Diagnóstico de influencia en modelos lineales generalizados con respuesta gamma: Una aplicación.

Trabajo presentado por:
Camila Rojas Merani

Trabajo de titulación para optar al título de:
Ingeniero en Estadística

Profesor guía:
Marco Riquelme Álamos, Ph.D.

10 de septiembre de 2019

Agradecimientos

En primer lugar quiero agradecer al profesor Dr. Marco Riquelme Álamos por aceptar guiarme en mi trabajo de título, por confiar en mí y siempre motivarme a seguir.

Agradezco a mis padres por ser personas fundamentales en mi formación, a mis hermanos por todo su apoyo y ayuda en mi vida.

Agradezco a los profesores del Instituto de Estadística de la Universidad de Valparaíso, debido a su dedicación con los estudiantes.

Finalmente quiero agradecer a mi hija Constanza, que es la persona más importante en mi vida y mi pilar fundamental para salir adelante.

Resumen

En este trabajo se estudian e implementan las técnicas de diagóstico en el modelo lineal generalizado con respuesta gamma. Un paso importante en el análisis de un ajuste de regresión es verificar posibles desviaciones de los supuestos hechos para el modelo, así como la existencia de observaciones discrepantes con alguna interferencia desproporcionada o inferencial con los resultados del ajuste.

A lo largo de este trabajo de título se estudiaron los modelos: (1) modelo lineal generalizado y (2) modelo lineal generalizado con respuesta gamma. También se presentan algunas técnicas de diagnósticos usando el método de influencia global para verificar posibles alejamientos en los supuestos establecidos sobre el modelo. Finalmente se aplican dichas técnicas de diagnóstico a un conjunto de datos reales .

Abstract

In this work, diagnostic techniques are studied and implemented in the generalized linear model with gamma response. An important step in the analysis of a regression adjustment is to verify possible deviations from the assumptions made for the model, as well as the existence of discrepant observations with some disproportionate or inferential interference with the results of the adjustment.

Throughout this title work the models were studied: (1) generalized linear model and (2) generalized linear model with gamma response. Some diagnostic techniques are also presented using the global influence method to verify possible departures in the assumptions established on the model. Finally, technical diagnostic techniques were analyzed in a set of real data.

Índice general

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

El modelo lineal generalizado con respuesta gamma es utilizado para ajustar datos positivos asimétricos. Los datos con asimetría positiva se llaman así porque la "cola"de la distribución apunta hacia la derecha y porque el valor de asimetría es mayor que 0. Por ejemplo, los datos sobre salarios suelen ser asimétricos de esta manera: muchos empleados de una empresa ganan relativamente poco, mientras que cada vez menos personas ganan salarios muy elevados. Sin embargo, es interesante estudiar el comportamiento de este tipo de variables con covariables. Para esto se utiliza el modelo lineal generalizado, el cual a partir de un conjunto de variables, logra explicar el comportamiento de una característica particular sobre la variable que se encuentra bajo estudio (Alcaide, 2015). Considendo lo anterior se ajusta el modelo como en el caso de Dennis y Costantino (1988) que realizan el análisis de poblaciones en estado estacionario con el modelo de abundancia gamma. Bringi, Huang, Chandrasekar y Gorgucci (2002) utilizan una metodología para estimar los parámetros de un modelo gamma de distribución sobre el tamaño de la gota de lluvia a partir de datos de radar polarimétricos. Lawless, y Crowder (2004) estudian las covariables y efectos aleatorios en un modelo de proceso gamma con aplicación a la degradación y el fracaso. Semeraro, (2008) aplican un modelos gamma de variabilidad multivariana para las aplicaciones financieras. Camargo (2018) utiliza la regresión gamma generalizada para el análisis de datos espaciales.

Una vez ajustado el modelo es importante estudiar la existencia de observaciones discrepantes en los datos que provoquen alguna interferencia sobre los resultados derivados en el ajuste de la regresión, y de este modo ver si estos datos atípicos son o no influyentes en la estimación. Esta etapa es conocida como análisis de diagnóstico (Paula, 2004), parte de este análisis incluye el diagnóstico de influencia.

El análisis de diagnóstico de influencia consta de evaluar dicha influencia estudiando el conjunto de datos sin una determinada observación a través de la elimiación de casos (influencia global) (Rivas, 2017)

1.1. Motivación

Es importante investigar las técnicas de diagnósticos en los modelos lineales generalizados con respuesta gamma. Como es mencionado por Rivas (2017) donde explica que un punto fundamental en el análisis de un modelo es la detección de observaciones influyentes, ya que estas, si es que existen, pueden afectar las conclusiones extraídas de los resultados del modelo ajustado. En diversos reportes de investigación científica es posible encontrar estudios aplicados de este tipo de modelos. Por ejemplo, Grover, Mittal y Sabharwal (2013) realizan una aplicación del modelo lineal generalizado gamma para la estimación de la función de supervivencia de los pacientes con nefropatía diabética. Guo et al. (2015) investigan los efectos de las variables climáticas, considerando respuesta gamma, en el área quemada por incendios forestales en el noreste de China. Cribble y Ng (2016) aplican el modelo lineal generalizado gamma para modelar variables continuas, sesgadas y heteroscedásticas en el área de la psicología. Sin embargo, los análisis de diagnósticos de influencia que se reportan son muy escasos.

Por lo tanto, se pretende desarrollar en este trabajo de título un estudio utilizando técnicas de influencia global propuesta por Cook y Weisberg (1982) en el modelo lineal generalizado con respuesta gamma.

1.2. Descripción de los objetivos

El objetivo general de este trabajo es estudiar e implementar las metodologías de técnicas de diagnóstico en el modelo lineal generalizado con respuesta gamma.

De este modo, los objetivos específicos de este trabajo de título son:

- Abordar la estimación por máxima verosimilitud en un modelo lineal generalizado con respuesta gamma.
- Desarrollar criterios de selección de bondad de ajuste en esta clase de modelos.
- Aplicar los métodos de diagnóstico de influencia global en el modelo ajustado con respuesta gamma. La idea también es aplicar estas técnicas a un conjunto de datos reales.

1.3. Hipótesis

A través de la aplicación de los métodos de diagnóstico en el modelo lineal generalizado con respuesta gamma es posible efectuar un análisis de sensibilidad y así observar la existencia de puntos aberrantes en los datos que puedan estar provocando alguna interferencia sobre los resultados.

Capítulo 2

Modelo lineal generalizado con respuesta gamma

Este capítulo se centra en las herramientas que se utilizarán en este trabajo de título. De este modo, en la Sección 2.1 se presenta el modelo lineal generalizado y las componentes que lo definen. La Sección 2.2 trata sobre la distribución gamma y la Sección 2.3 presenta al modelos lineal generalizado gamma que es utilizado para datos positivos asimétricos.

