del mismo modo, la sociedad establece prioridades en materias de ayuda, asistencia o cooperación hacia los distintos grupos (SERPLAC, 2018). La clasificación para los grupos etarios considerados por esta organización se muestran en la siguiente Tabla 2.

Grupos etarios	
Adolescente	0 a 17 años
Joven	18 a 30 años
Adulto	31 a 59 años
Adulto mayor	60 años o superior

Tabla 2: Clasificación de grupos etarios por rangos de edades (SERPLAC, 2018)

De este modo, es posible obtener resultados independientes para cada grupo etario, de tal forma que se puede conocer y diferenciar numéricamente entre los cuatro grupos. En la Tabla 3, se evidencia el detalle global del total de consultas médicas y, también, el total de veces que no hubo realizaciones de consultas médicas para cada uno de los cuatro grupos etarios a lo largo del año.

Grupo	Cantidad de consultas médicas	Sin consultas médicas
Adolescente	3.065.859	4.645
Joven	4.039.959	4.128
Adulto	2.491.740	4.395
Adulto mayor	3.018.624	4.355

Tabla 3: Cantidad de consultas médicas generales y cantidad de veces que no hubo consultas médicas por cada grupo etario, en el año 2015

Sin embargo, para desarrollar un trabajo de investigación con resultados representativos a nivel nacional, es necesario utilizar técnicas de muestreo probabilístico. Fuller W. (2009), estadístico americano, explica que existen muchas razones que hacen del muestreo probabilístico una opción muy importante antes de realizar cualquier inferencia estadística.

En muchas investigaciones sobre determinadas poblaciones es necesario realizar metodologías de muestreo, dado que existen múltiples factores que pueden perjudicar una buena representación de la realidad en los datos utilizados. Según Pérez C. (2010), todos los métodos probabilísticos de muestreo están sustentados por la estructura del azar, lo cual es muy ventajoso a la hora de realizar estudios de investigación, ya que elimina por completo la subjetividad que podría influir en la elección de las unidades que integrarán la muestra. Además, el autor explica que el hecho de que una muestra esté basada en la probabilidad es muy enriquecedor en términos de información, debido que permitirá la aplicación de la inferencia estadística, haciendo que las conclusiones obtenidas tengan validez.

Muestreo estratificado

Para tener un tamaño de muestra representativo de la población en general, se debe realizar un muestreo aleatorio. El método de muestreo probabilístico más apropiado para el conjunto de datos obtenido, es el muestreo aleatorio estratificado (MAE). Este método funciona muy bien cuando se tiene una población muy heterogénea que se divide en subpoblaciones muy homogéneas entre sí, las cuales se denominan estratos.

Por diversas razones, este tipo de muestreo aleatorio resulta ser de mucha utilidad en las investigaciones. Al realizar la partición de la población general de N elementos en L subpoblaciones, el muestreo estratificado puede aportar información más precisa de algunas subpoblaciones que varían bastante en tamaño, pero que son homogéneas entre sí. También, el uso adecuado del MAE puede generar ganancia de información, pues al dividir la población heterogénea en estratos homogéneos, la realización del muestreo en estos estratos tendrá poco error debido precisamente a la homogeneidad (Pérez C. 2010).

En relación al conjunto de datos obtenido desde el DEIS, los estratos o subpoblaciones serían todas aquellas regiones del país de Chile. A partir de esto, se procede a calcular el tamaño de muestra estratificado para la población.

Tamaño de muestra estratificado

Para determinar un tamaño de muestra global, primero se debe establecer el error de muestreo ε y el nivel de confianza $1 - \alpha$ y, luego, para tener el tamaño de muestra estratificado se debe determinar el tipo de afijación más óptimo para repartir la muestra global "n" en los tamaños muestrales $n_1, n_2, ..., n_L$ de cada estrato de la población.

Luego, para el cálculo del tamaño muestral global n, se tomará una muestra aleatoria simple en cada estrato y se calculará una varianza piloto s_i^2. Seguido de esto, a partir de esta varianza piloto se podrá calcular el tamaño de muestra global n.

Las subpoblaciones que se considerarán para la realización del MAE, serán todas las regiones del país de Chile. Además, se propone utilizar una afijación o diseño de muestreo proporcional, el cual es muy adecuado para estimar parámetros de la población, cuando se tienen estratos muy diferentes entre sí. Así mismo, estratificar la población permite reducir el tamaño de muestra requerido para lograr una estimación con un nivel de error determinado.

Región (estrato)	N_h	W_h
Tarapacá	369	0,013
Antofagasta	1.069	0,038
Atacama	550	0,019
Coquimbo	583	0,021
Valparaíso	2.157	0,076
Libertador B. o'Higgins	1.028	0,036
Maule	308	0,011
Biobío	2.371	0,084
Araucanía	871	0,031
Los Lagos	1.040	0,037
Aysén	68	0,002
Magallanes	399	0,014
Metropolitana	16.460	0,582
Los Rios	652	0,023
Arica y Parinacota	355	0,013
Tamaño total de la población (N): 28.280		

Tabla 4: Distribución del tamaño poblacional N por cada estrato

La población bajo estudio son todos los pacientes que realizaron consultas médicas en los establecimientos del sector privado y la unidad de muestreo es la cantidad de consultas médicas realizadas en un establecimiento para cada especialidad médica, dentro de una determinada región en el período de un mes. De esta forma, en la Tabla 4, se tienen los estratos que dividen la población ($L = 15$), los cuales consisten en las 15 regiones del país de Chile. El valor N_h $(h = 1, 2, ..., L)$ indica la cantidad de observaciones medibles para cada región del país y, por otro lado, el valor W_h especifica la proporción de las observaciones de cada estrato respecto del total de la población.

Región (estrato)	Adolescente	Joven	Adulto	Adulto mayor
Tarapacá	18.439	24.790	11.584	10.255
Antofagasta	45.636	44.101	39.068	48.787
Atacama	30.818	53.884	35.388	33.257
Coquimbo	53.030	70.064	47.791	53.274
Valparaíso	121.503	135.249	97.470	176.376
Libertador B. O'higgins	118.315	91.379	70.488	110.861
Maule	40.842	53.499	30.483	29.840
BioBio	150.683	182.265	111.911	125.920
Araucanía	58.629	95.030	56.170	64.067
Los Lagos	50.921	55.401	38.709	41.060
Aisén	131	2.439	2.285	758
Magallanes	5.622	13.218	8.264	10.582
Metropolitana	2.339.195	3.180.833	1.917.160	2.285.203
Los Rios	27.351	31.679	21.126	23.828
Arica y Parinacota	4.744	6.128	3.843	4.556

Tabla 5: Cantidad de consultas médicas en el sector privado de Chile del año 2015 segmentados por estrato y grupo etario

Cada unidad de observación representa un conteo de consultas y atenciones médicas, para los distintos grupos de pacientes. De esta forma, en la Tabla 5, se evidencia la suma total de atenciones médicas realizadas en cada región del país de Chile y separado por los cuatro grupos etarios anteriormente definidos.

El procedimiento para hallar el tamaño de muestra ideal, tanto para la población general como para cada estrato, es mediante el uso de la fórmula (1). Primero, se debe realizar un muestreo aleatorio simple para cada uno de los estratos en estudio, luego, calcular las varianzas muestrales iniciales (s_h^2) para cada una de las subpoblaciones. Luego de esto, se multiplican con los ponderadores W_h ($W_h = N_h/N$) de cada estrato. Finalmente, se establece un nivel de confianza $1-\alpha$ y una determinada precisión ε para calcular los tamaños de muestra para cada grupo etario.

$$n = \frac{\sum_{i=1}^{L} \frac{W_i^2 s_h^2}{W_i}}{\left(\frac{\varepsilon}{Z_{1-\alpha/2}}\right)^2} \tag{1}$$

El detalle del cálculo de tamaño muestral de cada grupo se puede ver en la Tabla 6, en donde se utilizó un nivel de confianza del $(1-\alpha) = 95\,\%$ y una precisión del $\varepsilon = 0,5$. Según Fuller W. (2009), la idea central del muestreo aleatorio estratificado, es obtener un tamaño de muestra totalmente representativo para cada una de las subpoblaciones. Una vez teniendo el tamaño muestral global, se puede calcular el tamaño de muestra en cada estrato usando la siguiente fórmula $n_h = n(N_h/N) = nW_h$. Este algoritmo corresponde a un tipo de diseño proporcional, el cual cumple la función de repartir el tamaño de muestra global a todos los estratos de la población de manera proporcional al tamaño real del estrato.

Grupo etario	Adolescente	Joven	Adulto	Adulto Mayor
Tamaño de muestra estratificada "n"	16.088	21.963	15.635	20.992

Tabla 6: Tamaño de muestra global para cada grupo etario

Luego de tener el tamaño de muestra global y el tamaño de muestra proporcionado por cada estrato, se puede estimar la media global utilizando la fórmula:

$$\overline{x} = \sum_{h=1}^{L} W_h \overline{x}_h \tag{2}$$

La anterior ecuación (2), describe la media ponderada de las medias de todos los estratos de la población y, además de esto, calculando la varianza de la media muestral $S^2_{\overline{x}}$ se puede encontrar un intervalo de confianza para la media poblacional de consultas médicas, con el nivel de confianza y precisión anteriormente fijados. El detalle sobre la media poblacional de consultas medicas, como de la varianza e intervalo de confianza para cada uno de los grupos etarios se puede ver en la Tabla 7.

Grupo etario	Estimador de $\overline{x}$	Estimador de $S^2_{\overline{x}}$	Intervalo de confianza
Adolescente	190,555	0,459	$[189,247\,;191,862]$
Joven	183,942	0,739	$[182,258\,;185,625]$
Adulto	159,374	2,775	$[156,110\,;162,637]$
Adulto mayor	143,794	0,490	$[142,422\,;145,166]$

Tabla 7: Estimadores poblacionales e intervalos de confianza para las consultas médicas de cada grupo etario

Con los estimadores de la media poblacional y la varianza, detallados en la Tabla 7, se puede construir un intervalo de confianza para la media poblacional de consultas médicas de la siguiente manera:

$$\mu \in \left[\overline{x} \pm Z_{1-\alpha/2} S_{\overline{x}}\right] \tag{3}$$

Con $Z_{1-\alpha/2} = 1{,}96$. Finalmente, se puede decir que con un 95 % de confianza, los intervalos calculados contienen el verdadero valor de la media de consultas y atenciones médicas realizadas en el sector privado de Chile en el año 2015, para los diferentes grupos etarios.

Factor de expansión

En muestreo probabilísticos donde el diseño muestral es complejo (estratificado y multi-etápico), la probabilidad de inclusión de cada unidad de observación es muy diferente según su procedencia, en cuanto al estrato y a la unidad de muestreo, por lo tanto el número de unidades que representan en la población es distinto. Según estudios históricos de la CASEN 2011, existen dos factores de expansión para cada unidad de estudio: el factor que expande a la proyección de población regional y otro factor que expande a la población comunal (CASEN 2011).

Para que el análisis del estudio tenga validez sobre la población objetivo, se utiliza un ponderador en la estimación de la variable de interés, el cual tiene relación con las probabilidades de selección de las distintas unidades de muestreo. Este ponderador, formalmente denominado factor de expansión, hace representación al individuo o unidad de observación que participa en el estudio, respecto al número total de individuos u observaciones de la población (CASEN 2015).

Mediante el análisis exploratorio del conjunto de datos, se pudieron analizar las proporciones de las observaciones para cada una de las regiones y comunas del país. Una vez hecho esto, se utilizaron las 15 regiones del país de Chile como estratos del MAE y, para tener mayor representatividad en la población, se consideró realizar factores de expansión para las 73 comunas del país. Sin embargo, dentro de todas las comunas que participaron del estudio, sólo 4 regiones del país contenían comunas con proporciones muy similares de observaciones. En cuanto a las 11 regiones restantes, más del 85 % de las observaciones pertenecen a una sola comuna y, por consiguiente, el resto de las comunas poseían una proporción muy baja de observaciones. Por lo tanto, al conocer esta gran desigualdad en las proporciones de las comunas que fueron objeto del estudio, se decidió no utilizar factores de expansión comunales, ya que no prestarían mayor representatividad para este estudio.

Una vez introducido la temática correspondiente al trabajo de investigación, en el siguiente capítulo se presentará la metodología utilizada para alcanzar los objetivos de trabajo.

Capítulo 2

Metodología

A continuación, en este capítulo se presentan la definición y propiedades de la distribución perteneciente a la familia de dispersión exponencial denominada Tweedie, como también una introducción a los modelos lineales generalizados.

En el área de la ciencia estadística, los modelos forman un conjunto de herramientas que ayudan a estudiar las relaciones existentes entre las variables aleatorias (Montgomery, Peck y Vining, 2006). Estas variables pueden relacionarse de manera lineal, a través de los modelos lineales clásicos, o bien, cuando los supuestos de normalidad y linealidad entre variables no se cumple, pueden relacionarse de manera no lineal. Dado esto último, McCullagh y Nelder (1989) propusieron una clase de modelos más general, denominada modelos lineales generalizados, los cuales amplían el marco de los modelos lineales clásicos a la clase de distribuciones de la familia exponencial.

Modelos lineales generalizados

En muchas ocasiones, cuando los supuestos del modelo no se cumplen debido a que la variable respuesta sigue otra distribución diferente a la normal, es muy frecuente utilizar transformaciones en la variable para poder corregir estos problemas. Sin embargo, estas transformaciones no siempre consiguen solucionar la falta de normalidad, la heterocedasticidad (varianza no constante) y la no linealidad de los datos. Una alternativa a la transformación de la variable de respuesta y a la falta de normalidad, es el uso de los modelos lineales generalizados, los cuales son una extensión de los modelos lineales clásicos. Los MLG son muy útiles, ya que permiten el uso de distribuciones con errores no normales y varianzas no constantes como por ejemplo la distribución Poisson, Gamma, entre otros (Cayuela, 2010).

Muchos tipos de variables sufren irremediablemente la violación de los supuestos para los modelos normales y los MLG otorgan una buena alternativa para poder tratarlos. Específicamente, se puede considerar la utilización del MLG cuando la variable respuesta es de tipo binaria (éxito o fracaso de un evento determinado), de proporciones (tasa de mortalidad, porcentaje de desempleo, entre otros) o de tipo conteo (cantidad

de reclamos, consultas u otros). Un supuesto central en los modelos lineales es que la varianza es constante. Sin embargo, en el caso de tener datos tipo conteo, en donde la variable respuesta está expresada con números enteros y, también, donde se hace muy recurrente la existencia de ceros en los datos, la varianza podría incrementar linealmente con la media (Cayuela, 2010).

Los autores McCullagh y Nelder (1989), explican que los MLG tienen propiedades muy importantes como lo es la estructura de sus errores y la función de enlace. Estas propiedades la definen en las siguientes forma:

- **Componente aleatoria**: corresponde a la variable aleatoria Y que sigue una distribución de la familia exponencial. Nelder y Wedderburn (1972) explican que la distribución de cada elemento Y_i con $i = 1, 2, ..., n$, pertenece a una familia exponencial, en el sentido de que la función de densidad de probabilidad para cada Y_i tiene la forma:

$$f(y; \theta, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{4}$$

  para algunas funciones específicas $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$, θ_i es el parámetro canónico para los modelos de la familia exponencial, ϕ es el parámetro de escala que permanece constante para todo i.

- **Componente sistemática**: denominada predictor lineal, matemáticamente denotada por η y corresponde al vector de n componentes, siendo cada una de ellas igual a:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} \tag{5}$$

- **Función de enlace**: es la encargada de relacionar la esperanza matemática de la variable respuesta con el predictor lineal, $g(\mu_i) = \eta_i, i = 1, 2, 3, ..., n$. De esta forma, la función de enlace, $g(\cdot)$, relaciona la componente aleatoria con la componente sistemática de la siguiente forma:

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij} \tag{6}$$

Existen funciones de enlace canónicos para algunas de las distribuciones más conocidas, las cuales se aplican por defecto a cada una de las distribuciones. Sin embargo, esto no significa que siempre se deba usar una única función de enlace para una determinada distribución. En la Tabla 8, se puede ver el detalle sobre las funciones de enlace más conocidas desarrollada por McCullagh y Nelder (1989).

Función de enlace	Fórmula	Contexto de uso
Identidad	μ	Datos continuos con distribución Normal
Logarítmica	$Log(\mu)$	Datos de conteos con distribución Poisson
Logit	$Log(\frac{\mu}{n-\mu})$	Datos de proporciones con distribución Binomial
Recíproca	$1/\mu$	Datos continuos con distribución Gamma

Tabla 8: Funciones de enlace canónicos más utilizadas para los modelos lineales generalizados (McCullagh y Nelder 1989 y Cayuela, 2010)

Los modelos generalizados se relacionan directamente con los modelos de dispersión exponencial, ya que los tipos de datos que puede tener la variable respuesta son casos específicos de la familia de distribuciones exponencial.

Modelos de dispersión exponencial

Los modelos lineales generalizados fueron desarrollados originalmente para la comprensión y análisis de aquellos datos que se escapan de los supuestos tradicionales de normalidad, es decir, para los datos que forman parte de la familia de distribuciones exponenciales. Sin embargo, las ideas principales pueden extenderse a modelos de clase más amplia conocidos como modelos de dispersión exponencial (MDE) (Jorgensen, 1997).

Los MDE, principal interés en el presente trabajo, son una clase de distribuciones que se caracterizan por tener dos parámetros de la familia exponencial lineal, con un parámetro adicional de dispersión. Estos modelos son muy importantes dentro de la estadística, ya que son las distribuciones de las variables de respuesta para los modelos lineales generalizados (Dunn y Smyth, 2005). La media y varianza están dados por:

$$E(Y) = \mu \tag{7}$$

$$Var(Y) = \phi V(\mu) \tag{8}$$

En donde V($\cdot$), es la función de varianza, la cual es una característica bien presente entre los modelos de dispersión exponencial, ya que describe la relación entre la media y la varianza de la distribución cuando la dispersión se mantiene constante (Jorgensen, 1987). Si la variable Y sigue una distribución perteneciente a los MDE con media μ, función de varianza V($\cdot$) y parámetro de dispersión ϕ, entonces la varianza de Y está dada por:

$$Var(Y) = \phi V(\mu) = \phi \mu^p \tag{9}$$

donde el valor de la constante p se encarga de dar forma a la distribución. Jorgensen (1997), también habla sobre la idea principal detrás de los MDE, la cual se centra en que la ubicación y escala pueden generalizarse a posición y dispersión, respectivamente. De manera similar, la suma residual de cuadrados del análisis de varianza, se puede generalizar a la noción de devianza, haciendo que el análisis de devianza esté disponible como una herramienta de inferencia general para una amplia gama de datos.

Considerar las observaciones independientes $y_1, ..., y_n$ con distribuciones $Y_i \sim \text{DM}(\mu,\sigma^2)$, donde $\text{DM}(\mu,\sigma^2)$ denota un modelo de dispersión, por lo que σ^2 es desconocido, pero común para todas las observaciones. Se define la devianza para un vector de parámetros $\mu = (\mu_1, ..., \mu_n)^\top$ en base a una muestra $y = (y_1, ..., y_n)^\top$ por:

$$D(y;\mu) = \sum_{i=1}^{n} d(y_i;\mu_i) \tag{10}$$

Dada la interpretación de la devianza $d(y;\mu)$ como análoga a la distancia de cuadrados $(y-\mu)^2$, la devianza total puede considerarse una generalización de la conocida suma de cuadrados residual proveniente de la teoría normal (Jorgensen, 1997).

Los modelos pertenecientes a la familia de dispersión exponencial se caracterizan por modelar datos de cualquier ámbito, ya sean datos discretos, continuos o mixtos. Algunas de las ditribuciones que son parte de esta familia son: Normal, Binomial, Poisson, Inversa Gausiana, Exponencial, Gamma y Tweedie. La distribución Binomial y Poisson son de índole discreta y utilizadas cuando los datos son de conteo. Por otro lado, las distribuciones Normal, Inversa Gausiana, Gamma y Exponencial, son utilizadas cuando los datos son de carácter continuo. Por último, la distribución Tweedie es de naturaleza mixta, esto significa que se pueden modelar datos con componentes tanto discretos como continuos, un ejemplo de ésto sería la distribución compuesta Poisson con Gamma (Swan, 2006).

Distribución Tweedie

La distribución Tweedie, introducida por el físico y estadístico Maurice Tweedie (1984), es un subconjunto de los MDE y se caracteriza por modelar datos de tipo discretos, continuos y mixtos. Esta distribución, según Glen (2018), puede tener un grupo de elementos de datos en cero, por tanto, es un buen candidato para modelar reclamos en la industria de seguros, pruebas médicas, riesgos en estudios actuariales, entre otros.

El estadístico danés Jorgensen (1987), solidificó el concepto de los modelos de dispersión exponencial y nombró la clase Tweedie en honor al científico Maurice Tweedie. De este modo, una variable aleatoria Y, perteneciente a los MDE, sigue esta distribución si en su función de varianza, la cual relaciona la media y la varianza de la distribución, existe una constante p denominada parámetro de potencia, la cual se encarga de proporcionar la forma a la distribución.

Entonces, en términos más formales, se tiene que la variable $Y \sim TW_p(\mu, \phi)$, la cual denota una variable aleatoria Tweedie con media μ y varianza $\phi\mu^p$, tal que $\phi > 0$ y $p \in (-\infty, 0] \cup [1, \infty)$ los cuales son los parámetros de dispersión y de potencia, respectivamente (Bonat, 2017).

Características de la distribución Tweedie

Según Kaas (2005) y Dunn (2005), la distribución Tweedie incluye la mayoría de las distribuciones más importantes y comúnmente asociadas con los modelos lineales generalizados, dado distintos valores del parámetro p, éstas son:

1. Distribución Normal, cuando el parámetro de potencia $p = 0$.
2. Distribución Poisson, cuando el parámetro de potencia $p = 1$.
3. Distribución Gamma, cuando el parámetro de potencia $p = 2$.
4. Distribución Inversa Gaussiana, cuando el parámetro de potencia $p = 3$.

Kaas (2005), afirma que el modelo Tweedie existe para todos los valores del parámetro de potencia p, excepto para el intervalo abierto $]0, 1[$. Por otro lado, el rango más interesante está entre los valores $1 < p < 2$. En este rango, el modelo Tweedie se expresa como una distribución compuesta Poisson con Gamma. Shono (2008) escribió matemáticamente esta distribución como sigue:

$$Y = \sum_{i=1}^{N} X_i \quad ; \ con \ \ N = 1, 2, 3, 4, ... \tag{11}$$

donde $N \sim \text{Poisson}(\lambda)$ y $X_i \sim \text{Gamma}(\alpha, \theta)$, siendo N una distribución Poisson con parámetro λ y, por otro lado, $X_1, X_2, ..., X_N$ son variables aleatorias independientes e idénticamente distribuidas Gamma con parámetro de forma α y parámetro de escala θ. Cuando $X = 0$, la densidad es una masa de probabilidad que se rige por la distribución Poisson. Por el contrario, si $X > 0$, la densidad es una mezcla de variables Gamma con Poisson.

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)} \qquad \alpha = \frac{2-p}{p-1} \qquad \theta = \phi(p-1)\mu^{p-1}$$

Los parámetros de la distribución mixta Poisson Gamma, recientemente detallados, están estrechamente relacionados con los parámetros naturales de la distribución Tweedie.

Densidad Tweedie

Los modelos de dispersión exponencial tienen funciones de densidad de probabilidad o masa de probabilidad, de la siguiente forma:

$$f(y;\mu,\phi) = a(y;\phi)\ exp\left[\frac{y\theta - k(\theta)}{\phi}\right] \quad ,\ y \in \mathbb{R} \tag{12}$$

para las funciones conocidas $a(\cdot)$ y $k(\cdot)$. El parámetro canónico θ pertenece al intervalo abierto que satisface $k(\theta) < \infty$ y el parámetro de dispersión ϕ es positivo. La función $k(\cdot)$ es denominada función acumulativa para los MDE, porque si $\phi = 1$, las derivadas para la función $k(\cdot)$ deben dar las las sucesiones acumulantes de la distribución (Dunn y Smyth, 2007). En particular, la media de la distribución es $\mu = k'(\theta)$ y la varianza es $\phi k''(\theta)$. El mapeo de θ a μ es invertible, por lo que se puede escribir $k''(\theta) = V(\mu)$, para una función adecuada $V(\cdot)$ llamada función de varianza de los MDE.

La distribución Tweedie generalmente no tiene una función de densidad que se pueda escribir de forma cerrada. No obstante, tiene funciones generadoras de momentos muy simples, por lo que la densidad se puede evaluar numéricamente mediante la inversión de Fourier en términos de su función generadora acumulante (Dunn y Smyth, 2007). La función generadora de momentos, está dada por:

$$M(t) = \int exp(ty) f(y;\mu,\phi) dy \tag{13}$$

sustituyendo la ecuación (12) en $M(t)$ y completando la integral, muestra que la función generadora cumulante es:

$$K(t) = logM(t) = \frac{[k(\theta + t\phi) - k(\theta)]}{\phi} \tag{14}$$

Dunn y Smyth (2007) explican que existe otra forma para la función de probabilidad que es más conveniente que (12) para algunos propósitos. Diferenciando $log\ f$ muestra que $f(y;\mu,\phi)$ se maximiza con respecto a μ en $\mu = y$. Este cálculo supone que el apoyo para y está contenido en el dominio de μ, que es cierto para los modelos de dispersión exponencial. Considérese $t(y,\mu) = y\theta - k(\theta)$, entonces la unidad de devianza $d(y,\mu) = 2(t(y,y) - t(y,\mu))$ puede verse como una medida de distancia que satisface $d(y,y) = 0$ y $d(y,\mu) > 0$, para $y \neq \mu$. Un ejemplo de unidad de devianza es el caso de la distribución normal, la cual se expresa como $d(y,\mu) = (y-\mu)^2$.

Entonces, la función de probabilidad puede ser reescrita en términos de la unidad de devianza, de la siguiente forma:

$$f(y;\mu,\phi) = b(y,\phi)exp\left\{-\frac{d(y,\mu)}{2\phi}\right\} \quad , \; y \in \mathbb{R} \tag{15}$$

donde la función $b(y,\phi) = f(y;y,\phi)$, el cual es denominado como modelo de dispersión de la función de probabilidad.

Identidad de re escalamiento

Una propiedad fundamental de las densidades del modelo Tweedie es la identidad de re escalamiento, la cual permite que las funciones puedan tener expresiones cerradas bajo una nueva escala. Considérese la transformación $Z = cY$ para algunos $c > 0$, donde la variable Y sigue una distribución de modelo Tweedie con media μ y función de varianza $V(\mu) = \mu^p$. Encontrar la función generadora acumulada para Z, revela que sigue una distribución Tweedie con el mismo parámetro p, media $c\mu$ y dispersión $c^{2-p}\phi$. Mientras tanto, el Jacobiano de la transformación es $1/c$ para todos los valores de $y > 0$. Estos dos eventos juntos otorgan una identidad de re escalamiento extremadamente útil (Dunn y Smyth, 2007).

$$f(y;\mu,\theta) = cf(cy;c\mu,c^{2-p}\phi) \tag{16}$$

para todo p, $y > 0$ y $c > 0$. Esta identidad permite seleccionar los valores de y, como también, los valores para los parámetros que son favorables en la evaluación numérica y obtener la densidad en otros valores mediante el re escalamiento.

Finalmente, los autores Dunn y Smyth (2007), explican que se puede evaluar la densidad Tweedie encontrando las funciones $a(\cdot)$ y $b(\cdot)$ para las ecuaciones (12) y (15), utilizando la identidad de re escalamiento y aplicando la inversión de Fourier.

Modelo Tweedie

El verdadero interés está en los MDE con función de varianza de la forma $V(\mu) = \mu^p$, para algunos $p \geqslant 1$. La función acumulativa $k(\cdot)$ para los modelos de dispersión exponencial Tweedie se puede encontrar al igualar $k''(\theta) = \partial\mu/\partial\theta = \mu^p$ y resolver para $k(\cdot)$.

Sin perdida de generalidad, se puede elegir $k(\cdot) = 0$ y $\mu = 1$, para $\theta = 0$, que daría como resultado:

$$\theta = \begin{cases} \frac{\mu^{1-p}-1}{1-p} & \text{si } p \neq 1 \\ log(\mu) & \text{si } p = 1 \end{cases} \tag{17}$$

con inversa, $\mu = \tau(\theta) = [\theta(1-p)+1]^{1/(1-p)}$ para $p \neq 1$ y además,

$$k(\theta) = \begin{cases} \frac{\mu^{2-p}-1}{2-p} & \text{si } p \neq 2 \\ log(\mu) & \text{si } p = 2 \end{cases} \tag{18}$$

téngase en cuenta que las definiciones para θ y $k(\theta)$ son continuas tanto en p como en θ. La anterior expresión para $k(\theta)$, implica que la función generadora acumulada (14) tiene una forma analítica muy simple, sin embargo, no es el caso para las funciones $a(\cdot)$ de la ecuación (12) y $b(\cdot)$ de la ecuación (15), ya que no poseen expresiones de forma cerrada.

Las únicas excepciones en las que $a(\cdot)$ y $b(\cdot)$ se pueden obtener analíticamente son las distribuciones conocidas para $p = 1$, $p = 2$, $p = 3$ y, también, en $y = 0$ para valores del parámetro de poder entre $1 < p < 2$, donde existe una masa de "ceros" igual a $f(0; \mu, \theta) = exp[-\mu^{(2-p)}/\phi(2-p)]$.

Duun y Smyth (2005), describen la notación del modelo Tweedie como $Y \sim ED_p(\mu, \phi)$, que indica una variable aleatoria Y que distribuye como un modelo de dispersión exponencial Tweedie con media μ, dispersión ϕ y parámetro de potencia $p \in \mathbb{R}$, excepto para el intervalo abierto entre 0 y 1.