2.1. Modelo lineal generalizado

Los modelos estadísticos se proponen para explicar un fenómeno que, a partir de la observación o de la experimentación, se logre explicar el comportamiento de una característica particular de los individuos o elementos de una población bajo estudio en base a las diferencias existentes entre las características asociadas. (Alcaide, 2015)

En cuanto al estudio, la variable que se desea explicar se conoce cómo variable respuesta o variable objetivo, mientras que las variables en las que se desea basar la explicación se denominan variables explicativas o covariables. Inicialmente, se trata de explicar una o varias variables objetivos, a través de un conjunto de variables explicativas, requiriendo la elección de un modelo que describa la estructura de la relación entre las variables. Para la elección del modelo es importante distinguir entre el tipo de variables que intervienen (continuas, de conteo, cualitativas, entre otros). Según el tipo de variables que intervienen y de la relación entre ellas, se dispone de un conjunto de posibles modelos más o menos adecuados, capaces de explicar la realidad. Generalmente, el modelo de regresión estándar es el modelo de regresión lineal, es decir, se modeliza la relación tratando de expresar la variable o variables explicativas, o alguna característica de ellas, a través de una combinación lineal de las covariables. El modelo lineal clásico consiste en expresar la esperanza condicionada de la variable objetivo como combinación lineal de las variables explicativas bajo la suposición de normalidad y homocedasticidad. Sin embargo, durante años se intentó describir la mayoría de los fenómenos aleatorios, incluso cuando el fenómeno en estudio no presentó una respuesta a la que el supuesto de normalidad fuera razonable (Alcaide, 2015)

Esta modelización clásica se puede extender a una familia más general, propuesta por Nelder (1972) y ampliada por McCullagh y Nelder (1989), conocida como modelos lineales generalizados (MLG). Los MLG se crearon para reunir varios modelos estadísticos que se trataron por separado (Paula, 2004). En general, los análisis de regresión buscaron algún tipo de transformación que lle-

varía a la normalidad, como Box-Cox (1964) dada a continuación

$$Z = \begin{cases} \frac{y^{\lambda}-1}{\lambda} \; si \; \lambda \neq 0 \\ log \; y \; si \; \lambda = 0 \end{cases},$$

en que $y > 0$ y λ es una constante desconocida.

De esta forma los MLG son una extensión de la teoría de los modelos lineales, pero se agrega la posibilidad de modelar variables respuestas, no exigiéndose necesariamente normalidad, sino que la idea básica es abrir el rango de opciones para la distribución de la variable de respuesta, lo que le permite pertenecer a la familia exponencial de distribuciones, así como dar mayor flexibilidad a la relación funcional entre la media de la variable de respuesta y el predictor lineal η (Paula, 2004)

En los modelos de regresión lineal se considera el supuesto de independencia para las observaciones, sin embargo, para esta nueva familia a diferencia del modelo clásico, las distribución de la componente aleatoria no necesariamente es homocedástica, es decir no se requiere de un supuesto de homogeneidad de varianzas (Alcaide, 2015)

2.1.1. Definición

A continuación se presentan las diferentes componentes que definen un modelo lineal generalizado y son la componente aleatoria, componente sistemática y la función de enlace (McCullagh y Nelder, 1989)

Componente aleatoria

Sea $Y_1, ..., Y_n$ variables aleatorias independientes e idénticamente distribuidas, cada una con una función de densidad o función de probabilidad dada de la forma:

$$f\left(y_i; \theta_i, \phi\right) = e^{\phi\{y_i\theta_i - b(\theta_i)\}} + c\left(y_i, \phi\right), \tag{2.1}$$

en que $f(\cdot)$ denota la función de probabilidad en el caso que Y sea una variable discreta, o la función de densidad en el caso que Y sea una variable continua, θ es el parámetro de localización o canónico, ϕ el parámetro escala.

Según Alcaide (2015), se verifica que:

$$E(Y_i) = \mu_i = b'(\theta_i), \tag{2.2}$$

$$Var(Y_i) = \phi^{-1}V(\mu_i), \quad i = 1, 2, ..., n, \tag{2.3}$$

en que $b'(\theta_i)$ es la primera derivada, y donde $V_i = V(\mu_i)$ se denomina función de varianza y $\phi^{-1} > 0$ ($\phi > 0$) es el parámetro de dispersión (precisión)

Componente sistemática

La componente sistemática recoge la variabilidad de Y_i expresada a través de p variables explicativas $X_1, ..., X_p$, que denotaremos por $\mathbf{X}$, y de sus correspondientes parámetros $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$. La componente sistemática, también denominado predictor lineal, se denota η_i, dada por

$$\eta_i = \mathbf{X}_i^\top \beta, \quad i : 1, 2, ..., n. \tag{2.4}$$

Función enlace

Es necesario la inclusión de una función que relacione el valor esperado con las variables explicativas. Esta función se denomina función enlace y se simboliza por $g(\mu_i)$.

$$g(\mu_i) = \eta_i. \tag{2.5}$$

La inversa de la función enlace dada por:

$$\mu_i = g^{-1}(\eta_i) = h(\eta_i) = h(\mathbf{X}_i\beta),$$

definiendo h como $h = g^{-1}$, función inversa.

2.1.2. Enlaces canónicos

En particular para cada elemento de la familia exponencial existe una función enlace denominada canónica o natural, que consiste en relacionar el parámetro natural directamente con el predictor lineal:

$$\theta_i = \theta(\mu_i) = \eta_i = \mathbf{X}_i\beta = g(\mu_i). \tag{2.6}$$

Suponiendo ϕ conocido, el logaritmo de la función de verosimilitud de un MLG con respuestas independientes se puede expresar de la forma (McCullagh y Nelder, 1991)

$$L(\beta) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.7}$$

Un caso particular importante ocurre cuando el parámetro canónico (θ) coincide con el predictor lineal, es decir, cuando $\theta_i = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. En ese caso, $L(\beta)$ está dado por:

$$L(\beta) = \sum_{i=1}^{n} \phi\{y_i \sum_{j=1}^{p} x_{ij}\beta_j - b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.8}$$

Definiendo la estadística $S_j = \phi \sum_{i=1}^{n} Y_i x_{ij}$, $L(\beta)$ se vuelve a reexpresar en la forma

$$L(\beta) = \sum_{j=1}^{p} s_j \beta_j - \phi \sum_{i=1}^{n} b\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.9}$$

Por lo tanto, por el teorema de la factorización la estadística $S = (S_1, ., , S_p)^\top$ es suficientemente pequeño para el vector $\beta = (\beta_1, ..., \beta_p)^\top$. Los enlaces que corresponden a tales estadísticas son llamadas canónicas y desempeñan un papel importante en la teoría de los MLG.