Estimación de parámetros

Para poder realizar un ajuste del modelo a un conjunto de datos reales, se necesitan las estimaciones de los valores que toman los parámetros. El método de estimación máxima verosimilitud (EMV), se utiliza para estimar los parámetros de los modelos lineales y no lineales (Dobson, 2002). Primero que todo, para obtener los estimadores de máxima verosimilitud de los parámetros β_j, se necesita la función de probabilidad. En general, la función de probabilidad se escribe de la siguiente forma,

$$L(\beta; y) = \prod_{i=1}^{n} f(y, \beta) \tag{19}$$

donde n es el tamaño de muestra para el conjunto de datos reales y β es el parámetro de interés. Para la estimación de los parámetros, resulta mejor utilizar la función de log-máxima verosimilitud, que se define del siguiente modo,

$$\begin{aligned} l(\beta, y) &= logL(\beta; y) \\ &= log \prod_{i=1}^{n} f(y, \beta) \\ &= \sum_{i=1}^{n} logf(y; \beta) \end{aligned} \tag{20}$$

Para poder utilizar esta teoría en los modelos lineales generalizados, la función de probabilidad logarítmica debe aplicarse a los modelos de dispersión exponencial dada en (12). Entonces, la log-máxima verosimilitud para los MDE está dado por,

$$l(\theta, \phi; y) = \sum_{i=1}^{n} a(\phi, y) + \frac{1}{\phi}[y\theta - k(\theta)] \tag{21}$$

Los estimadores de probabilidad máxima para los parámetros β_j se pueden hallar derivando la ecuación (21) con respecto a β_j. Esto se puede encontrar a través de las siguientes derivadas parciales,

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{22}$$

La combinación de estas cuatro derivadas muestra que la ecuación (22), se puede escribir como la ecuación *score* para los MLG,

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - \mu_i)}{V(\mu_i)}\frac{x_{ij}}{g'(\mu_i)} \tag{23}$$

Entonces, los EMV para los modelos lineales generalizados, se pueden hallar igualando la ecuación (23) a 0 y resolviendo para $j = 1, 2, ..., r$. Cuando $\partial l/\partial \beta_j = 0$, el valor del parámetro ϕ no necesita ser conocido. Esto es un concepto muy importante para los MLG, ya que se puede encontrar una estimación para β_j sin conocer ϕ (Dobson, 2002).

Software R

Para aplicar el modelo Tweedie se utilizó el programa de libre distribución *R project*. Este *software* contiene una serie de funciones muy útiles para la aplicación del modelo lineal generalizado Tweedie, como las librerías "tweedie" y "statmod" desarrolladas por Dunn y Smyth en el año 2005. Principalmente, se utilizaron las funciones "*tweedie.profile*" y "*tweedie family*". La función *tweedie profile*, se encarga de encontrar la distribución Tweedie más apropiada para el conjunto de datos, utilizando métodos de máxima verosimilitud. Esta función sólo trabaja para valores de $p \geqslant 1$ y proporciona el valor de probabilidad máxima de p y ϕ y un intervalo de confianza del 95 % para p. El procedimiento para ajustar un modelo lineal generalizado con una distribución Tweedie es, primero que todo, especificar la potencia de variación, es decir, el valor de p máximo encontrado a través de la función "*tweedie.profile*" y, también, la función enlace más apropiada para el modelo. La función enlace predeterminada y natural para este tipo de modelos es la canónica $1 - p$ y la otra función de enlace es la logarítmica, que se muestra en la ecuación (17).

Diagnóstico del modelo

En los estudios y análisis de resultados obtenidos a través de la regresión lineal y no lineal, es muy importante examinar exhaustivamente el modelo antes de interpretar cualquier resultado. El propósito de ajustar un modelo es resumir adecuadamente las características importantes de los datos al encontrar un modelo adecuado que explique lo que está sucediendo en los datos. Al postular un modelo, este puede presentar desviaciones de los datos y, por lo tanto, no ajustarse a la distribución predeterminada (Smolárová 2017). Es por esto que existen las pruebas de diagnóstico descriptivas, para determinar si el modelo se ajusta adecuadamente a los datos. Para los MLG existen una serie de pruebas de diagnóstico disponibles, como por ejemplo: gráficas *Q-Q plot*, diagramas de dispersión de los residuos y covariables, comparación de tamaños residuales, desviaciones residuales y en general, análisis de los residuos del modelo para investigar sobre posibles patrones en los datos (Swan, 2006). Estas técnicas permiten evaluar la idoneidad de la función de enlace y la distribución supuesta, así como la prueba de los datos para valores influyentes, valores atípicos o posibles patrones. Hay varios casos en los que un MLG ajustado puede no representar de manera apropiada a los datos, uno de estos casos se presenta cuando un modelo se ajusta bien a la mayoría de los datos. Sin embargo, algunos casos aislados no lo hacen, estos se denominan valores atípicos. Otros casos en el que un modelo generalizado ajustado no representa bien a los datos, se debe a la incorrecta especificación de la función de enlace o la variable de respuesta (McCullagh y Nelder, 1989).

Residuos del modelo

Los residuos constituyen la principal herramienta para el diagnóstico del modelo, puesto que son las estimaciones de las perturbaciones. Entonces, la comprobación de la idoneidad o adecuación del modelo, se puede realizar representando los residuos (Swan, 2006). En los modelos de regresión simple, se utilizan los residuos en bruto $(y - \hat{y})$. Sin embargo, estos son inadecuados cuando se tiene un MLG. Los tipos de residuos más comunes para el caso de MLG son los residuos de Pearson y los residuos de desviación. Los residuos de Pearson, tienen una distribución aproximadamente Normal $N(0, \phi)$, por otro lado, los residuos de la desviación están relacionados con los conceptos de desviación $D(y; \mu)$, y también poseen una distribución aproximadamente Normal (McCullagh y Nelder, 1989). A partir de la información obtenida del modelo ajustado, se puede realizar una serie de gráficos con respecto a los residuos de Pearson o residuos de devianza. A partir de esto, y sabiendo que los residuos del modelo deberían ser aleatorios, cualquier patrón que se observe en los gráficos de los residuos es un indicador de que hay problemas con el modelo ajustado (Swan, 2006). Los gráficos más apropiados para los residuos son los de Pearson *versus* valores predichos, residuos naturales *versus* observaciones y residuos de Pearson *versus* el predictor lineal.

A continuación, en el capítulo siguiente, se presenta la aplicación de la metodología estudiada a un conjunto de datos proveniente del área administrativa de la salud, relacionada con conteos de consultas médicas.

Capítulo 3

Aplicación

En este capítulo se presenta los resultados obtenidos mediante el uso del modelo lineal generalizado con familia de distribuciones Tweedie. Para la aplicación de este modelo, se cuenta con un conjunto de datos en donde la variable respuesta es de tipo discreta con una gran cantidad de ceros. La exploración de los datos y el análisis del modelo se realizará mediante el uso de los *software* estadísticos *R project* y Stata.

Definición de las variables

Mes: variable categórica que indica el número de mes en el cual se realizaron cierta cantidad de consultas médicas.

Región: variable cualitativa nominal que representa las 15 regiones de Chile.

Código establecimiento: variable cualitativa nominal que indica el código del establecimiento en donde realizaron consultas médicas.

Nombre del establecimiento: variable cualitativa nominal que especifica el establecimiento, de determinada región y comuna, donde realizaron consultas médicas.

Nombre comuna: variable cualitativa nominal que especifica las comunas de cada región de Chile.

Glosa prestación: variable cualitativa nominal que determina a qué tipo de especialidad médica pertenece un determinado número de consultas médicas.

Total consultas: variable numérica sobre el total de consultas médicas durante determinado mes.

Consultas por hombres: variable numérica sobre el total de consultas médicas, realizadas por hombres, durante determinado mes.

Consultas por mujeres: variable numérica sobre el total de consultas médicas, realizadas por mujeres, durante determinado mes.

Consultas menores de 10 años: variable numérica que indica la cantidad de consultas médicas en menores de 10 años.

Consultas entre 11 y 14 años: variable numérica que indica la cantidad de consultas médicas en personas entre 11 y 14 años.

Consultas entre 15 y 19 años: variable numérica que indica la cantidad de consultas médicas en personas entre 15 y 19 años.

Consultas entre 20 y 24 años: variable numérica que indica la cantidad de consultas médicas en personas entre 20 y 24 años.

Consultas entre 25 y 39 años: variable numérica que indica la cantidad de consultas médicas en personas entre 25 y 39 años.

Consultas entre 40 y 54 años: variable numérica que indica la cantidad de consultas médicas en personas entre 40 y 54 años.

Consultas entre 55 y 64 años: variable numérica que indica la cantidad de consultas médicas en personas entre 55 y 64 años.

Consultas mayores de 65 años: variable numérica que indica la cantidad de consultas médicas en personas mayores de 65 años.

Variables a utilizar

Las variables más importantes para el análisis, son las variables "Mes", "Región" y los recuentos de consultas médicas que corresponden a las variables de consultas para todas las edades. Como se explicó anteriormente, esta variable se encuentra segmentada en muchos rangos de edades, es por esto que se decidió agrupar las edades de acuerdo a cuatro grupos sociales, estos son: adolescente, joven, adulto y adulto mayor (SERPLAC, 2008). El detalle sobre los rangos de edades para estos cuatro grupos etarios se puede ver en la Tabla 2. Esta agrupación permitirá conocer y comparar las realidades para cada uno de los grupos etarios en estudio y, de este modo, se trabajará con estos cuatro vectores de datos como variables de respuesta para ajustar el modelo Tweedie.

Administración del conjunto de datos

Antes de realizar el análisis descriptivo de los datos, se efectuó una limpieza y administración del conjunto de datos. Esto se realizó mediante la exploración de los datos, en el cual se pudo detectar la existencia de algunos datos mal tabulados que impedían

la ejecución del análisis estadístico. Una vez corregido esto, se procedió a organizar y agrupar los vectores de datos a utilizar, es decir, aquellos vectores correspondientes a los grupos etarios anteriormente establecidos. Mediante la utilización de funciones matemáticas y estructuras de control del *software*, se pudo sumar las columnas de datos correspondientes a todos los rangos de edades y, por consiguiente, agruparlos en las cuatro clases sociales.

Análisis exploratorio

A continuación se realizó un análisis descriptivo de los datos pertenecientes al DEIS. Para este análisis, se usaron los tamaños de muestra obtenidos mediante el procedimiento del MAE (Tabla 6), ya que al utilizar estos datos, se puede tener información totalmente representativa a la hora de hacer los análisis estadísticos.

En la Tabla 9 y Figura 1, se hizo un análisis de frecuencia para los conteos de consultas y atenciones médicas realizadas por cada uno de los grupos etarios. Es de especial interés observar que, para cada uno de los grupos, la cantidad de datos correspondientes a ceros es muy alta. Si bien el máximo de consultas médicas sobrepasa los 2 mil, la gran cantidad de ceros hace que el promedio de consultas médicas, por grupo etario, disminuya considerablemente. En la Tabla 10, se puede observar la cantidad de veces que no hubo realización de consultas (ceros exactos) para cada uno de los grupos. En especial, se puede destacar que los grupos sociales de jóvenes y adultos mayores tienen más inasistencias de consultas médicas a comparación de los adolescentes y adultos.

Grupo	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
Adolescente	0,0	2,0	13,0	185,2	72,0	6804,0
Joven	0,0	5,0	30,0	178,9	107,0	6874,0
Adulto	0,0	6,0	27,0	156,4	83,0	2887,0
Adulto mayor	0,0	7,0	36,0	141,6	112,0	3389,0

Tabla 9: Resumen estadístico de consultas médicas según grupos etarios del sector privado

Grupo	Suma de consultas médicas	Cantidad de veces sin consultas médicas
Adolescente	1.691.294	2629
Joven	3.045.059	3197
Adulto	1.359.048	2351
Adulto mayor	2.243.945	3253

Tabla 10: Número de veces que no hubo consultas médicas para cada grupo etario

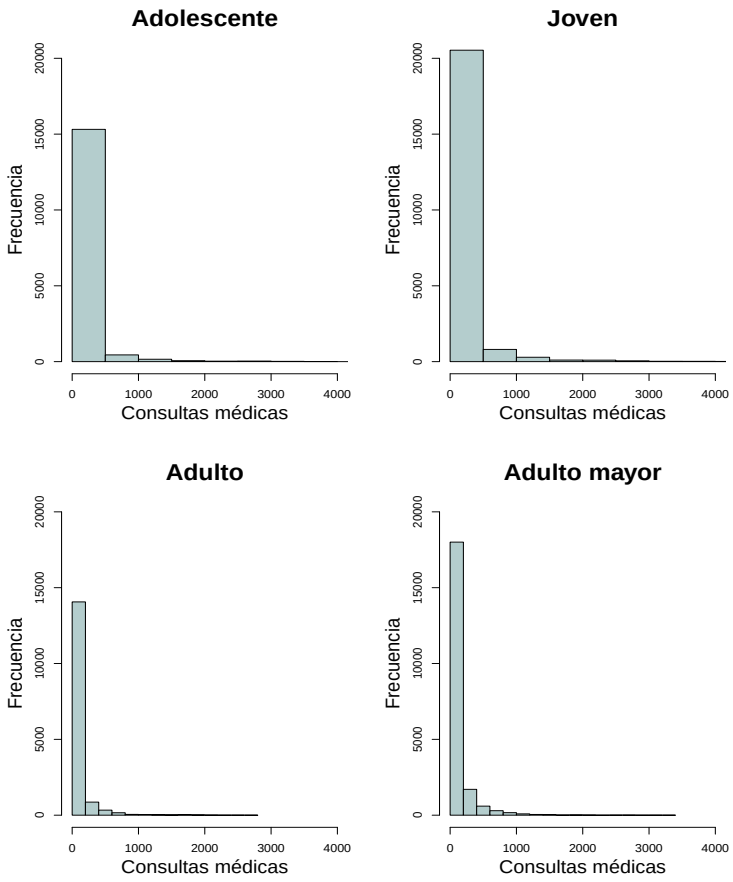

Figura 1: Distribución de frecuencia para los conteos de consultas médicas según grupo etario

Por otro lado, es interesante conocer el comportamiento que tienen las consultas médicas a medida que pasan los meses del año. A continuación, en la Figura 2 y Figura 3, se puede observar la distribución de ocurrencia y no ocurrencia de consultas médicas por cada grupo etario a medida que van pasando los meses del año, con el fin de poder encontrar algún tipo de tendencia o patrón y, sobre todo, poder diferenciar de manera gráfica el comportamiento de las realizaciones de consultas y atenciones médicas mensuales de cada grupo social.