2.1.3. Función de devianza

La calidad del ajuste de un MLG se evalúa a través de la devianza y se expresa a continuación (McCullagh y Nelder, 1989):

$$D\left(y; \hat{\mu}\right) = 2\left\{L(y; y) - L(\hat{\mu}; y)\right\}. \tag{2.10}$$

El logaritmo de la función de verosimilitud es definido por (Paris, 2010)

$$L(\hat{\mu}; y) = \sum_{i=1}^{n} L(\mu_i; y_i), \tag{2.11}$$

en que, $\mu = g^{-1}(\eta_i)$. Luego el modelo con un parámetro por observación es el modelo saturado. Para este modelo en que, $(p = n)$ la funcion $L(\mu; y)$ está estimada por:

$$L(y; y) = \sum_{i=1}^{n} L(y_i; y_i). \tag{2.12}$$

Es decir la estimación por máxima verosimilitud (definida más adelante en (2.1.5)) de μ_i está dada por $\tilde{\mu}_i = y_i$. Cuando $p < n$, se denota la estimación de $L(\mu; y)$ por $L(\hat{\mu}; y)$. Entonces la calidad del ajuste del MLG se evalúa a través de la función de devianza.

$$D^*(y; \hat{\mu}) = \phi D(y; \hat{\mu}) = 2\{L(y; y) - L(\hat{\mu}; y)\}, \tag{2.13}$$

que es una distancia entre el logaritmo de la función de verosimilitud del modelo saturado (con n parámetros) y del modelo a estimar (con p parámetros). Si la diferencia es pequeña, se indica que con un número menor de parámetros el ajuste seria igual de bueno que el modelo saturado (Paula, 2004)

Finalmente para ver si esta diferencia es o no significativa, es comparado el valor con el percentil de alguna distribución de probabilidad.

2.1.4. Función score e información de Fisher

A través de la función score e información de Fisher se puede hallar el estimador de máxima verosilimitud a través de la primera y segunda derivada de la función de verosimilitud o log-verosimilitud.

Score y Fisher para β

Se considera la partición $\theta = (\beta^\top, \phi)^\top$ y denota el logaritmo de la función de verosimilitud por $L(\theta)$.

La función de score para el parámetro β está dada por (Paula, 2004)

$$U_\beta(\theta) = \frac{\partial L(\theta)}{\partial \beta} = \phi X^\top W^{1/2} V^{-1/2}(y - \mu), \tag{2.14}$$

en que X es una matriz $n \times p$ de puesto completo cuyas líneas serán denotadas por $x_i^\top$, $i = 1, ..., n$, $W = diag\{w_1, ..., w_n\}$ es la matriz de pesos, $V = diag\{V_1, ..., V_n\}$, $y = (y_1, ., y_n)^\top$ y $\mu = (\mu_1, ., \mu_n)^\top$.

La matriz de información de Fisher para β está dada por

$$K_{\beta\beta}(\theta) = E\{-\frac{\partial^2 L(\theta)}{\partial \beta \partial \beta^\top}\} = \phi X^\top W X. \tag{2.15}$$

Score y Fisher para ϕ

La función score para el parámetro ϕ está dado por (Paula, 2004)

$$\begin{aligned} U_\phi(\theta) &= \frac{\partial L(\theta)}{\partial \phi} \\ &= \sum_{i=1}^{n} \{y_i \theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c'(y_i, \phi), \end{aligned} \tag{2.16}$$

en que $c'(y_i, \phi) = dc(y_i, \phi)/d\phi$.

La información de Fisher para el parámetro ϕ está dado por

$$K_{\phi\phi}(\theta) = -\sum_{i=1}^{n} E\{c''(Y_i, \phi)\}. \tag{2.17}$$

2.1.5. Inferencia

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad (Codeiro, 2000)

Estimación de los parámetros

Es el procedimiento utilizado para conocer las características de un parámetro poblacional a partir del conocimiento de la muestra (Codeiro, 2000)

Dos de los métodos más comunes en la estimación estadística de parámetros son el método de Mínimos Cuadrados Ordinarios y el método de Máxima Verosimilitud. Sin embargo, el más adecuado es el Método de Máxima Verosimilitud, que tiene las propiedades de consistencia y eficiencia asintótica (Codeiro, 2000)

- Estimación por máxima verosimilitud

La estimación de máxima verosimilitud procura encontrar los valores más probables de los parámetros de la distribución para un conjunto de datos (McCullagh y Nelder, 1991). Entonces, siendo $y = (y_1, ..., y_n)^\top$ un conjunto de n observaciones aleatorias independientes cuya función de probabilidad es $f(y_i; \theta_i, \phi)$ y depende de un vector de parámetros θ_i y ϕ, su función de probabilidad conjunta es:

$$l(y_i; \theta_i, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi). \tag{2.18}$$

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

$$Ln(l(\theta_i, \phi)) = \sum_{i=1}^{n} Lnf(y_i; \theta_i, \phi). \tag{2.19}$$

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

$$Ln(l(\theta_i, \phi)) = \sum_{i=1}^{n} \left\{ Ln(c(y_i, \phi)) + \frac{y_i\theta_i - b(\theta_i)}{\phi} \right\}. \tag{2.20}$$

- Estimación de β

Paula (2004) señala que:

$$\beta^{(m+1)} = (X^\top W^{(m)} X)^{-1} X^\top W^{(m)} z^{(m)}, \tag{2.21}$$

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

- Estimación de ϕ

$$\sum_{i=1}^{n} c'(y_i, \hat{\phi}) = \frac{1}{2} D(y; \hat{\mu}) - \sum_{i=1}^{n} \{y_i \tilde{\theta}_i - b(\tilde{\theta}_i)\}, \tag{2.22}$$

en que $D(y; \hat{\mu})$ denota la devianza del modelo bajo investigación.

Intervalos de confianza

Se llama intervalo de confianza a un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto.

Según Codeiro (2000) un intervalo de confianza asintótico de coeficiente $1 - \alpha$ puede ser construída para $\mu(z) = g^{-1}(z^\top \beta)\ \forall z \epsilon \mathbb{R}^p$. Asintóticamente se tiene que $(\hat{\beta} - \beta) \sim N_p(0, \phi^{-1}(X^\top W X)^{-1})$. Luego un intervalo de confianza de $1 - \alpha$ para $z^\top \beta\ \forall z \epsilon \mathbb{R}^p$, está dado por

$$z^\top \hat{\beta} \pm \sqrt{\phi^{-1} c_\alpha} \{z^\top (X^\top W X)^{-1} z\}^{\frac{1}{2}} \quad \forall z \epsilon \mathbb{R}^p, \tag{2.23}$$

en que c_α es tal que $Pr\{\chi_p^2 \leq c_\alpha\} = 1 - \alpha$.

Test de hipótesis

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra (Codeiro, 2000)

"Buse (1982) presenta de manera muy didáctica la interpretación geométrica de la razón de verosimilitud, la puntuación y las pruebas de Wald para el caso de Hipótesis simples. Las siguientes son generalizaciones para MLG"(Paula, 2004)

- Test de razón de verosimilitud

La prueba de razón de verosimilitud, se puede expresar para los MLG como la diferencia entre dos funciones de devianza.

$$\xi_{RV} = \phi\{D(y; \hat{\mu}^0) - D(y; \hat{\mu})\}, \tag{2.24}$$

en que $\hat{\mu}^0$ es la estimación de máxima verosimilitud bajo H_0 y $\hat{\mu}^0 = g^{-1}(\hat{\eta}^0)$, $\hat{\eta}^0 = X\beta^0$.

- Test de Wald

Para los MLG el estadístico de Wald se expresa de la forma

$$\xi_W = \phi[\hat{\beta} - \beta^0]^\top (X^\top \hat{W} X)[\hat{\beta} - \beta^0]. \tag{2.25}$$

- Test score

el test score, también conocido como prueba de Rao para los MLG tenemos que

$$\xi_{SR} = \phi^{-1} U_\beta(\beta^0)^\top (X^\top \hat{W}_0 X)^{-1} U_\beta(\beta^0), \tag{2.26}$$

en que $\hat{W}_0$ se estima en H_0, aunque tiene la forma del modelo en H_1.

- Test F

El estadístico F toma la siguiente forma para el caso de hipótesis simples

$$F = \frac{\{D(y;\hat{\mu}^0) - D(y;\hat{\mu})\}/p}{D(y;\hat{\mu})/(n-p)},$$

que para $\phi \to \infty$ y bajo H_0 sigue una $F_{p,(n-p)}$.

2.1.6. Selección del modelo

Existen varios procedimientos para seleccionar modelos de regresión, aunque ninguno de ellos es consistente, es decir, incluso para muestras grandes selecciona con probabilidad uno las variables explicativas con coeficiente de regresión distinto de cero. (Paula, 2004)

Algunos métodos para la selección del modelo son método forward, backward, stepwise y el criterio de akaike (Garcia, Castellana, Rapelli, Koegel y Catalano, 2014), los cuales son definidos a continuación:

Método forward

El método comienza por el modelo $\mu = \alpha$.

$$\mu = \alpha + \beta_j x_j, (j = 1, ..., q).$$

Sea P el nivel descriptivo más bajo entre las pruebas q. Si $P \leq P_E$, la variable correspondiente entra en el modelo.

Suponiendo que la variable entra en el modelo, entonces en el siguiente paso ajustamos el modelo

$$\mu = \alpha + \beta_j x_j, (j = 2, ..., q).$$

Sea P el nivel descriptivo más bajo entre las pruebas $q-1$. Si $P \leq P_E$, la variable correspondiente entra en el modelo. Se repite el procedimiento hasta que ocurra $P > P_E$

Método backward

$$\mu = \alpha + \beta_1 x_1 + ... + \beta_q x_q,$$

Sea P el nivel descriptivo más grande entre las pruebas q. Si $P > P_S$, la variable correspondiente sale del modelo.

Suponiendo que la variable sale de el modelo, entonces en el siguiente paso ajustamos el modelo

$$\mu = \alpha + \beta_2 x_2 + ... + \beta_q x_q.$$

Sea P el nivel descriptivo más alto entre las pruebas $q-1$. Si $P > P_S$, la variable correspondiente sale de el modelo. Se repite el procedimiento hasta que ocurra $P \leq P_S$

Método stepwise

Es una mezcla de los dos procedimientos anteriores. Comenzamos el proceso con el modelo $\mu = \alpha$.

Luego de que dos variables se incluyeron en el modelo, se debe verificar si la primera no ha salido del modelo. El proceso continúa hasta que no se incluye ninguna variable o retirada del modelo.

Se sugiere usar $P_E = P_S = 0,20$.

Método akaike

El criterio de información de Akaike, propuesto por Akaike (1974), proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos experimentales.

Como el logaritmo de la función de verosimilitud crece con el aumento del número de parámetros del modelo, una propuesta razonable sería encontrar el modelo con menor valor para la función

Este criterio se define como:

$$AIC = k - 2Ln\hat{L},$$

siendo k el número de parámetros del modelo estadístico estimado y $\hat{L}$ es el logaritmo de la función de máxima verosimilitud, que permite determinar los valores de los parámetros libres de un modelo estadístico

2.2. Distribución gamma

En este caso se asume que Y es una variable aleatoria con distribución gama de media μ y coeficiente de variación $\phi^{-1/2}$, se denota $Y \sim G(\mu, \phi)$, y cuya función de densidad se expresa en la forma

$$\begin{aligned} f(y;\mu,\phi) &= \frac{1}{\Gamma(\phi)}\left(\frac{\phi y}{\mu}^{\phi}\right) exp\left(-\frac{\phi y}{\mu}\right)(logy) \\ &= exp[\phi\{(-y/\mu) - log\mu\}log\Gamma(\phi) + \phi log(\phi y) - logy], \end{aligned} \tag{2.27}$$

en que $y > 0$, $\phi > 0$, $\mu > 0$ y $\Gamma(\phi) = \int_0^\infty t^{\phi-1}e^{-t}dt$ es la función gamma. Se puede notar que a medida que ϕ aumenta la distribución gamma es más simétrica en torno a la media y Y se aproxima a una distribución normal de media μ y varianza $\mu^2\phi^{-1}$. Por lo tanto, la distribución de gama se vuelve atractiva para el estudio de variables aleatorias asimétricas.

Los momentos centrales de Y se expresan de la siguiente manera:

$$E(Y-\mu)^r = \frac{(r-1)!\mu^r}{\phi^{r-1}},$$

para $r = 1, 2, ..$ Por lo tanto, expandiendo $logY$ en serie de Taylor alrededor de μ hasta el segundo orden se obtiene que

$$logY \cong log\mu + \frac{1}{\mu}(Y-\mu) - \frac{1}{2\mu^2}(Y-\mu)^2.$$

Por lo tanto, para ϕ grande se tiene

$$E(logY) \cong log\mu - \frac{1}{2\mu^2}E(Y-\mu)^2 = log\mu - \frac{1}{2\mu^2}\frac{\mu^2}{\phi} = log\mu - (2\phi)^{-1}e,$$

$$Var(logY) \cong \phi^{-1}.$$

Es decir, la transformación $logY$ estabiliza la varianza a medida que el coeficiente de variación de Y es pequeño.

2.3. Modelo lineal generalizado con respuesta gamma (MLGG)

El modelo lineal generalizado con respuesta gamma es utilizado para datos positivos asimétricos. Los datos con asimetría positiva se llaman así porque la "cola"de la distribución apunta hacia la derecha y porque el valor de asimetría es mayor que 0. Los datos positivos asimétricos se observan particularmente en tiempos de supervivencia (o duración), este tipo de datos se observan con fuerte énfasis en diferentes áreas de la medicina, ingeniería, pesca, meteorología, finanzas, seguros, entre otros (Paula, (2004)

2.3.1. Definición

A continuación se recogen las diferentes componentes que definen un modelo lineal generalizado con respuesta gamma, que son la componente aleatoria, componente sistemática y la función de enlace (McCullagh y Nelder, 1989)

Componente aleatoria

Suponiendo que $Y_1, ..., Y_n$ son variables aleatorias independientes tales que:

$$Y_i \sim G(\mu_i, \phi),\ i = 1, 2, ..., n. \tag{2.28}$$

Es decir, se asume que estas variables poseen medias diferentes e incluso coeficiente de variación $\phi^{-1/2}$.

Componente sistemática

Dado μ_i y un predictor lineal que se presenta a continuación:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} = X_i^\top \beta, \tag{2.29}$$

en que $x_i = (x_{i1}, ..., x_{ip})^\top$ contiene valores de variables explicativas y $\beta = (\beta_1, ..., \beta_p)^\top$ siendo el vector de parámetros de interés.