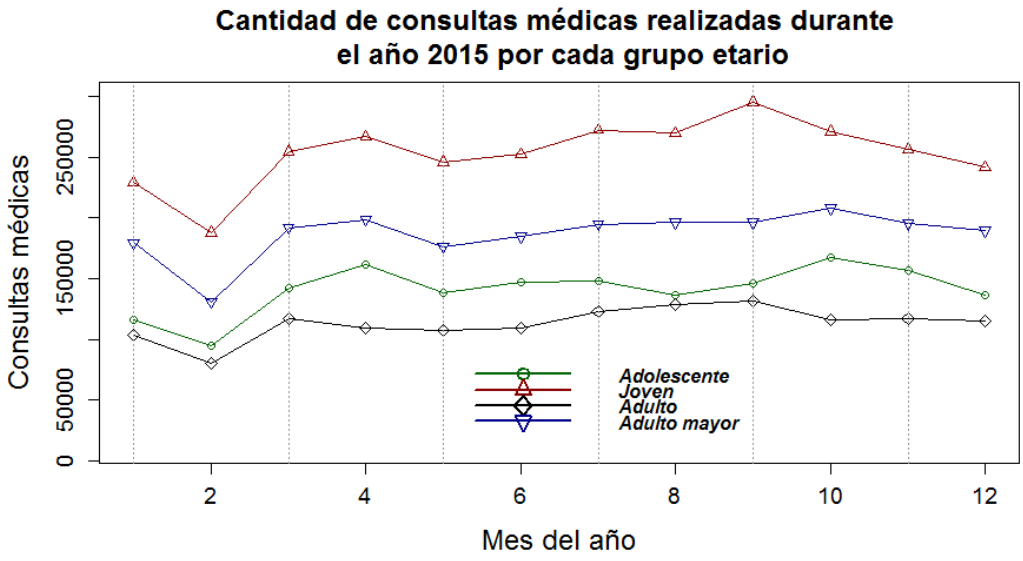

Figura 2: Distribución de frecuencia mensual sobre la realización de consultas médicas por cada grupo etario

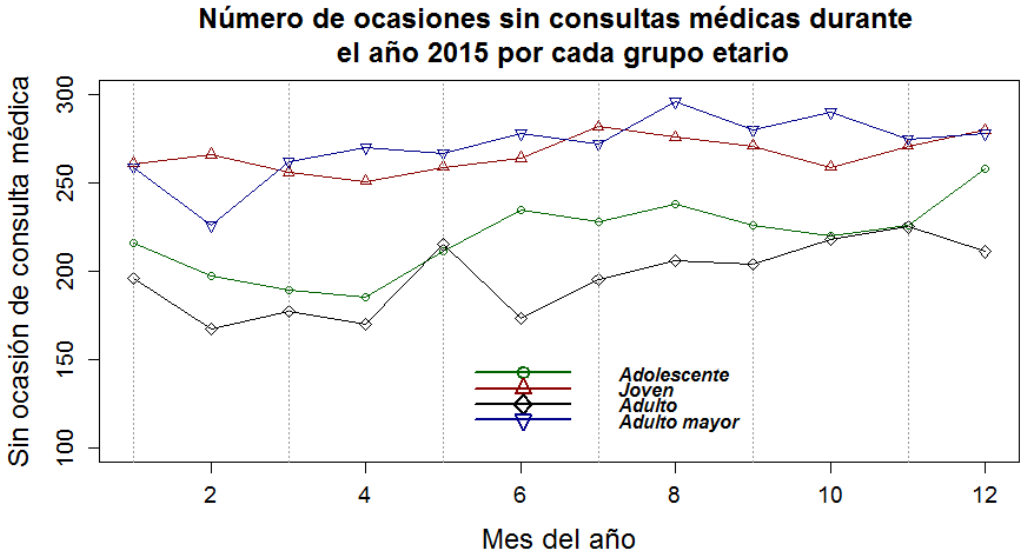

Figura 3: Distribución de frecuencia mensual sobre la no realización de consultas médicas por cada grupo etario

En la Figura 2, se puede ver que, independiente de la diferencia entre las cantidades de consultas realizadas entre los grupos, existe un comportamiento casi similar para la mayoría de los meses. En especial, los primeros y últimos meses del año, que presenta una gran baja en las realizaciones de consultas médicas. Muy por el contrario, en los meses correspondientes de marzo hasta septiembre, presentan un comportamiento casi

homogéneo. Por otro lado, con respecto a las no realizaciones de consultas médicas (Figura 3), ocurre algo ligeramente distinto, ya que se puede observar un comportamiento positivo a medida que pasan los meses del año. Esto quiere decir que empiezan el año con una cantidad menor de ausencia de consultas médicas (ceros exactos) y terminan el año con una mayor cantidad de ausencias sobre las mismas, en especial los jóvenes, adolescentes y adultos mayores.

La existencia de estas grandes cantidades de ceros, hacen que la distribución tome una forma asimétrica con una gran cola de datos a la derecha. El hecho de contener una gran cantidad de ceros exactos es, de hecho, una característica muy común en datos de conteos y, al mismo tiempo, una característica muy deseable para los modelos Tweedie, ya que esta distribución permite modelar grandes cantidades de ceros exactos, para cierto rango de valores del parámetro de poder.

Estimación de los parámetros p y ϕ

El objetivo principal es probar que los datos pueden ajustarse a un modelo Tweedie. Para esto, primero que todo se debe realizar la estimación de los parámetros del modelo. Estos parámetros son el de dispersión ϕ (estrictamente positivo) y el parámetro de potencia p, el cual debe estar entre $1 < p < 2$, en donde la distribución Tweedie es una mezcla Poisson-Gamma. A continuación, en la Tabla 11, se puede ver los valores estimados para cada parámetro de la distribución Tweedie.

Grupo etario	Estimación del parámetro $\hat{p}$	Estimación del parámetro $\hat{\phi}$
Adolescente	1,802	7,632
Joven	1,785	7,381
Adulto	1,736	6,839
Adulto mayor	1,720	7,422

Tabla 11: Estimación de los parámetros p y ϕ para los grupos etarios

Los valores detallados en la anterior Tabla 11, fueron obtenidos mediante el método de EMV utilizando los cuatro vectores de grupos etarios, en donde se puede ver que los parámetros estimados de potencia se aproximan a 2, por consiguiente se puede decir que la distribución de los datos se acerca a una Gamma ($p = 2$). La utilización de la librería *tweedie* del *software R project*, implementada por el autor Dunn en 2008, contiene una serie de funciones muy útiles para tratar este tipo de datos. En específico, para estimar los parámetros del modelo Tweedie, se utiliza la función *tweedie.profile*, el cual arroja los valores estimados de los parámetros de potencia y de dispersión. En la Figura 4, se pueden ver los gráficos en donde el parámetro de potencia obtiene su valor máximo,

para cada uno de los grupos etarios considerados en el estudio.

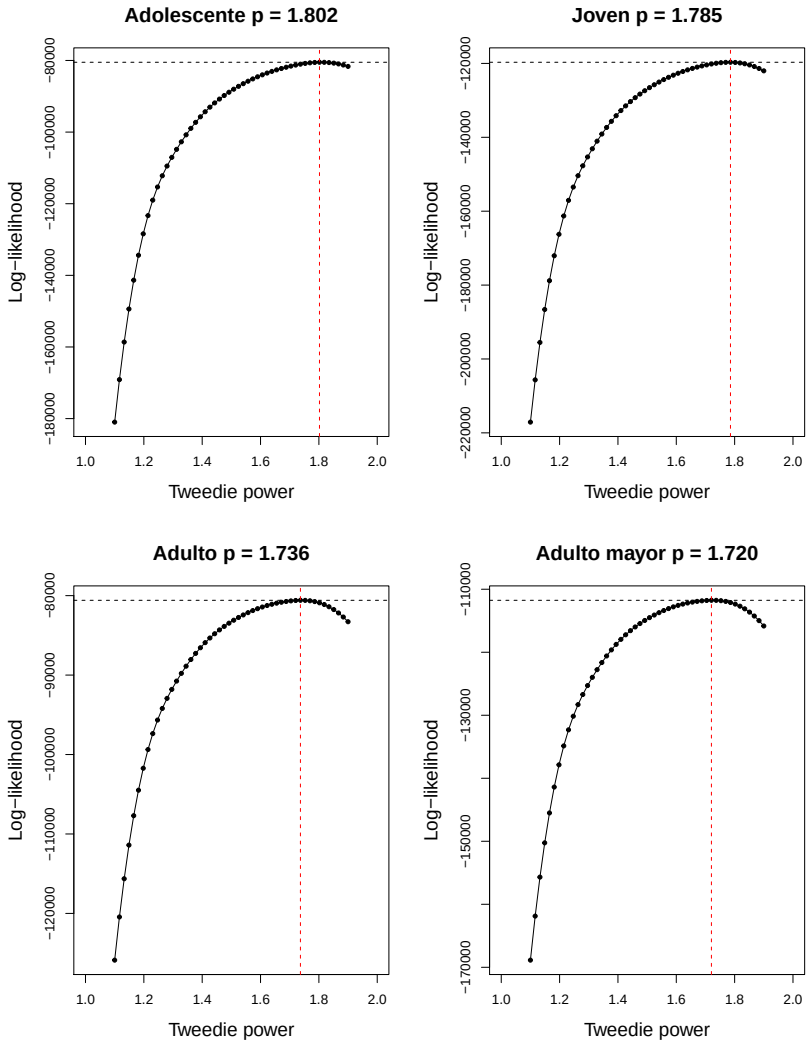

Figura 4: Estimación máxima verosímil para el parámetro de potencia de los cuatro grupos etarios

Una vez teniendo la estimación de los parámetros, es importante comprobar si los datos reales se ajustan a la distribución deseada. Para verificar que la distribución elegida es adecuada para los datos, se puede hacer uso del método de probabilidad Cuantil-Cuantil, ilustrado en la Figura 5, el cual hace una relación lineal entre los datos reales y los cuantiles de la distribución deseada. Para realizar este ajuste, primero se tuvieron que ordenar los datos de conteos, de menor a mayor cantidad, luego se creó un nuevo vector de datos con la misma cantidad de datos para cada grupo etario. Seguido de esto, se utilizaron los nuevos vectores de datos y se construyeron cuantiles Tweedie mediante la generación de datos pseudo aleatorios Tweedie con los parámetros estimados en la Tabla 11. Finalmente, se hizo la relación lineal entre los datos reales de cada

grupo etario y los datos simulados. Una vez hecho ésto, si los datos de conteo se ajusta bien a la distribución, entonces la gráfica debería acercarse a la linea recta de 45 grados.

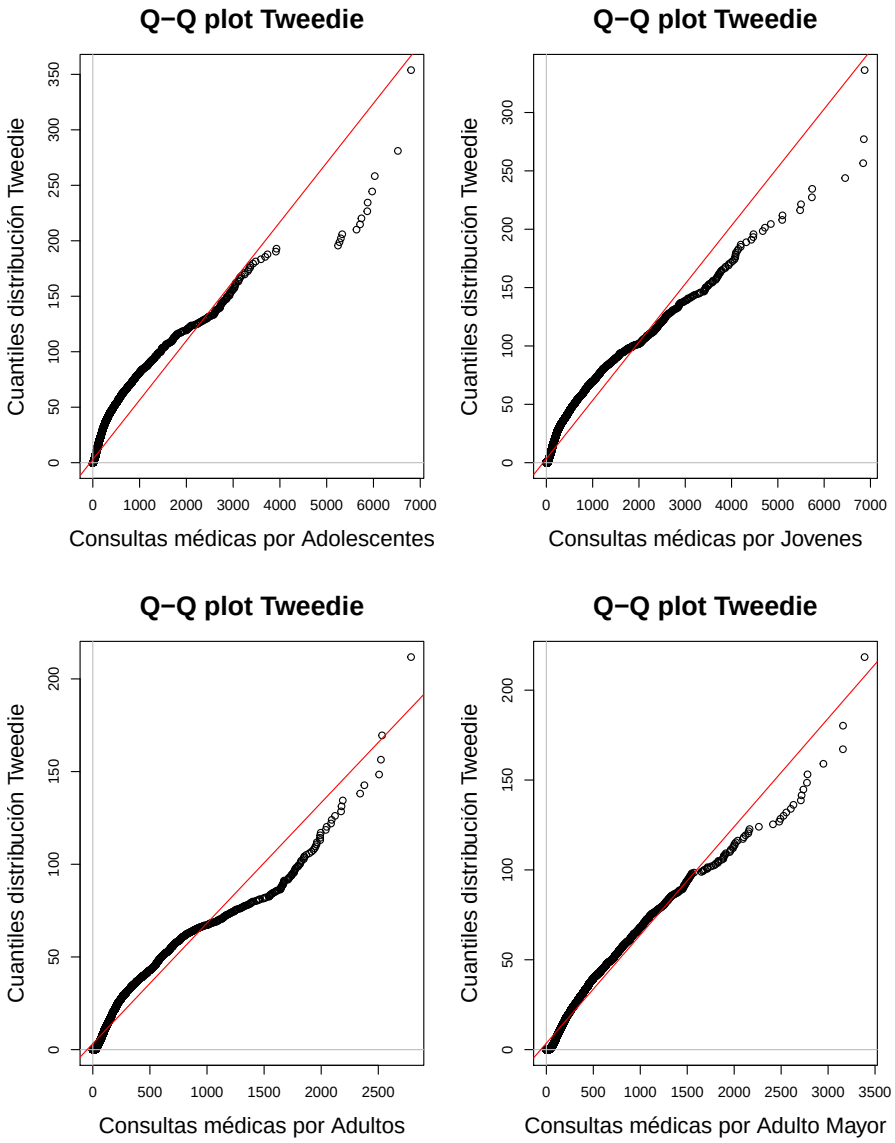

Figura 5: Probabilidad Cuantil-Cuantil para los grupos etarios

De esta forma se puede ver que de los grupos etarios, las consultas médicas realizadas por adultos mayores se ajustan mucho mejor a la distribución Tweedie que con respecto a los demás grupos etarios. Por otro lado, en las cuatro gráficas de probabilidad Cuantil-Cuantil, se observa la existencia de muchos datos que presentan sobredispersión, es decir, datos que tienen una mayor variabilidad de lo que se podría esperar.

Ajuste del modelo

Para ajustar un modelo lineal generalizado se tiene que tener en cuenta la "familia" y la "función de enlace" para la variable respuesta. La especificación de la familia indica cuál es la distribución de los errores del modelo y, la función vínculo, va a relacionar el valor esperado de la variable respuesta con el predictor lineal. En este caso, se ajustó un modelo lineal generalizado con familia tweedie y función de enlace logarítmica.

Se han realizado cuatro diferentes ajustes de modelo, es decir, uno para cada grupo etario, utilizando una variable explicativa, la variable categórica del "mes". Al momento de aplicar el modelo, se utilizaron los valores estimados del parámetro de potencia $\hat{p}$ para cada grupo etario, el cual atribuye la distribución apropiada a cada vector de datos. El modelo matemático es el siguiente:

$$Consultas\, medicas_{ij} \sim \beta_{0j} + \beta_{1j} Mes_{ij}$$

En la Tabla 12, se encuentran los resultados obtenidos mediante la regresión lineal generalizada, en donde se tienen los coeficientes estimados y el nivel de significancia de cada predictor, para cada grupo etario. También, en la Tabla 13, se puede ver el análisis descriptivo de la devianza residual para cada modelo.

	Grupos etarios							
Coeficientes	Adolescente		Joven		Adulto		Adulto mayor	
	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p
Intercepto (β_0)	4,445	$< 0,000$	4,842	$< 0,000$	4,373	$< 0,000$	4,659	$< 0,000$
Mes 2 (β_1)	-0,115	0,358	-0,146	0,113	-0,173	0,057	-0,288	$< 0,000$
Mes 3 (β_2)	0,235	0,052	0,107	0,236	0,154	0,080	0,077	0,026
Mes 4 (β_3)	0,363	0,003	0,127	0,156	0,104	0,242	0,060	0,358
Mes 5 (β_4)	0,185	0,125	0,056	0,535	0,011	0,898	-0,392	0,006
Mes 6 (β_5)	0,242	0,044	0,084	0,348	0,084	0,343	0,098	0,019
Mes 7 (β_6)	0,242	0,044	0,168	0,060	0,164	0,059	0,043	0,517
Mes 8 (β_7)	0,176	0,147	0,141	0,115	0,197	0,023	0,094	0,008
Mes 9 (β_8)	0,256	0,034	0,210	0,018	0,228	0,088	0,037	0,177
Mes 10 (β_9)	0,358	0,003	0,147	0,100	0,100	0,249	0,089	0,171
Mes 11 (β_{10})	0,299	0,012	0,080	0,371	0,083	0,338	0,058	0,066
Mes 12 (β_{11})	0,154	0,200	0,041	0,651	0,069	0,429	0,021	0,478

Tabla 12: Estimación de efectos en el modelo Tweedie que explica el conteo de consultas médicas en el sector privado de Chile en el año 2015

A *priori* se podría decir que, los grupos de adolescente y adulto mayor podrían ajustarse mucho mejor al modelo Tweedie, ya que gracias a la información que otorga los valores-p, la gran mayoría de los meses (factores) presentan una adición significativa al modelo, es decir, al tener un efecto significativo al modelo quiere decir que los cambios en los valores del mes tienen una fuerte relación a las realizaciones de consultas médicas. Sin embargo, antes de interpretar, hay que comprobar la calidad del modelo a través del diagnóstico de los residuos.