Función enlace

La componente aleatoria y sistemática se combinan en el modelo y es expresado a través de la función de enlace

$$g(\mu_i) = \eta_i. \tag{2.30}$$

2.3.2. Enlaces canónicos en MLGG

La función de enlace canónico para el MLG gamma se denomina función de vínculo recíproca y está dada por:

$$\frac{1}{\mu_i} = \eta_i. \tag{2.31}$$

2.3.3. Función de devianza del MLGG

La calidad del ajuste de un MLG se evalúa a través de la devianza y para el caso gamma se expresa de la siguiente forma (McCullagh y Nelder, 1989)

Cuando todos los valores son positivos

$$D\left(y;\hat{\mu}\right)=2\sum_{i=1}^{n}\left\{-log\left(\hat{\mu}_i/y_i\right)+\left(y_i-\hat{\mu}_i\right)/\hat{\mu}_i\right\}. \tag{2.32}$$

Con $\hat{\mu}_i=g^{-1}(\hat{\eta}_i$ y $\hat{\eta}_i=X_i^{\top}\hat{\beta}$. Si algún componente de y_i es igual a cero la devianza es indeterminada. Se sugiere sustituir $D\left(y;\hat{\mu}\right)$ por (McCullagh y Nelder, 1989)

$$D^{*}\left(y;\hat{\mu}\right)=2\phi\sum_{i=1}^{n}log(\hat{\mu}_i/y_i). \tag{2.33}$$

2.3.4. Función score e información de Fisher para MLGG

La función score e información de Fisher son utiles para hallar el EMV a través de la primera y segunda derivada de la función de vesosimilitud o la log-verosimilitud.

Score y Fisher para β

Considerando $V(\mu)=\mu^2$. Luego $w=\mu^2\left(d\theta/d\eta\right)^2$.
La función de score para el parámetro β está dada por

$$U_{\beta}(\theta)=\phi X^{\top}V^{-1/2}(y-\mu). \tag{2.34}$$

La matriz de información de Fisher para β está dada por

$$K_{\beta\beta}=\phi X^{\top}X. \tag{2.35}$$

Score y Fisher para ϕ

La función score para el parámetro ϕ está dado por

$$U_{\phi}=-\sum_{i=1}^{n}\left(\frac{y_i}{\mu_i}+log\mu_i\right)+\sum_{i=1}^{n}c'(y_i,\phi), \tag{2.36}$$

en que, $c'(y_i,\phi)=logy_i+log\phi+1-\psi\left(\phi\right)$ y $\psi\left(\phi\right)=\Gamma'(\phi)/\Gamma(\phi)$.

Luego $c''(y_i, \phi) = 1/\phi - \psi(\phi)$, por lo tanto

La matriz de información de Fisher para ϕ está dada por

$$K_{\phi\phi} = -\sum_{i=1}^{n} E\left\{c''(Y_i, \phi)\right\} = n\left\{\phi\psi'(\phi) - 1\right\}/\phi, \tag{2.37}$$

en que, $\psi'(\phi) = d\psi(\phi)/d\phi$.

2.3.5. Inferencia en MLGG

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra (Codeiro, 2000). La inferencia en el modelo lineal generalizado gamma es presentada a continuación:

Estimación de los parámetros

Es el procedimiento utilizado para conocer las características de un parámetro poblacional a partir del conocimiento de la muestra (Codeiro, 2000)

- Estimación por máxima verosimilitud

  La función de log-verosimilitud para el MLGG se representa de la siguiente forma:

$$Ln(l(\mu_i, \phi)) = \sum_{i=1}^{n} \left\{(\phi - 1)logy + \phi log\phi + log\Gamma(\phi) + \frac{y_i\mu_i + log(-\mu_i)}{\phi}\right\}. \tag{2.38}$$

- Estimación de β

Sea $\beta = (\beta_1, ..., \beta_p)^\top$ el vector de parámetros de interés. Las conexiones más usadas en el caso gamma son identidad ($\mu_i = \eta_i$), logarítmica ($log\mu_i = \eta_i$) y recíproca ($\mu_i = \eta_i^{-1}$), esta última siendo la conexión canónica. El proceso iterativo para la estimación de β esta dado por

$$\beta^{(m+1)} = (X^\top W^{(m)} X)^{-1} X^\top W^{(m)} z^{(m)}, \tag{2.39}$$

en que $m = 0, 1, ...$, es una variable dependiente modificada $z = \eta + W^{-1/2} V^{-1/2}(y - \mu)$, $\eta = (\eta_1, ..., \eta_n)^\top$, $y = (y_1, ..., y_n)^\top$, $\mu = (\mu_1, ..., \mu_n)^\top$, $V = diag\{\mu_1, ..., \mu_n\}$ y $W = diag\{w_1, ..., w_n\}$ con $w_i = (d\mu_i/d\eta_i)^2/\mu_i$.

Es interesante notar que bajo conexión logarítmica los pesos del proceso iterativo para la obtención de β quedan dados por $w_i = \mu_i^2/\mu_i^2 = 1$, de modo que el proceso iterativo asume la forma simplificada

$$\beta^{(m+1)} = (X^\top X)^{-1} X^\top z^{(m)}, \tag{2.40}$$

en que $z = (z_1, ..., z_n)^\top$ con $z_i = \eta_i = (y_i - \mu_i)/\mu_i$ y $\mu_i = exp(\eta_i)$. La varianza asintótica de $\hat{\beta}$ esta dada por $Var(\hat{\beta}) = \phi^{-1}(X^\top X)^{-1}$. En particular, si las columnas de la matriz X son ortogonales, esto es $X^\top X = I_p$, en que I_p es la matriz de orden p, entonces $Var(\hat{\beta}_j) = \phi^{-1}$ y $Cov(\hat{\beta}_j, \hat{\beta}_\ell) = 0$, para $j \neq \ell$, es decir, $\hat{\beta}_j$ y $\hat{\beta}_\ell$ son asintóticamente independientes.

Test de hipótesis

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra (Codeiro, 2000)

- Test de Wald

Para los MLG gamma el estadístico de Wald se expresa de la forma

$$\xi_W = \phi[\hat{\beta} - \beta^0]^\top (X^\top \hat{W} X)[\hat{\beta} - \beta^0]. \tag{2.41}$$

- Test score

el test score, también conocido como prueba de Rao para los MLG gamma tenemos que

$$\xi_{SR} = \phi^{-1} U_\beta(\beta^0)^\top (X^\top \hat{W}_0 X)^{-1} U_\beta(\beta^0), \tag{2.42}$$

en que $\hat{W}_0$ se estima en H_0, aunque tiene la forma del modelo en H_1.

2.3.6. Selección del modelo

Existen varios procedimientos para seleccionar modelos de regresión, aunque ninguno de ellos es consistente, es decir, incluso para muestras grandes selecciona con probabilidad uno las variables explicativas con coeficiente de regresión distinto de cero (Paula, 2004)

Método akaike

El criterio de información de Akaike, propuesto por Akaike (1974), proporciona un método simple y objetivo que selecciona el modelo más adecuado para caracterizar los datos experimentales. Se describe la relación entre el sesgo y varianza en la construcción del modelo, o hablando de manera general acerca de la exactitud y complejidad del modelo (Garcia, Castellana, Rapelli, Koegel y Catalano, 2014)

$$AIC = D^*(y; \hat{\mu}) + 2p, \tag{2.43}$$

en que $D^*(y; \hat{\mu})$ denota la devianza del modelo y p el número de parámetros.