Grupo etario	Mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Máximo
Modelo Adolescente	-5,116	-3,360	-2,279	-1,750	-0,553	19,408
Modelo Joven	-5,249	-3,170	-1,943	-1,579	-0,424	18,922
Modelo Adulto	-5,050	-2,969	-1,649	-1,346	-0,089	14,185
Modelo Adulto mayor	-5,195	-3,187	-1,673	-1,351	0,113	15,754

Tabla 13: Resumen estadístico de la devianza residual de los modelos

A través del resumen estadístico de la devianza residual, presentado en la Tabla 13, se puede ver que los residuos tienen una fuerte asimetría a la derecha, es decir, la distribución de los datos residuales se separa fuertemente de la media hacia la cola derecha, habiendo una mayor concentración de datos residuales por debajo del cero. Esta fuerte asimetría se puede visualizar de mejor manera en el diagnóstico del modelo ilustrado en la Figura 6 y Figura 7.

Diagnóstico

El estudio y verificación de la idoneidad del modelo es muy importante antes de realizar la interpretación de los resultados, ya que de este modo se puede verificar, mediante el análisis de los residuos, si el modelo es adecuado para los datos utilizados. Los residuos más utilizados para el diagnóstico del modelo son los residuos de Pearson y los residuos de devianza.

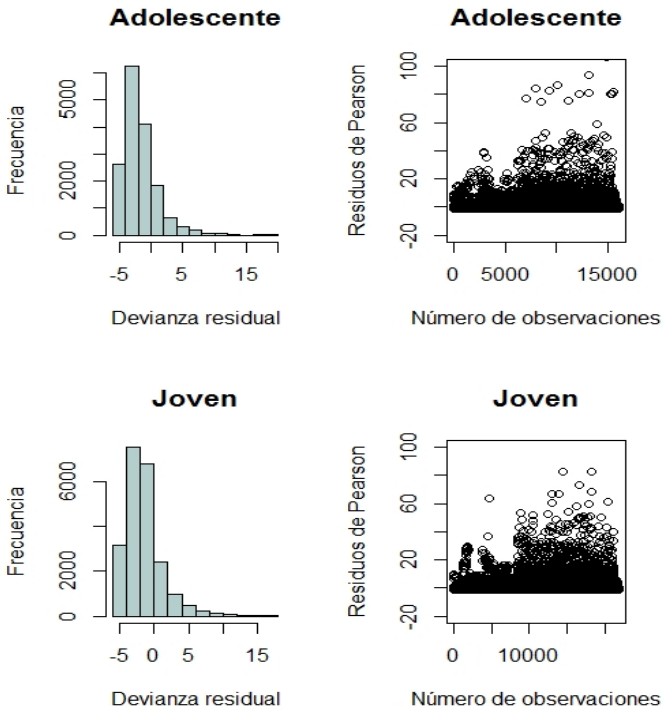

Figura 6: Análisis residual para los modelos Adolescente y Joven

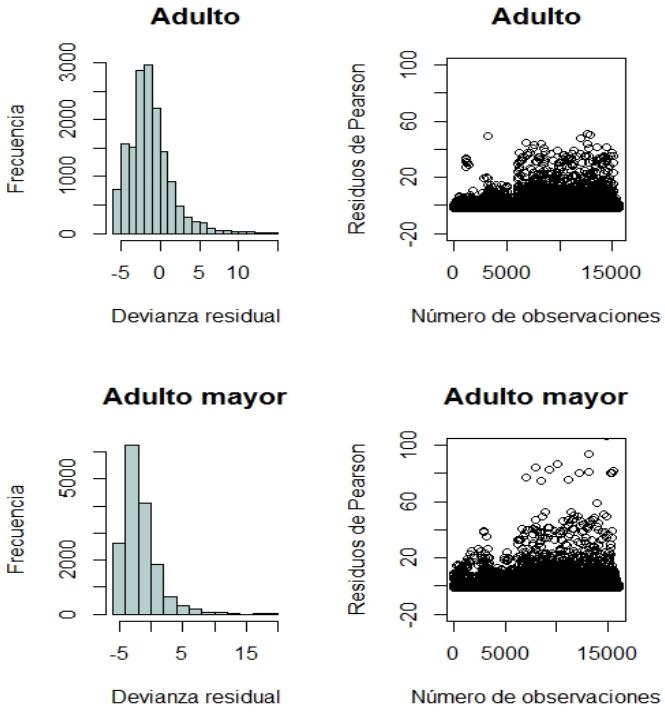

Figura 7: Análisis residual para los modelos Adulto y Adulto mayor

En particular, al realizar el análisis residual de Pearson, es deseable tener una distribución simétrica de residuos alrededor del cero. Sin embargo, para ninguno de los cuatro modelos se cumple, ya que la magnitud de los datos positivos es muy distinta a la magnitud de los datos negativos, lo cual indica que probablemente los modelos no pueden predecir con mucha precisión los datos atípicos. Esto se puede ver en los gráficos ilustrados en las Figuras 6 y 7, que muestran la elevada dispersión de los residuos respecto al cero. Para tratar de arreglar esto, se va a realizar un análisis de datos influyentes con el fin de poder mejorar el modelo propuesto. Según el autor Cook (1982), las medidas de influencia son aquellas medidas estadísticas que permiten detectar e identificar las observaciones influyentes. Estas observaciones son individual o colectivamente influyentes en el ajuste del modelo lineal, tanto en las estimaciones de los parámetros como en los residuos del modelo de regresión.

Datos atípicos

Para tener un mejor ajuste se van a tener que eliminar las observaciones influyentes, es por esto que se va a trabajar con los datos correspondientes a la Región Metropolitana de Santiago, ya que esta región presenta la gran mayoría de datos influyentes. A continuación, en la Figura 8 y Figura 9 se muestra la relación lineal de las consultas médicas correspondientes a la Región Metropolitana de Santiago y los cuantiles de la distribución Tweedie, para los conjuntos de datos con y sin datos atípicos.

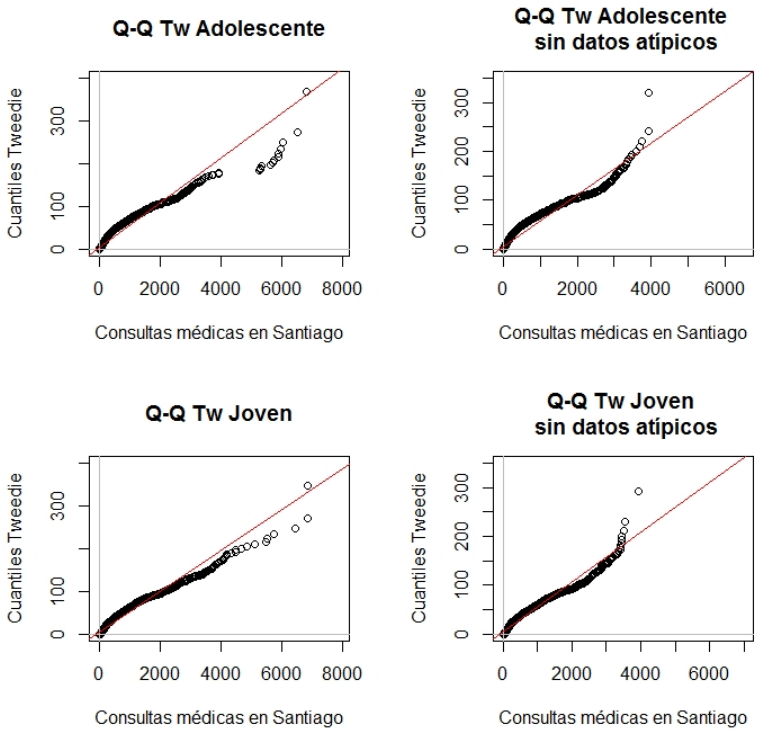

Figura 8: Ajuste de las consultas médicas Adolescente y Joven en Santiago

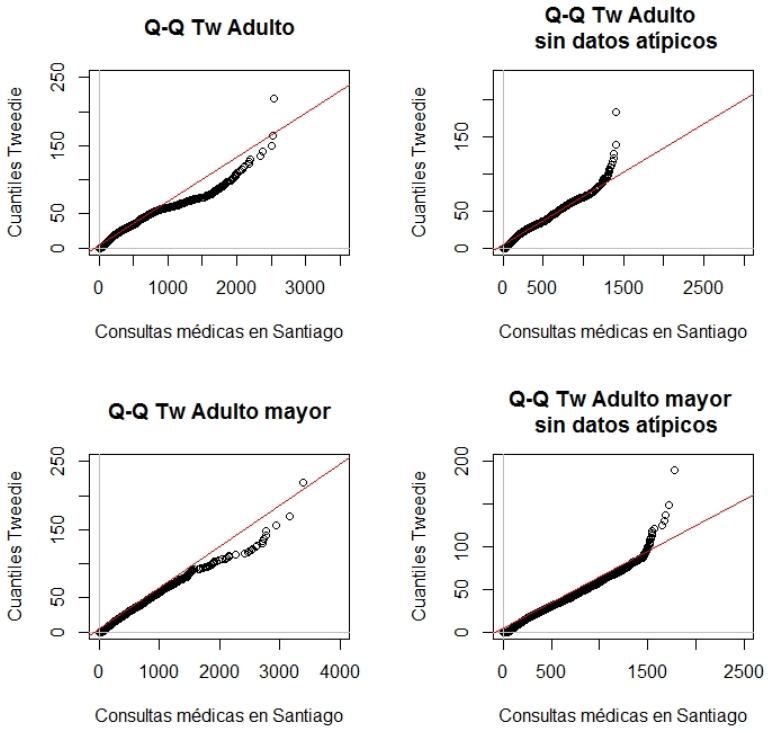

Figura 9: Ajuste de las consultas médicas Adulto y Adulto mayor en Santiago

Los gráficos de la izquierda corresponden a las observaciones de consultas médicas realizadas en Santiago por cada grupo etario y los gráficos de la derecha representan las mismas observaciones sin datos atípicos. Como se puede ver, al eliminar las observaciones atípicas hay un mejor ajuste a la distribución Tweedie en cada uno de los cuatro grupos etarios. El criterio que se utilizó para encontrar y eliminar los datos atípicos fue el método de las distancias de Cook. La función que se utilizó para hallar los datos atípicos es *cook.distance* del paquete "*car*" aplicado en el *software R*. Esta función identifica cuáles son los datos más influyentes en el modelo de regresión y, a partir de esta información, se eliminan del modelo.

El propósito de eliminar los datos atípicos del modelo es para comparar el ajuste del modelo con la presencia y ausencia de estos datos atípicos y comprobar si la ausencia de estos datos mejoran el ajuste del modelo propuesto. Una vez eliminados los datos atípicos, se procede a realizar el ajuste del modelo de regresión con y sin los datos atípicos con el propósito de comparar sus resultados. A continuación, en la Tabla 14 y Tabla 15, se encuentra el detalle sobre la regresión lineal generalizada con los datos completos y otro sin datos atípicos, respectivamente.

Grupos etarios en la Región Metropolitana de Santiago								
Coeficientes	Adolescente		Joven		Adulto		Adulto mayor	
	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p
Intercepto (β_0)	4,714	< 0000	5,089	< 0000	4,673	< 0000	4,941	< 0000
Mes 2 (β_1)	-0,194	0,205	0,039	0,722	-0,213	0,051	-0,338	< 0000
Mes 3 (β_2)	0,239	0,108	0,052	0,636	0,163	0,126	0,077	0,095
Mes 4 (β_3)	0,386	0,009	0,041	0,704	0,077	0,473	0,046	0,155
Mes 5 (β_4)	0,220	0,135	0,090	0,404	0,001	0,994	-0,056	0,081
Mes 6 (β_5)	0,295	0,044	0,065	0,547	0,041	0,697	-0,008	0,216
Mes 7 (β_6)	0,288	0,050	0,068	0,529	0,138	0,188	0,057	0,465
Mes 8 (β_7)	0,180	0,222	0,033	0,764	0,159	0,127	0,030	0,103
Mes 9 (β_8)	0,186	0,205	-0,113	0,297	0,149	0,156	-0,023	0,767
Mes 10 (β_9)	0,412	0,005	0,225	0,037	0,098	0,349	0,097	0,217
Mes 11 (β_{10})	0,307	0,036	0,098	0,366	0,083	0,430	0,052	0,507
Mes 12 (β_{11})	0,186	0,204	-0,016	0,880	0,054	0,604	0,000	0,099

Tabla 14: Estimación de efectos en el modelo Tweedie que explica el conteo de consultas médicas en el sector privado de la Región Metropolitana de Santiago (datos completos)

En la Tabla 14, se puede ver que los datos de consultas médicas perteneciente a los grupos sociales de adolescentes y adultos mayores siguen teniendo un muy buen ajuste al modelo Tweedie (igual que en el ajuste detallado en la Tabla 12). Por otra parte, en la Tabla 15, se puede ver que el ajuste de los datos al modelo mejora aún más cuando no hay presencia de datos atípicos, haciendo que los factores tengan un efecto aún más significativo al modelo.

Grupos etarios en la Región Metropolitana de Santiago sin datos atípicos								
Coeficientes	Adolescente		Joven		Adulto		Adulto mayor	
	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p	Estimación	Valor-p
Intercepto (β_0)	3,923	< 0000	4,335	< 0000	4,396	< 0000	4,236	< 0000
Mes 2 (β_1)	-0,127	0,160	0,158	0,003	-0,325	< 0000	-0,252	< 0000
Mes 3 (β_2)	0,244	0,005	0,018	0,738	0,176	0,025	0,139	0,014
Mes 4 (β_3)	0,315	0,000	0,018	0,726	0,087	0,272	0,130	0,020
Mes 5 (β_4)	0,232	0,007	0,059	0,265	-0,070	0,369	0,010	0,859
Mes 6 (β_5)	0,278	0,001	0,009	0,865	0,026	0,743	0,091	0,104
Mes 7 (β_6)	0,170	0,050	0,124	0,019	0,068	0,385	0,131	0,019
Mes 8 (β_7)	0,226	0,009	0,038	0,469	0,012	0,874	0,063	0,260
Mes 9 (β_8)	0,141	0,103	-0,381	< 0000	0,055	0,478	0,017	0,470
Mes 10 (β_9)	0,398	0,000	0,164	0,003	0,058	0,456	0,167	0,003
Mes 11 (β_{10})	0,270	0,002	0,106	0,047	0,013	0,864	0,122	0,029
Mes 12 (β_{11})	0,116	0,180	-0,093	0,082	0,002	0,985	0,066	0,236

Tabla 15: Estimación de efectos en el modelo Tweedie que explica el conteo de consultas médicas en el sector privado de la Región Metropolitana de Santiago (sin datos atípicos)

Además, en la Tabla 16, se puede ver el criterio de información de Akaike (AIC), el cual proporciona información cuantificada sobre la idoneidad de los modelos utilizados. A través de esto, se puede concluir que los modelos sin datos atípicos son más buenos y adecuados que los modelos con los datos completos, ya que al eliminar los datos influyentes se disminuyó considerablemente el AIC. El hecho de tener menor AIC es muy deseable en la elección de modelos, ya que a través de esto no se perderá tanta información a la hora de obtener los resultados (Cook, 1982).