Capítulo 3

Técnicas de diagnóstico en modelo con respuesta gamma

El capítulo 3 trata sobre las técnicas de diagnósticos en los modelos lineales generalizados. La Sección 3.1 define las técnicas de diagnóstico del MLG, la Sección 3.2 define las técnicas de diagnóstico del MLG gamma y la Sección 3.3 trata del diagnóstico de influencia, específicamente influencia global.

3.1. Técnicas de diagnóstico

Un paso importante en el análisis de un ajuste de regresión es verificar posibles desviaciones de los supuestos hechos para el modelo, especialmente para la componente aleatoria y la parte sistemática del modelo, así como la existencia de observaciones discrepantes con alguna interferencia desproporcionada o inferencial con los resultados del ajuste. Este paso, conocido como análisis de diagnóstico, se ha implementado durante mucho tiempo y comenzó con el análisis de residuos para detectar la presencia de puntos aberrantes y evaluar la adecuación de la distribución propuesta para la variable de respuesta. Una referencia importante en este tema es el artículo de Cox y Snell (1968) en el que se presenta una forma general de definir residuos. Belsley, Cook y Weisberg (1982) discuten la estandarización de los residuos para el caso normal lineal. Pregibon (1981) propone el componente de desviación como un residuo en la clase de modelos lineales generalizados y sugiere una estandarización probada por Cordeiro (1982) utilizando los enfoques propuestos por Cox y Snell (1968). Atkinson (1981) propone la construcción mediante la simulación de Monte Carlo de una banda de confianza para los residuos de la regresión lineal normal, que denominó "Q-Q Plot", que permite una mejor comparación entre los residuos y los percentiles de la distribución normal estándar.

Otro tema importante en el análisis de diagnóstico es la detección de observaciones influyentes, es decir, puntos que ejercen un peso desproporcionado en las estimaciones de los parámetros del modelo. Han surgido varias propuestas relacionadas con la influencia de las observaciones en los coeficientes estimados del modelo normal lineal. El estudio de la diagonal principal de la matriz de proyección presentada por Hoaglin y Welsch (1978) motivó la definición de los puntos de apalancamiento (también conocidos como puntos Leverage y son definidos a continuación), estos puntos tienen un perfil diferente de los otros puntos con respecto a los valores de las variables explicativas que dependiendo de la ubicación, pueden influir fuertemente en las estimaciones del coeficiente de regresión. Wei, Hu y Fung (1998) extienden la definición de los puntos de apalancamiento a modelos

generalizados cuya variable de respuesta es continua. Esta generalización incluye otros métodos de estimación además de la máxima verosimilitud y otros enfoques, como el enfoque bayesiano. Paula (1999) analiza los puntos de apalancamiento en modelos lineales normales restringidos con extensiones a MLG.

Sin embargo, la eliminación de puntos es quizás la técnica más conocida para evaluar el impacto de retirar una observación particular en las estimaciones de regresión. Cook, Peña y Weisberg (1988) comparan la distancia basada en la probabilidad con medidas tradicionales de eliminación de puntos como la distancia de Cook.

3.1.1. Matriz sombrero (Leverage)

La idea principal que está detrás del concepto de punto de apalanca o matriz "sombrero" $P_{n\times n}$ es de evaluar la influencia de y_i sobre el propio valor ajustado $\hat{y}_i$ (McCullagh y Nelder, 1989). En especial su diagonal juega un rol importante en las técnicas de diagnóstico de influencia global.

Los elementos diagonales de P están limitados de la siguiente forma:

$$\frac{1}{n} \leq p_{ii} \leq 1.$$

Y proveen una medida de distancia desde el i-ésimo caso de la media, definida por:

$$p_{ii} = \frac{1}{n} + (x_i - \bar{x})^t (X^T X)^{-1} (x_i - \bar{x}),$$

en que $x_i - \bar{x}$ denota el vector de orden $(p \times 1)$ de predictores de la matriz de diseño para el i-ésimo caso de la matriz sombrero (p_{ii}).

La matriz sombrero para el MLG es:

$$\hat{GL} = \hat{V}(X^\top \hat{V} X)^{-1} X^\top.$$

3.1.2. Residuos

El análisis de residuos es una inspección gráfica que ayudan a encontrar casos desviantes y puntos influyentes.

Residuo de Anscombe

Propuesto por Anscombe (1972), este trata de normalizar la diferencia entre los valores observados y ajustados de manera que la heterogeneidad en los datos sea identificable.

Se define de la siguiente manera:

$$t_{Ai} = \frac{\phi^{1/2}\{\psi(y_i) - \psi(\hat{\mu}_i)\}}{\hat{V}^{1/2}(\hat{\mu}_i)\psi'(\hat{\mu}_i)},$$

en que, $\psi(\mu) = \int_0^\mu V^{-1/3}(t)dt$, y que corresponde a una transformación utilizada para normalizar la distribución de Y.

Sin embargo, los residuos más utilizados en MLG son definidos a partir de las componentes de la función de devianza. La versión estandarizada es la siguiente:

Residuo estandarizado

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

en que $d(y_i;\hat{\mu}_i) = \pm\sqrt{2}\{y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\}^{1/2}$.

Se verificó mediante simulaciones que la distribución de t_{D_i} tiende a estar más cerca de lo normal que las distribuciones de otros residuos (Williams, 1984)

3.1.3. Gráfico de variable agregada

La siguiente es la versión gráfica de la variable agregada para el MLG. Supongamos un MLG con p parámetros, $\beta_1, ..., \beta_p$ ϕ conocido, y que se está incluyendo un parámetro adicional γ en el modelo. La idea es probar $H_0 : \gamma = 0$ contra $H_1 : \gamma \neq 0$ (Paula, 2004)

Entonces, sea $\eta(\beta, \gamma)$ un predictor lineal con $p + 1$ parámetros, esto es:

$$\eta(\beta, \gamma) = X^\top \beta + \gamma Z. \tag{3.1}$$

Una función score para γ está dada por:

$$U_\gamma = \frac{\partial L(\beta, \gamma)}{\partial \gamma} = \phi^{1/2} Z^\top W^{1/2} r_p, \tag{3.2}$$

en que $z = (z_1, ..., z_n)^\top$, del resultado anterior se tiene que :

$$Var(\hat{\gamma}) = \phi^{-1}[Z^\top W^{1/2} M W^{1/2} Z]^{-1}, \tag{3.3}$$

en que, $M = I_n - H$. Luego $Var(\hat{\gamma}) = \phi^{-1}(R^\top W R)^{-1}$ con $R = Z - XC$ y $C = (X^\top W X)^{-1} X^\top W Z$.

Por lo tanto, el estadístico score para probar $H_0 : \gamma = 0$ v/s $H_1 : \gamma \neq 0$ viene dado por:

$$\xi_{SR} = \frac{(\hat{r}_P^\top \hat{W}^{1/2} Z)^2}{(Z^\top \hat{W}^{\frac{1}{2}} \hat{M} \hat{W}^{1/2} Z)},$$

en que $\hat{W}$, $\hat{r}_p$ y $\hat{M}$ son evaluados en $\hat{\beta}$ (bajo H_0). Bajo H_0, $\xi_{SR} \sim \chi^2$, cuando $n \to \infty$

3.1.4. Técnicas gráficas

Las técnicas gráficas más recomendadas para los MLG son

- gráficos de t_{Di} contra el orden de las observaciones, contra los valores ajustados y contra las variables explicativas.
- gráfico normal de probabilidades para t_{Di}.
- gráfico de $\hat{z}_i$ contra $\hat{\eta}_i$ para verificar la adecuación de la función de enlace.
- gráficos de la distancia de Cook cuando la i-ésima observación es excluida.

3.2. Técnicas de diagnóstico del MLGG

Como se mencionó en la sección anterior, las técnicas de diagnóstico son un paso importante en el ajuste de la regresión al verificar la existencia de posibles desviaciones en los supuestos obtenidos en el modelo. A continuación se presentan algunas técnicas de diagnóstico del MLG con respuesta gamma:

3.2.1. Residuos

Como ya se sabe el análisis de residuos es una inspección gráfica que ayudan a encontrar puntos aberrantes y puntos influyentes.

Resíduo de Anscombe

Propuesto por Anscombe (1972). Los residuos de Anscombre utiliza la función de varianza propia del modelo y para el MLGG está dado de la siguiente manera:

$$t_{Ai} = \frac{3\phi^{1/2}\{y_i^{1/3} - \hat{\mu}_i^{1/3}\}}{\hat{\mu}_i^{1/3}}, \quad i = 1, 2, ...n.$$

Resíduo estandarizado

El resíduo componente de la desviación estandarizada asume para los modelos gamma la forma

$$t_{D_i} = \pm\frac{\sqrt{2\phi}}{\sqrt{1-\hat{h}_{ii}}}\{log(\hat{\mu}_i/y_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i\}^{1/2}, \tag{3.4}$$

en que $y_i > 0$ y $\hat{h}_{ii}$ es el i-ésimo elemento de la diagonal principal de la matriz.

$H = W^{1/2}X(X^\top WX)^{-1}X^\top W^{1/2}$ con $w_i = (d\mu_i/d\eta_i)^2/\mu_i^2$. En particular cuando hay un intercepto en η_i el residuo componente de la desviación t_{D_i} asume la forma reducida

$$t_{D_i} = \pm\frac{\sqrt{2\phi}}{\sqrt{1-\hat{h}_i i}}\{log(\hat{\mu}_i/y_i)\}^{1/2}. \tag{3.5}$$

Los estudios de simulación indican que el residuo t_{D_i} se aproxima a la normalidad, particularmente para ϕ grande.

3.2.2. Técnicas gráficas

Gráficos de t_{D_i} y $\hat{h}_{ii}$ contra los valores ajustados μ_i como también gráficos de los índices de LD_i se recomiendan para el análisis de diagnóstico.

3.3. Diagnóstico de influencia

Codeiro (2000) menciona que una etapa importante en el modelado estadístico es verificar posibles alejamientos en las suposiciones establecidas sobre el modelo, así como la existencia de observaciones discrepantes con alguna interferencia desproporcional sobre los resultados derivados del ajuste de la regresión. En la literatura estadística esta etapa se conoce como análisis de diagnóstico. En este contexto, se han desarrollado diversos procedimientos para detectar la presencia de observaciones discrepantes. entre las técnicas iniciales más usadas se encuentran en el análisis de residuos y eliminación de casos. El análisis de residuos sugiere el uso de una inspección gráfica de los residuos estandarizados. La eliminación de casos propone evaluar el impacto de cada observación sobre las estimaciones de la regresión mediante la retirada individual de cada observación del conjunto de datos (influencia global). Una observación es influyente si el efecto de excluirla del conjunto de datos produce diferencias significativas en el análisis.

Cook (1977) propone un importante procedimiento para la detección de observaciones influyentes en la regresión lineal; Cook y Weisberg (1982) desarrollan algunas medidas de diagnóstico para el modelo de regresión basadas en los residuos e influencia global; Hawkins (1980) y Leroy y Rousseeaw (1987) trata el problema de identificación de observaciones aberrantes.

3.3.1. Diagnóstico de influencia global

Dada la importancia que tienen los estimadores de los parámetros de un modelo en el análisis de regresión, es común estudiar el cambio que se puede producir en dichos estimadores al eliminar un caso o un subconjunto de ellos del conjunto de datos.

La idea básica es utilizar una medida que permita comparar la distancia entre el estimador de máxima verosimilitud obtenido luego de eliminar un subconjunto de observaciones, y el correspondiente estimador sin eliminar observaciones (Rivas, 2019)

Sean $l(\theta|y_c)$ y $l(\theta|y_{c[i]})$, la función log-verosimilitud del vector q-dimensional de los parámetros θ para los datos completos y para los datos con la eliminación del i-ésimo caso, respectivamente, donde un subíndice $[i]$ significa la cantidad original con el caso eliminado.

En que, $Q(\cdot)$ para el conjunto de datos sin la i-ésima fila está dado por

$$Q_i(\theta|\hat{\theta}) = E\left\{l(\theta|y_{c[i]})|y_{0[i]}, \hat{\theta}\right\}, \tag{3.6}$$

cuyo máximo se denota por $\hat{\theta}_{[i]}$, $i = 1, ..., n$ y $\hat{\theta}$ es el estimado de θ. Zhu et al. (2001) han propuesto el uso de la distancia de Cook generalizada de la $Q-$función. En este caso, la distancia entre $\hat{\theta}$ y $\hat{\theta}_{[i]}$ viene dada por

$$LD_i = \frac{(\hat{\theta} - \hat{\theta}_{[i]})^\top \left\{\ddot{Q}(\hat{\theta} \mid \hat{\theta})\right\} (\hat{\theta} - \hat{\theta}_{[i]})}{q}, \quad \forall i = 1, ..., n. \tag{3.7}$$

Usando aproximaciones lineales de un paso $\hat{\theta}[i]^{(1)}$ de $\hat{\theta}[i]$ (Pregibon, 1981), se obtiene

$$\hat{\theta}[i]^{(1)} = \hat{\theta} + \left\{-\ddot{Q}\left(\hat{\theta} \mid \hat{\theta}\right)\right\}^{-1} \dot{Q}_{[i]}\left(\hat{\theta} \mid \hat{\theta}\right), \tag{3.8}$$

en que $\dot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta}) = \frac{\partial Q_{[i]}(\theta|\hat{\theta})}{\partial \theta}_{|\theta=\hat{\theta}}$ y $\ddot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta})_{|\theta=\hat{\theta}}$ del algoritmo estimado es reemplazado por $\ddot{Q}(\hat{\theta} \mid \hat{\theta})$.