AIC		
Grupo etario	Modelo Santiago con datos completos	Modelo Santiago sin datos atípicos
Adolescente	99.487	84.763
Joven	145.310	89.278
Adulto	99.254	91.948
Adulto mayor	138.209	98.208

Tabla 16: Criterio de información de Akaike para los modelos de cada grupo etario

Por último, se investigará más profundamente la idoneidad del modelo a través del diagnóstico. En la Figura 10 y Figura 11, se ilustra el análisis residual de los modelos sin presencia de datos atípicos. En esta ocasión, existe una buena estabilidad en la varianza y no hay tanta dispersión de residuos como en los modelos anteriores. Sin embargo, aún no hay una distribución homogénea de los residuos en torno al cero, esto puede deberse a la existencia de asimetría en los datos.

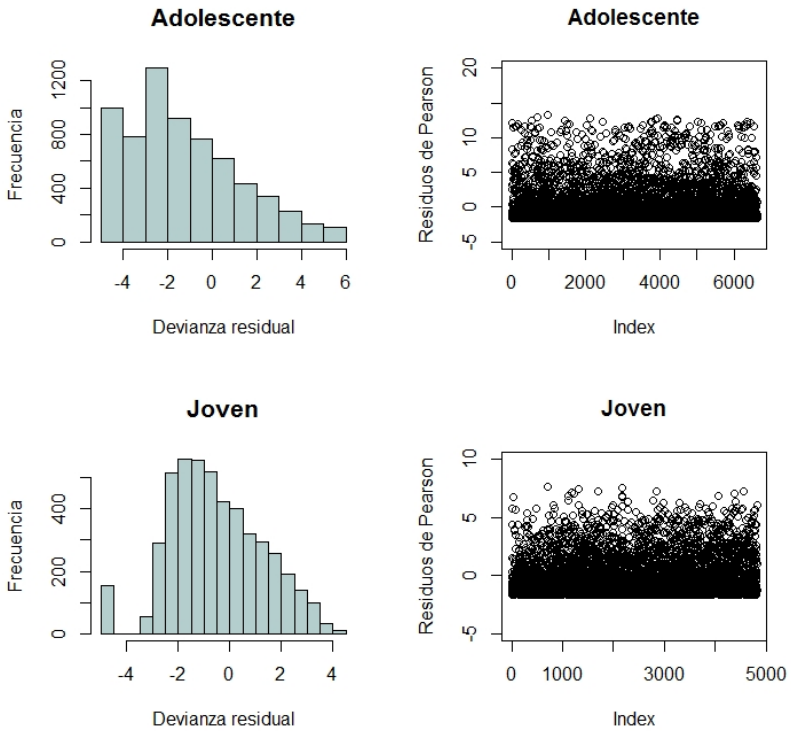

Figura 10: Residuos de los modelos Adolescente y Joven sin datos atípicos

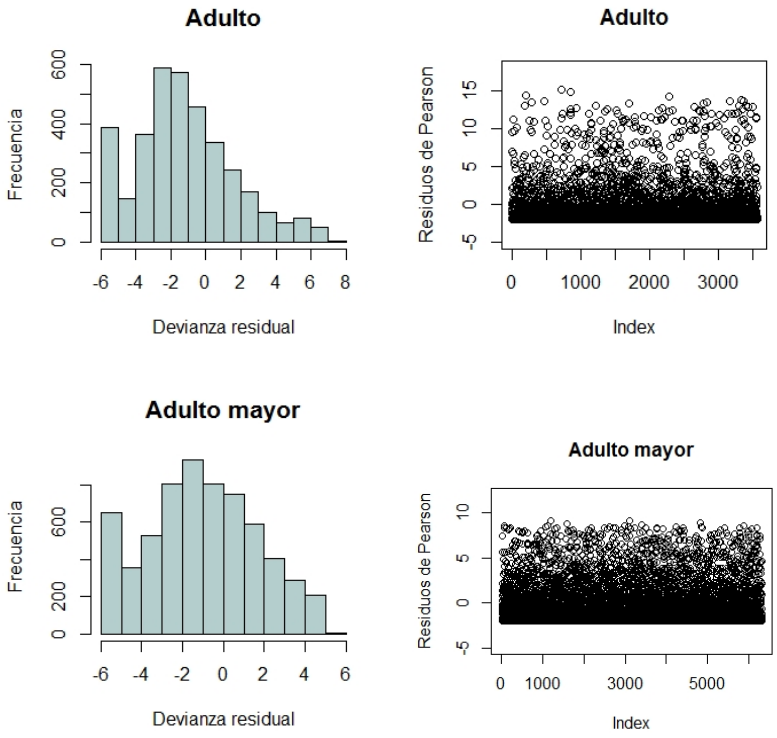

Figura 11: Residuos de los modelos Adulto y Adulto mayor sin datos atípicos

Conclusión

Observando los resultados obtenidos mediante la realización del ajuste de un MLG, con familia de dispersión exponencial Tweedie, a un conjunto de datos sobre conteos de consultas médicas, si bien no dieron los resultados que se esperaba tener en el diagnóstico del modelo, el modelo Tweedie sí logra ajustar de manera adecuada los datos relacionados con conteos de consultas médicas. Al realizar la eliminación de aquellos datos influyentes, se pudo lograr un buen ajuste de los datos a la distribución Tweedie, en especial los conteos de consultas médicas realizados por el Adulto mayor, ya que contenían una mayor cantidad de ceros exactos, que permiten ajustar mucho mejor la distribución Tweedie. Sin embargo, en el diagnóstico del modelo no se puede ver una distribución homogénea de los residuos al rededor del cero, debido precisamente a la existencia de una fuerte asimetría en los datos. No obstante, si se puede ver una buena estabilidad en la varianza.

En relación al ajuste del modelo Tweedie y el análisis de los residuos, se propone seguir estudiando el modelo Tweedie. Sin embargo, utilizando otros métodos de diagnóstico y, también, otro tipo de errores como los Anscombe, los cuales son usados para modelos lineales generalizados bajo distribuciones más tradicionales.

Referencias

Bonat, W. y Kokonendji, C. (2017). Flexible Tweedie regression models for continuous data. *Journal of Statistical Computation ans Simulation*. 87(11), 2138-2152.

Cayuela, L. (2010). *Modelos Lineales Generalizados (GLM)*. Univeridad de Granada.

Cook, R. D., and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Dobson, A. (2002). *An Introduction to Generalized Linear Models*. (Second edition). New York: CHAPMAN & HALL/CRC

Dunn, P. y Smyth, G. (2001). *Tweedie Family Densities: Methods of Evaluation*. Recuperado de https://pdfs.semanticscholar.org/aa7a/5b57d880b112f17987b865d7485eb8c641e1.pdf

Dunn, P. y Smyth, G. (2005). Series evaluation of Tweedie exponential dispersión models densities. *Statistics and computing 15*, 267-280.

Dunn, P. y Smyth, G. (2007). *Evaluation of Tweedie exponential dispersion model densities by Fourier inversion*. Spinger Science.

Fuller, A. W. (2009). *Sampling Statistics*. Canada: A John Wiley & Sons, Inc., Puclication.

Glen, S. (2018). *Tweedie Distribution: Definition and Examples*. Statistics How To. Recuperado de http://www.statisticshowto.com/tweedie-distribution/

Jorgensen, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistics Society*. 49(2), 127-162. Recuperado de http://www.jstor.org/stable/2345415?seq=1#page_scan_tab_contents

Jorgensen, B. (1997). *The Theory of the Dispersion Models*. London, New York: Chapman & Hall.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer Science.

Kaas, R. (2005). Compound Poisson distribution and GLMs Tweedie distribution. *MATHEMATICS DAY*. vol 3.

Lee, Y., Nelder, J. y Pawitan, Y. (2006). Generalized Linear Models with Random Effect. London, New York: Chapman & Hall/CRC.

Martínez, A. y Morales, J. (2001). *Modelos Lineales Generalizados*.

Meyers, G. (2009). *Predictive Modeling with the Tweedie Distribution*. Recuperado de https://www.casact.org/education/annual/2009/handouts/c25-meyers.pdf

Ministerio de desarrollo social. (2012). *Metodología del diseño muestral y factores de expansión CASEN 2011*. Recuperado de http://observatorio.ministeriodesarrollosocial.gob.cl/layout/doc/casen/Informe %20Diseno % 20Muestral_ Revision_ 13sep12.pdf

Ministerio de desarrollo social. (2015). *Metodología del diseño muestral y factores de expansión CASEN 2015*. Recuperado de http://observatorio.ministeriodesarrollosocial.gob.cl/casen multidimensional/casen/docs/Metodologia_ de_ Diseno_ Muestral_ Casen_ 2015.pdf

Ministerio de salud. (2015). *Departamento de estadísticas e información de salud*. Santiago de Chile: Gobierno de Chile. Recuperado de http://www.deis.cl

Mood, A. y Graybill, F. (1978). *Introducción a la Teoría de la Estadística*. (Cuarta edición). Madrid: Aguilar S A de ediciones.

Moshitch, D. y Nelken, I. (2014). Using Tweedie distribution for fitting spike count data. *Journal of Neuroscience Methods*. 225, 13-28. Recuperade de https://www.sciencedirect.com/science/article/pii/S0165027014000156

Montgomery, D. C., Peck, E. A. y Vining, G. G. (2002). *Introducción al Análisis de Regresión Lineal*. México: Compañía española de comunicaciones S A.

McCullagh, P. y Nelder, J. (1989). *Generalized Linear Models*. (Second edition). London, New York: Chapman and Hall.

Nelder, J. y Wedderburn, R. (1972). *Generalized Linear Models*. Journal of the Royal Statistical Society. vol 135 No 3. pag 370-385.

Pérez, C. (2010). *Técnicas de muestreo estadístico*. Madrid: Ibergaceta Publicaciones S.L.

Peña Sánchez, M. I. (2017). *Tarificación de Microseguros: Una Aplicación del Modelo Tweedie*.

Rickert, J. (2014). *A Note on Tweedie*. Revolutions. Recuperado de http://blog.revolutionanalytics.com /2014/10/a-note-on-tweedie.html

Tweedie, M.C.K. (1984). "An index which distinguishes between some important exponential families". In Ghosh, J.K.; Roy, J. *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*. Calcutta: Indian Statistical Institute. pp. 579?604.

A. J. Dobson. *An introduction to generalized linear models*. Chapman and Hall, London, 2nd edition, 2002.

SERPLAC. (2008). Grupos Sociales Específicos. *Un techo para Chile*. Recuperado de http://www.gobiernobiobio.cl/Documentos/Genero/Cuadernillo %203.pdf

Sidi, A. (1988). A user friendly extrapolation method for oscillatory infinite integrals. *Math. Comp.* 51(183), 249-266.

Shono, H. (2008). Application of the Tweedie Distribution to Zero-Catch Data in CPUE Analysis. *Fisheries Research*. 93, 154-162. Recuperado de https://www.sciencedirect.com/science/article/pii/S0165783608000945

Smolárová, T. (2017). *Tweedie models for pricing and reserving* (Master thesis). Faculty of Mathematics and Physics, Prague.

Swan, T. (2006). *Generalized estimating equations when the response variable has a Tweedie distribution: An application for multi-site rainfall modelling* (Doctoral dissertation, University of Southern Queensland).

Rutina en *software R*

```
rm(list=ls()) #limpiar

# Paquetes para las funciones utilizadas
library(readxl)
library(dplyr)
library(tweedie)
library(faraway)
library(statmod)
library(car)

# Para cambiar el máximo de impresiones en el output
getOption("max.print")
options(max.print = 999999)

# ----------------------------------------------------------------------
# Gráfico del total de consultas médicas por grupo etario y mes
# Vector de mes
m=c(1,2,3,4,5,6,7,8,9,10,11,12)

# Vector grupo adolescente
grupo1=c(115980, 95026, 141880, 161283, 138868, 147007, 148577, 136409,
146514, 167144, 156565, 136041)

# Vector grupo joven
grupo2=c(229321, 187909, 254886, 266921, 246214, 252528, 272574, 269888,
295456, 271204, 256518, 241640)

# Vector grupo adulto
grupo3=c(103388, 80540, 117519, 109295, 107516, 109214, 122707, 128462,
131909, 116154, 117515, 114829)

# Vector grupo adulto mayor
grupo4=c(179774, 131089, 192024, 198613, 176525, 184960, 194995, 196484,
196350, 207915, 195666, 189550)
```

```
plot(m, grupo1 , ylim=c(10000, 300000), pch=1, type="overplotted",
     col="darkgreen", cex.lab=1.5, ylab="Consultas médicas",
     xlab="Meses del año", cex.axis=1.2, las=0,  cex.main=1.5,
     mgp=c(3,1,0), main="Cantidad de consultas médicas realizadas
     durante \n el año 2015 por cada grupo etario")

lines(m, grupo2, pch=2, type="overplotted", col="red4")
lines(m, grupo3, pch=5, type="overplotted", col="black")
lines(m, grupo4, pch=6, type="overplotted", col="blue4")

legend("bottom",legend=c("Adolescente","Joven", "Adulto",
       "Adulto mayor"), y.intersp=0.35, pch=c(1,2,5,6),
       col=c("darkgreen","red4", "black", "blue4"), cex=1,
       text.width = 0.2, inset=0.03, lwd=2, pt.cex=1.5,
       text.font=4, bty="n")

abline(v=c(1,3,5,7,9,11), lty=3, col="gray61")


# Gráfico para el total de veces sin ocurrencia de consultas médicas
# por grupo etario y mes

# Vector grupo adolescente
c1=c(216,197, 189, 185, 211, 235, 228, 238, 226, 220, 226, 258)

# Vector grupo joven
c2=c(261,266, 256, 251, 259, 264, 282, 276, 271, 259, 271, 280)

# Vector grupo adulto
c3=c(196, 167, 177, 170, 215, 173, 195, 206, 204, 218, 225, 211)

# Vector grupo adulto mayor
c4=c(259, 226, 262, 270, 267, 278, 272, 296, 280, 290, 275, 278)

plot(m, c1 , ylim=c(100, 300), pch=1, type="overplotted", col="darkgreen",
     cex.axis=1.2, las=0, ylab="Sin ocasión de consulta médica",
     xlab="Meses del año", cex.main=1.5, cex.lab=1.5, main="Número de
     ocasiones sin consultas médicas durante \n el año 2015 por cada
     grupo etario")

lines(m, c2, pch=2, type="overplotted", col="red4")
lines(m, c3, pch=5, type="overplotted", col="black")
lines(m, c4, pch=6, type="overplotted", col="blue4")
```

```
legend("bottom",legend=c("Adolescente","Joven", "Adulto", "Adulto mayor"),
       y.intersp=0.35, pch=c(1,2,5,6),col=c("darkgreen","red4", "black",
       "blue4"), cex=1, text.width = 0.2, inset=0.03, lwd=2, pt.cex=1.5,
       text.font=4, bty="n")

abline(v=c(1,3,5,7,9,11), lty=3, col="gray61")


# ---------------------------------------------------------------------
# Resumen descriptivo de cada grupo
summary(adolescente)
summary(joven)
summary(adulto)
summary(adulto_mayor)

# ---------------------------------------------------------------------
# Histogramas para los conteos de consultas médicas generales de cada grupo
par(mfrow=c(2,2))
hist(adolescente, ylim = c(0,20000), xlim = c(0,4000), main="Adolescente",
     cex.main=2, cex.lab=1.6, ylab="Frecuencia", xlab="Consultas
     médicas", col="lightcyan3", mgp=c(2.5,1,0))
hist(joven, ylim = c(0,20000), xlim = c(0,4000), main="Joven",
     cex.main=2,cex.lab=1.6, ylab="Frecuencia", xlab="Consultas
     médicas", col="lightcyan3", mgp=c(2.5,1,0))
hist(adulto, ylim = c(0,20000), xlim = c(0,4000), main="Adulto",
     cex.main=2, cex.lab=1.6, ylab="Frecuencia", xlab="Consultas
     médicas",col="lightcyan3", mgp=c(2.5,1,0))
hist(adulto_mayor, ylim = c(0,20000), xlim = c(0,4000), main="Adulto
     mayor", cex.main=2, cex.lab=1.6, ylab="Frecuencia", xlab="Consultas
     médicas", col="lightcyan3", mgp=c(2.5,1,0))


# ---------------------------------------------------------------------
# Gráfico Cuantil Cuantil "adolescente"
phi=pow$phi.max
pow$p.max
xj=sort(datos$adolescente)
length(datos$adolescente)
p <- seq(from = 0, to = 0.99999, length.out =  16084)
qj = qtweedie(p, xi=NULL, mu, phi, power=pow$p.max)
QQmod1=glm(xj ~ qj, family = tweedie(var.power = pow$p.max,
           link.power = 0), data=datos)



plot(xj,qj, pch=1, main="Q-Q plot Tweedie", col="black",
```

```
     xlab="Conteo de consultas médicas por Adolescentes",
     ylab="Cuantiles de la distribución Tweedie",
     cex.main=1.5
)
abline(h=0,v=0, col="gray")
abline(QQmod1, lty=7, col="red")


# Gráfico Cuantil Cuantil "joven"
pow2$p.max
phi1=pow2$phi.max
xj1=sort(datos2$joven)
length(datos2$joven)
p1 <- seq(from = 0, to = 0.99999, length.out =  21956)
qj1 = qtweedie(p1, xi=NULL, mu1, phi1, power=pow2$p.max)
QQmod2=glm(xj1 ~ qj1, family = tweedie(var.power = pow2$p.max,
           link.power = 0), data=datos2)
plot(xj1,qj1, pch=1, main="Q-Q plot Tweedie", col="black",
     xlab="Conteo de consultas médicas por Jovenes",
     ylab="Cuantiles de la distribución Tweedie",
     cex.main=1.5
)
abline(h=0,v=0, col="gray")
abline(QQmod2, lty=7, col="red")


# Gráfico Cuantil Cuantil "adulto"
phi3=pow3$phi.max
xj3=sort(datos3$adulto)
length(datos3$adulto)
p3 <- seq(from = 0, to = 0.99999, length.out =  15631)
qj3 = qtweedie(p3, xi=NULL, mu3, phi3, power=pow3$p.max)
QQmod3=glm(xj3 ~ qj3, family = tweedie(var.power = pow3$p.max,
           link.power = 0), data=datos3)
plot(xj3,qj3, pch=1, main="Q-Q plot Tweedie", col="black",
     xlab="Conteo de consultas médicas por Adultos",
     ylab="Cuantiles de la distribución Tweedie",
     cex.main=1.5
)
abline(h=0,v=0, col="gray")
abline(QQmod3, lty=7, col="red")




# Gráfico Cuantil Cuantil "adulto mayor"
```

```
phi4=pow4$phi.max
xj4=sort(datos4$adulto_mayor)
length(datos4$adulto_mayor)
p4 <- seq(from = 0, to = 0.99999, length.out =  20992)
qj4 = qtweedie(p4, xi=NULL, mu4, phi4, power=pow4$p.max)
QQmod4=glm(xj4 ~ qj4, family = tweedie(var.power = pow4$p.max,
           link.power = 0), data=datos4)
plot(xj4,qj4, pch=1, main="Q-Q plot Tweedie", col="black",
     xlab="Conteo de consultas médicas por Adulto Mayor",
     ylab="Cuantiles de la distribución Tweedie",
     cex.main=1.5
)
abline(h=0,v=0, col="gray")
abline(QQmod4, lty=7, col="red")


# ------------------------------------------------------------------
# Gráfico de EMV del parámetro p para los grupos etarios

par(mfrow=c(2,2))
# Estimacion del parámetro p para "adolescente"
plot(pow, xlim=c(1:2), pch=20, col="gray0", main="Adolescente p = 1.802",
     xlab="Tweedie power", ylab="Log-likelihood", cex.lab=1.4, cex.main=1.5)
lines(pow, type="l")
abline(pow, v=pow$p.max, lty=2, col="red")
abline(pow, h=pow$ht, lty=2, col="black")

# Estimacion del parámetro p para "joven"
plot(pow2, xlim=c(1:2) , pch=20, col="gray0", main="Joven p = 1.785",
     xlab="Tweedie power", ylab="Log-likelihood", cex.lab=1.4, cex.main=1.5)
lines(pow2, type="l")
abline(pow2, v=pow2$p.max, lty=2, col="red")
abline(pow2, h=pow2$ht, lty=2, col="black")


# Estimacion del parámetro p para "adulto"
plot(pow3, xlim=c(1:2) , pch=20, col="gray0", main="Adulto p = 1.736",
     xlab="Tweedie power", ylab="Log-likelihood", cex.lab=1.4, cex.main=1.5)
lines(pow3, type="l")
abline(pow3, v=pow3$p.max, lty=2, col="red")
abline(pow3, h=pow3$ht, lty=2, col="black")




# Estimacion del parámetro p para "adulto mayor"
```

```
plot(pow4, xlim=c(1:2) , pch=20, col="gray0", main="Adulto mayor p = 1.720",
     xlab="Tweedie power", ylab="Log-likelihood", cex.lab=1.4, cex.main=1.5)
lines(pow4, type="l")
abline(pow4, v=pow4$p.max, lty=2, col="red")
abline(pow4, h=pow4$ht, lty=2, col="black")




# ------------------------------------------------------------------
# Planilla de datos para el grupo Adolescente
datos<- read_excel("C:/Users/FrancoJaime/Desktop/Seminario de tesis/
                   MuestraDeis.xlsx", sheet="Adolescente")


# Exploración del conjunto de datos
attach(datos)
length(adolescente)
View(datos)
head(datos)
summary(adolescente)
table(adolescente, mes_a)


# Transformación de la variable mes a variable factor
mes_aa <- as.factor(datos$mes_a)


# Estimación del parámetro de poder y phi
pow <- tweedie.profile(adolescente~1, p.vec=seq(1.1, 1.9, by=0.1),
                       do.plot=TRUE, data=datos)


# Modelo de regresión
mlg=glm(adolescente~mes_aa, family=tweedie(var.power = pow$p.max,
        link.power=0), data=datos)
summary(mlg)


# ------------------------------------------------------------------
# Planilla de datos para el grupo Joven
datos2<- read_excel("C:/Users/FrancoJaime/Desktop/Seminario de tesis/
                    MuestraDeis.xlsx", sheet="Joven")


# Exploración del conjunto de datos
```

```
attach(datos2)
length(joven)
summary(joven)
View(datos2)
head(datos2)
table(joven, mes_j)

# Transformación de la variable mes a variable factor
mes_jj    <- as.factor(datos2$mes_j)


# Estimación del parámetro de poder y phi
pow2 <- tweedie.profile(joven~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=datos2)

# Modelo de regresión
mlg2=glm(joven~mes_jj, family=tweedie(var.power = pow2$p.max,
         link.power=0), data=datos2)
summary(mlg2)



# ---------------------------------------------------------------------
# Planilla de datos para el grupo Adulto
datos3<- read_excel("C:/Users/FrancoJaime/Desktop/Seminario de tesis/
                    MuestraDeis.xlsx", sheet="Adulto")


# Exploración del conjunto de datos
attach(datos3)
length(adulto)
summary(adulto)
table(adulto, mes_ad)
View(datos3)
head(datos3)


# Transformación de la variable mes a variable factor
mes_add    <- as.factor(datos3$mes_ad)


# Estimación del parámetro de poder y phi
pow3 <- tweedie.profile(adulto~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=datos3)
```

```
# Modelo de regresión
mlg3=glm(adulto~mes_add, family=tweedie(var.power = pow3$p.max,
         link.power=0), data=datos3)
summary(mlg3)



# ------------------------------------------------------------------
# Planilla de datos para el grupo Adulto mayor
datos4<- read_excel("C:/Users/FrancoJaime/Desktop/Seminario de tesis/
                    MuestraDeis.xlsx", sheet="Adulto mayor")


# Exploración del conjunto de datos
attach(datos4)
summary(adulto_mayor)
table(adulto_mayor, mes_adm)
View(datos4)
head(datos4)



# Transformación de la variable mes a variable factor
mes_admm    <- as.factor(datos4$mes_adm)

# Estimación del parámetro de poder y phi
pow4 <- tweedie.profile(adulto_mayor~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=datos4)

# Modelo de regresión
mlg4=glm(adulto_mayor~mes_admm, family=tweedie(var.power=pow4$p.max,
         link.power = 0),data=datos4)
summary(mlg4)


# ------------------------------------------------------------------
# Diagnóstico de los cuatro modelos de regresión


# Histograma residuos de Pearson
par(mfrow=c(2,2))
hist(residuals(mlg, type="pearson"))
hist(residuals(mlg2, type="pearson"))
hist(residuals(mlg3, type="pearson"))
hist(residuals(mlg4, type="pearson"))
```

```
# Histograma de residuos de devianza
par(mfrow=c(2,2))
dev.new()
hist(residuals(mlg, type="deviance"), xlab="Residuos devianza",
     ylab="Frecuencia", main="Adolescente", cex.main=2, cex.lab=1.5)
hist(residuals(mlg2, type="deviance"), xlab="Residuos devianza",
     ylab="Frecuencia", main="Joven", cex.main=2, cex.lab=1.5)
hist(residuals(mlg3, type="deviance"), xlab="Residuos devianza",
     ylab="Frecuencia", main="Adulto", cex.main=2, cex.lab=1.5)
hist(residuals(mlg4, type="deviance"), xlab="Residuos devianza",
     ylab="Frecuencia", main="Adulto mayor", cex.main=2, cex.lab=1.5)


# Gráficos para los residuos de Pearson
par(mfrow=c(2,2))
plot(residuals(mlg, type="pearson"), ylim=c(-20,100),main="Adolescente",
     xlab="Numero observacion", ylab="Residuos de Person")
plot(residuals(mlg2, type="pearson"), ylim=c(-20,100), main="Joven",
     xlab="Numero observacion", ylab="Residuos de Person")

plot(residuals(mlg3, type="pearson"), ylim=c(-20,60), main="Adulto",
     xlab="Numero observacion", ylab="Residuos de Person")
plot(residuals(mlg4, type="pearson"), ylim=c(-20,60),main="Adulto mayor",
     xlab="Numero observacion", ylab="Residuos de Person")



# Gráficos de residuos de Pearson vs valores ajustados
par(mfrow=c(2,2))
plot(mlg$fitted.values, residuals(mlg, type="pearson"), main="Adolescente",
     xlab="Valores ajustados", ylab="Residuos de Pearson")
plot(mlg2$fitted.values, residuals(mlg2, type="pearson"), main="Joven",
     xlab="Valores ajustados", ylab="Residuos de Pearson")
plot(mlg3$fitted.values, residuals(mlg3, type="pearson"), main="Adulto",
     xlab="Valores ajustados", ylab="Residuos de Pearson")
plot(mlg4$fitted.values, residuals(mlg4, type="pearson"),main="Adulto
     mayor", xlab="Valores ajustados", ylab="Residuos de Pearson")


# Gráficos de residuos de Pearson vs predictor lineal
par(mfrow=c(2,2))
plot(mlg$linear.predictors, residuals(mlg, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adolescente")
plot(mlg2$linear.predictors, residuals(mlg2, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Joven")
```

```
plot(mlg3$linear.predictors, residuals(mlg3, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adulto")
plot(mlg4$linear.predictors, residuals(mlg4, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adulto mayor")


# Agrupación de gráficos de los residuos para el grupo adolescente
par(mfrow=c(2,2))
hist(residuals(mlg, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adolescente", cex.lab=1.5, cex.main=2)
plot(residuals(mlg, type="pearson"), ylim=c(-20,100), xlab="número
     de observaciones", ylab="Residuos de Pearson", main="Adolescente",
     cex.lab=1.5, cex.main=2)
plot(mlg$fitted.values, residuals(mlg, type="pearson"), xlab="Valores
     ajustados", ylab="Residuos Pearson", main="Adolescente",
     cex.lab=1.5, cex.main=2)
plot(mlg$linear.predictors, residuals(mlg, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adolescente",
     cex.lab=1.5, cex.main=2)


# Agrupación de gráficos de los residuos para el grupo joven
par(mfrow=c(2,2))
hist(residuals(mlg2, type="deviance"), xlab="Devianza residual", ylab=
     "Frecuencia", main="Joven")
plot(residuals(mlg2, type="pearson"), ylim=c(-20,100), xlab="número de
     observaciones", ylab="Residuos de Pearson", main="Joven")
plot(mlg2$fitted.values, residuals(mlg2, type="pearson"), xlab="Valores
     ajustados", ylab="Residuos Pearson", main="Joven")
plot(mlg2$linear.predictors, residuals(mlg2, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Joven")


# Agrupación de gráficos de los residuos para el grupo adulto
par(mfrow=c(2,2))
hist(residuals(mlg3, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adulto")
plot(residuals(mlg3, type="pearson"), ylim=c(-20,100), xlab="número de
     observaciones", ylab="Residuos de Pearson", main="Adulto")
plot(mlg3$fitted.values, residuals(mlg3, type="pearson"), xlab="Valores
     ajustados", ylab="Residuos Pearson", main="Adulto")
plot(mlg3$linear.predictors, residuals(mlg3, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adulto")


# Agrupación de gráficos de los residuos para el grupo adulto mayor
```

```
par(mfrow=c(2,2))
hist(residuals(mlg4, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adulto mayor")
plot(residuals(mlg4, type="pearson"), ylim=c(-20,100), xlab="número
     de observaciones", ylab="Residuos de Pearson", main="Adulto mayor")
plot(mlg4$fitted.values, residuals(mlg4, type="pearson"), xlab="Valores
     ajustados", ylab="Residuos Pearson", main="Adulto mayor")
plot(mlg4$linear.predictors, residuals(mlg4, type="pearson"), xlab=
     "Predictor lineal", ylab="Residuos Pearson", main="Adulto mayor")



# ------------------------------------------------------------------
# Región Metropolitana

# Conjunto de datos "adolescente" de la región metropolitana
dat1 <- datos %>% filter(nreg1=="13")
attach(dat1)
length(dat1$adolescente)


# Factor mes "adolescente" región Metropolitana
factor=as.factor(dat1$mes_a)

# Estimación de parámetros "adolescente" región Metropolitana
pow5 <- tweedie.profile(dat1$adolescente~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=dat1)

# Modelo de regresión "adolescente" región Metropolitana
modelo=glm(dat1$adolescente~factor, family=tweedie(var.power=pow5$p.max,
           link.power = 0),data=dat1)
summary(modelo)


# Datos atípicos
cooksd=cooks.distance(modelo)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")


# Gráfico de distancia de cook
sample_size <- nrow(dat1)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4/sample_size, col="red")  # add cutoff line
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4/sample_size,
     names(cooksd),""), col="red")  # add labels
```

```
# Datos influyentes "adolescentes" región Metropolitana
influential <- as.numeric(names(cooksd)[(cooksd > (4/sample_size))])

# Conjunto de datos no influyentes
dat1_screen <- dat1[-influential, ]

# Estimación de parámetros sin valores influyentes
pow5_1 <- tweedie.profile(dat1_screen$adolescente~1, p.vec=seq(1.1, 1.9,
                          by=0.1), do.plot=TRUE, data=dat1_screen)

# Factor mes sin datos atípicos
factor_SA=as.factor(dat1_screen$mes_a)

# Modelo de regresión "adolescente" región Metropolitana  sin datos atípicos
modelo_SA=glm(dat1_screen$adolescente~factor_SA, family=tweedie(var.power=
              pow5_1$p.max, link.power = 0),data=dat1_screen)
summary(modelo_SA)



# Conjunto de datos "joven" de la región metropolitana
dat2 <- datos2 %>% filter(nreg2=="13")
attach(dat2)
length(joven)

# Factor mes "joven" región metropolitana
factor2=as.factor(dat2$mes_j)

# Estimación de parámetros "joven" región metropolitana
pow7 <- tweedie.profile(dat2$joven~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=dat2)

# Modelo de regresión "joven" región Metropolitana
modelo2=glm(dat2$joven~factor2, family=tweedie(var.power=pow7$p.max,
            link.power = 0),data=dat2)
summary(modelo2)

# Datos atípicos
cooksd2=cooks.distance(modelo2)
plot(cooksd2, pch="*", cex=2, main="Influential Obs by Cooks distance")


# Gráfico de distancia de cook
sample_size2 <- nrow(dat2)
plot(cooksd2, pch="*", cex=2, main="Influential Obs by Cooks distance")
```

```
abline(h = 4/sample_size2, col="red")  # add cutoff line
text(x=1:length(cooksd2)+1, y=cooksd2, labels=ifelse(cooksd2>4/sample_size2,
     names(cooksd2),""), col="red")  # add labels

# Datos influyentes "joven" región Metropolitana
influential2 <- as.numeric(names(cooksd2)[(cooksd2 > (4/sample_size2))])
length(influential2)


# Conjunto de datos no influyentes
dat2_screen <- dat2[-influential2, ]


# Estimación de parámetros sin valores influyentes
pow7_1 <- tweedie.profile(dat2_screen$joven~1, p.vec=seq(1.1, 1.9, by=0.1),
                          do.plot=TRUE, data=dat2_screen)


# Factor mes sin datos atípicos
factor2_SJ=as.factor(dat2_screen$mes_j)


# Modelo de regresión "joven" región Metropolitana sin datos atípicos
modelo2_SJ=glm(dat2_screen$joven~factor2_SJ, family=tweedie(var.power=
               pow7_1$p.max, link.power = 0),data=dat2_screen)
summary(modelo2_SJ)


# Conjunto de datos "adulto" región metropolitana
dat3 <- datos3 %>% filter(nreg3=="13")
attach(dat3)
length(adulto)

# Estimación de parámetros "adulto" región Metropolitana
pow9 <- tweedie.profile(dat3$adulto~1, p.vec=seq(1.1, 1.9, by=0.1),
                        do.plot=TRUE, data=dat3)

# Factor mes "adulto" región Metropolitana
factor3=as.factor(dat3$mes_ad)

# Modelo de regresión "adulto" región Metropolitana
modelo3=glm(dat3$adulto~factor3, family=tweedie(var.power=pow9$p.max,
            link.power = 0),data=dat3)
summary(modelo3)

# Datos atípicos
```

```
cooksd3=cooks.distance(modelo3)
plot(cooksd3, pch="*", cex=2, main="Influential Obs by Cooks distance")


# Gráfico de distancia de cook
sample_size3 <- nrow(dat3)
plot(cooksd3, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4/sample_size3, col="red")  # add cutoff line
text(x=1:length(cooksd3)+1, y=cooksd3, labels=ifelse(cooksd3>4/sample_size3,
     names(cooksd3),""), col="red")  # add labels

# Datos influyentes "adulto" región Metropolitana
influential3 <- as.numeric(names(cooksd3)[(cooksd3 > (4/sample_size3))])

# Conjunto de datos no influyentes
dat3_screen <- dat3[-influential3, ]

# Estimación de parámetros "adulto" sin valores influyentes
pow9_1 <- tweedie.profile(dat3_screen$adulto~1, p.vec=seq(1.1, 1.9, by=0.1),
                          do.plot=TRUE, data=dat3_screen)

# Factor mes "adulto" region Metropolitana sin datos atípicos
factor3_SAd=as.factor(dat3_screen$mes_ad)

# Modelo de regresión "adulto" región metropolitana sin datos atípicos
modelo3_SAd=glm(dat3_screen$adulto~factor3_SAd, family=tweedie(var.power=
                pow9_1$p.max, link.power = 0),data=dat3_screen)
summary(modelo3_SAd)



# Conjunto de datos "adulto mayor" región Metropolitana
dat4 <- datos4 %>% filter(nreg4=="13")
attach(dat4)
length(adulto_mayor)


# Estimación de parámetros "adulto mayor" región metropolitana
pow11 <- tweedie.profile(dat4$adulto_mayor~1, p.vec=seq(1.1, 1.9, by=0.1),
                         do.plot=TRUE, data=dat4)


# Factor mes "adulto mayor" región Metropolitana
factor4=as.factor(dat4$mes_adm)
```

```
#Modelo de regresión "adulto mayor" región Metropolitana
modelo4=glm(dat4$adulto_mayor~factor4, family=tweedie(var.power=
            pow11$p.max, link.power = 0),data=dat4)
summary(modelo4)

# Datos atípicos
cooksd4=cooks.distance(modelo4)
plot(cooksd4, pch="*", cex=2, main="Influential Obs by Cooks distance")

# Gráfico de distancia de cook
sample_size4 <- nrow(dat4)
plot(cooksd4, pch="*", cex=2, main="Influential Obs by Cooks distance")
abline(h = 4/sample_size4, col="red")  # add cutoff line
text(x=1:length(cooksd4)+1, y=cooksd4, labels=ifelse(cooksd4>4/sample_size4,
     names(cooksd4),""), col="red")  # add labels

# Datos influyentes "adulto mayor" región Metropolitana
influential4 <- as.numeric(names(cooksd4)[(cooksd4 > (4/sample_size4))])

# Conjunto de datos no influyentes
dat4_screen <- dat4[-influential4, ]


# Estimación de parámetros "adulto mayor" sin valores influyentes
pow11_1 <- tweedie.profile(dat4_screen$adulto_mayor~1, p.vec=
                           seq(1.1, 1.9, by=0.1), do.plot=TRUE,
                           data=dat4_screen)

# Factor mes "adulto mayor" sin datos atípicos
factor4_SAdm=as.factor(dat4_screen$mes_adm)

# Modelo de regresión "adulto mayor" Metropolitana sin datos atípicos
modelo4_SAdm=glm(dat4_screen$adulto_mayor~factor4_SAdm, family=tweedie
           (var.power=pow11_1$p.max, link.power = 0),data=dat4_screen)
summary(modelo4_SAdm)


# ------------------------------------------------------------------
# Criterio de información de Akaike para los modelos
# AIC datos adolescente
AICtweedie(mlg)         # modelo completo
AICtweedie(modelo)      # modelo santiago
AICtweedie(modelo_SA)   # modelo santiago sin atípicos


# AIC datos joven
```

```
AICtweedie(mlg2)        # modelo completo
AICtweedie(modelo2)     # modelo santiago
AICtweedie(modelo2_SJ)  # modelo santiago sin atípicos


# AIC datos adulto
AICtweedie(mlg3)        # modelo completo
AICtweedie(modelo3)     # modelo santiago
AICtweedie(modelo3_SAd) # modelo santiago sin atípicos


# AIC datos adulto mayor
AICtweedie(mlg4)         # modelo completo
AICtweedie(modelo4)      # modelo santiago
AICtweedie(modelo4_SAdm) # modelo santiago sin atípicos


# ------------------------------------------------------------------
# Gráficos Cuantil Cuantil con y sin datos atípicos


# Gráfico Cuantil Cuantil "adolescente" región Metropolitana
 fi1=pow5$phi.max
 yt1=sort(dat1$adolescente)
 length(dat1$adolescente)
 b1 <- seq(from = 0, to = 0.99999, length.out =  9360)
 v1 = qtweedie(b1, xi=NULL, miu1, fi1, power=pow5$p.max)
 QQ1=glm(yt1 ~ v1, family = tweedie(var.power = pow5$p.max,
         link.power = 0), data=dat1)
 plot(yt1,v1, pch=1, main="Q-Q Tw Adolescente", col="black",
      xlab="Consultas médicas en Santiago",
      ylab="Cuantiles Tweedie",
      cex.main=1.2, xlim=c(0,7900), ylim=c(0,400)
 )
 abline(h=0,v=0, col="gray")
 abline(QQ1, lty=7, col="red")


# Gráfico Cuantil Cuantil "adolescente" región Metropolitana
# sin datos atípicos
fi2=pow5_1$phi.max
yt2=sort(dat1_screen$adolescente)
length(dat1_screen$adolescente)
b2 <- seq(from = 0, to = 0.99999, length.out =  9347)
v2 = qtweedie(b2, xi=NULL, miu2, fi2, power=pow5_1$p.max)
QQ2=glm(yt2 ~ v2, family = tweedie(var.power = pow5_1$p.max,
```

```
        link.power = 0), data=dat1_screen)
plot(yt2,v2, pch=1, main="Q-Q Tw Adolescente \n sin datos atípicos",
     col="black", xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie", cex.main=1.2, ylim=c(0,350),
     xlim=c(0,6500)
)
abline(h=0,v=0, col="gray")
abline(QQ2, lty=7, col="red")



# Gráfico Cuantil Cuantil "joven" región Metropolitana
fi3=pow7$phi.max
yt3=sort(dat2$joven)
length(dat2$joven)
b3 <- seq(from = 0, to = 0.99999, length.out =  12778)
v3 = qtweedie(b3, xi=NULL, miu3, fi3, power=pow7$p.max)
QQ3=glm(yt3 ~ v3, family = tweedie(var.power = pow7$p.max,
        link.power = 0), data=dat2)
plot(yt3,v3, pch=1, main="Q-Q Tw Joven", col="black",
     xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie",
     cex.main=1.2, xlim=c(0,7900), ylim=c(0,400)
)
abline(h=0,v=0, col="gray")
abline(QQ3, lty=7, col="red")


# Gráfico Cuantil Cuantil "joven" región Metropolitana
# sin datos atípicos
fi4=pow7_1$phi.max
yt4=sort(dat2_screen$joven)
length(dat2_screen$joven)
b4 <- seq(from = 0, to = 0.99999, length.out =  12739)
v4 = qtweedie(b4, xi=NULL, miu4, fi4, power=pow7_1$p.max)
QQ4=glm(yt4 ~ v4, family = tweedie(var.power = pow7_1$p.max,
        link.power = 0), data=dat2_screen)
plot(yt4,v4, pch=1, main="Q-Q Tw Joven \n sin datos atípicos",
     col="black", xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie", cex.main=1.2,
     ylim=c(0,350), xlim=c(0,7000)
)
abline(h=0,v=0, col="gray")
abline(QQ4, lty=7, col="red")
```

```
# Gráfico Cuantil Cuantil "adulto" región Metropolitana
fi5=pow9$phi.max
yt5=sort(dat3$adulto)
length(dat3$adulto)
b5 <- seq(from = 0, to = 0.99999, length.out = 9100)
v5 = qtweedie(b5, xi=NULL, miu5, fi5, power=pow9$p.max)
QQ5=glm(yt5 ~ v5, family = tweedie(var.power = pow9$p.max,
        link.power = 0), data=dat3)
plot(yt5,v5, pch=1, main="Q-Q Tw Adulto", col="black",
     xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie",
     cex.main=1.2, xlim = c(0,3500), ylim=c(0,250)
)
abline(h=0,v=0, col="gray")
abline(QQ5, lty=7, col="red")


# Gráfico Cuantil Cuantil "adulto" región Metropolitana
# sin datos atípicos
fi6=pow9_1$phi.max
yt6=sort(dat3_screen$adulto)
length(dat3_screen$adulto)
b6 <- seq(from = 0, to = 0.99999, length.out =  9016)
v6 = qtweedie(b6, xi=NULL, miu6, fi6, power=pow9_1$p.max)
QQ6=glm(yt6 ~ v6, family = tweedie(var.power = pow9_1$p.max,
        link.power = 0), data=dat3_screen)
plot(yt6,v6, pch=1, main="Q-Q Tw Adulto \n sin datos atípicos",
     col="black", xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie", cex.main=1.2,
     ylim=c(0,230), xlim=c(0,3000)
)
abline(h=0,v=0, col="gray")
abline(QQ6, lty=7, col="red")



# Gráfico Cuantil Cuantil "adulto mayor" región Metropolitana
fi7=pow11$phi.max
yt7=sort(dat4$adulto_mayor)
length(dat4$adulto_mayor)
b7 <- seq(from = 0, to = 0.99999, length.out = 12218)
v7 = qtweedie(b7, xi=NULL, miu7, fi7, power=pow11$p.max)
QQ7=glm(yt7 ~ v7, family = tweedie(var.power = pow11$p.max,
        link.power = 0), data=dat4)
plot(yt7,v7, pch=1, main="Q-Q Tw Adulto mayor", col="black",
     xlab="Consultas médicas en Santiago",
```

```
     ylab="Cuantiles Tweedie",
     cex.main=1.2, xlim=c(0,4000), ylim=c(0,250)
)
abline(h=0,v=0, col="gray")
abline(QQ7, lty=7, col="red")


# Gráfico Cuantil Cuantil "adulto mayor" región Metropolitana
# sin datos atípicos
fi8=pow11_1$phi.max
yt8=sort(dat4_screen$adulto_mayor)
length(dat4_screen$adulto_mayor)
b8 <- seq(from = 0, to = 0.99999, length.out =  12177)
v8 = qtweedie(b8, xi=NULL, miu8, fi8, power=pow11_1$p.max)
QQ8=glm(yt8 ~ v8, family = tweedie(var.power = pow11_1$p.max,
        link.power = 0), data=dat4_screen)
plot(yt8,v8, pch=1, main="Q-Q Tw Adulto mayor \n sin datos atípicos",
     col="black", xlab="Consultas médicas en Santiago",
     ylab="Cuantiles Tweedie", cex.main=1.2,
     ylim=c(0,200), xlim=c(0,2500)
)
abline(h=0,v=0, col="gray")
abline(QQ8, lty=7, col="red")


# ------------------------------------------------------------------
# Diagnóstico de los modelos sin datos atípicos
hist(residuals(modelo_SA, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adolescente", cex.lab=1, cex.main=1.3,
     col="lightcyan3")
plot(residuals(modelo_SA, type = "pearson"), ylim=c(-5,20),
     main="Adolescente", ylab="Residuos de Pearson")


hist(residuals(modelo2_SJ, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Joven", cex.lab=1, cex.main=1.3,
     col="lightcyan3")
plot(residuals(modelo2_SJ, type = "pearson"), ylim=c(-5,10),
     main="Joven", ylab="Residuos de Pearson")


hist(residuals(modelo3_SAd, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adulto", cex.lab=1, cex.main=1.3,
     col="lightcyan3")
plot(residuals(modelo3_SAd, type = "pearson"), ylim=c(-5,18),
     main="Adulto", ylab="Residuos de Pearson")
```

```
hist(residuals(modelo4_SAdm, type="deviance"), xlab="Devianza residual",
     ylab="Frecuencia", main="Adulto mayor", cex.lab=1, cex.main=1.3,
     col="lightcyan3")
plot(residuals(modelo4_SAdm, type = "pearson"), ylim=c(-5,12),
     main="Adulto mayor", ylab="Residuos de Pearson")
```