Luego se obtiene

$$LD_i^{(1)} = \frac{\left(\dot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta})^\top\right)\left\{-\ddot{Q}\left(\hat{\theta} \mid \hat{\theta}\right)\right\}^{-1}(\dot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta}))}{q}$$

$$\approx \left(\dot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta})^\top\right)\left\{-\ddot{Q}\left(\hat{\theta} \mid \hat{\theta}\right)\right\}^{-1}(\dot{Q}_{[i]}(\hat{\theta} \mid \hat{\theta})).$$

Zhu et al. (2001) definen la Q-distancia como:

$$QD_i = 2\left\{Q(\hat{\theta} \mid \hat{\theta}) - Q(\hat{\theta}_{[i]} \mid \hat{\theta})\right\}. \tag{3.9}$$

Luego obtienen una aproximación de QD_i

$$QD_i^{(1)} = 2\left\{Q(\hat{\theta} \mid \hat{\theta}) - Q(\hat{\theta}_{[i]} \mid \hat{\theta})\right\}. \tag{3.10}$$

3.3.2. Diagnóstico de influencia global para el MLGG

Cuando la i-ésima observación es excluida la distancia de Cook aproximada para el MLG gamma está dado por (Pregibon, 1981)

$$LD_i = \frac{\phi\hat{h}_{ii}}{(1-\hat{h}_{ii})^2}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i^2}. \tag{3.11}$$

Capítulo 4

Aplicación

En este capítulo se aplican las técnicas de diagnóstico definidas en el Capítulo 3 a un conjunto de datos reales en donde se evaluó la resistencia (en horas) de un tipo particular de vidrio de acuerdo con cuatro niveles de voltaje (en kilovoltios) y dos temperaturas (en grados Celsus). El análisis es realizado en el paquete estadístico R. El problema puede ser encontrado en Lawless, 1982, p. 338.

4.1. Descripción de los datos

Uno de los objetivos principales de este trabajo es implementar las metodologías de técnicas de diagnóstico en el modelo lineal generalizado con respuesta gamma. Para llevar esto a cabo se analiza la resistencia de un tipo particular de vidrio en relación con el voltaje y temperatura. Los datos están compuestos por el tiempo de resistencia, el voltaje (1: 200kV, 2: 250kV, 3: 300kV y 4: 350kV) y la temperatura (1: 170°C y 2: 180°C). En este caso la variable respuesta es la resistencia del vidrio, mientras que los factores son el voltaje y la temperatura.

Sea Y_{ijk} el tiempo de resistencia de la k-ésima muestra de vidrio sometida a la i-ésima temperatura y en el j-ésimo voltaje. Es habitual en este tipo de estudio asumir respuestas con alguna distribución asimétrica. Entonces supongamos $Y_{ijk} \sim G(\mu_{ij}, \phi)$.

En la Tabla 4.1 se presentan los resultados de un experimento en el que se evaluó la resistencia (en horas) de un tipo particular de vidrio de acuerdo con cuatro niveles de voltaje (en kilovoltios) y dos temperaturas (en grados Celsus).

Tabla 4.1: Tiempo de resistencia del vidrio de acuerdo con los niveles de voltaje y temperatura.

	Voltaje			
Temperatura	200(kV)	250(kV)	300(kV)	350(kV)
170 °C	439	572	315	258
	904	690	315	258
	1092	904	439	347
	1105	1090	628	588
180 °C	959	216	241	241
	1065	315	315	241
	1065	455	332	435
	1087	473	380	455

4.2. Análisis descriptivo

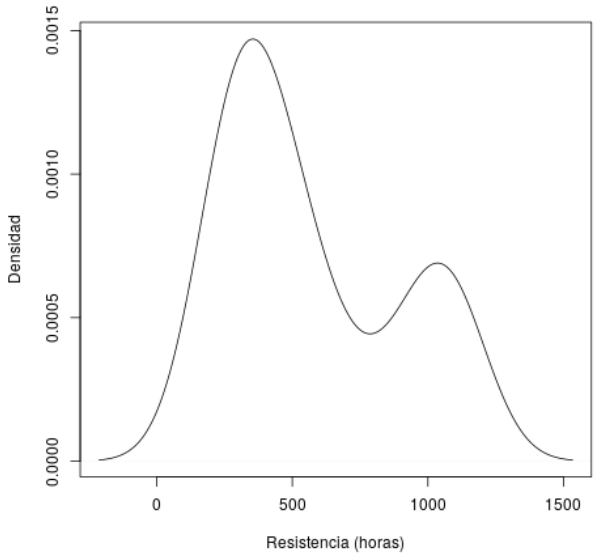
A continuación se realiza un análisis descriptivo de los datos, el que ayuda a comprender la estructura que tienen, de tal manera detectar tanto un patrón de comportamiento general como alejamientos de alguna observación. Una forma de realizar ésto es mediante gráficos de sencilla realización e interpretación. Otra forma de describir los datos es obtener medidas de resumen clásicas (Codeiro, 2000)

Las variables consideradas en estudio son las siguientes:

- Variable Respuesta: tiempo de resistencia de los vidrios.
- Variables Independientes (Factores): voltaje y temperatura.
  - **Niveles de voltaje**: 1=200kV, 2=250kV, 3=300kV e 4=350kV;
  - **Niveles de temperatura**: 1=170°C e 2=180°C.

Se nota por la Figura 4.1 (ignorando los niveles de voltaje y temperatura) una cierta asimetría a la derecha para la distribución del tiempo de resistencia de los vidrios. Sin embargo, se puede observar un comportamiento bimodal de la distribución empírica.

Figura 4.1: Densidad empírica del tiempo de resistencia de los vidros.



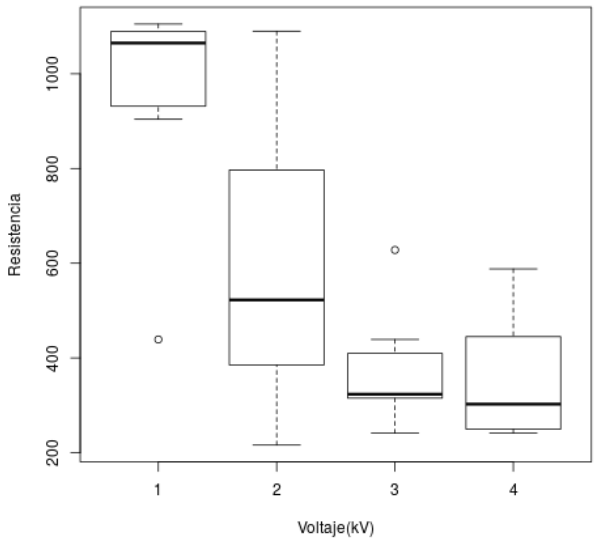
En la Tabla 4.2 se tienen algunas medidas descriptivas como: Mínimo, 1^o, 2^o e 3^o Cuartiles (Q1, Mediana y Q3, respectivamente), Media, Máximo y Coeficiente de Variación (CV). Se puede notar de la Tabla 4.2 que las muestras de vidrio presentan una mayor variabilidad para el tiempo de resistencia cuando son evaluadas en relación a la temperatura en comparación cuando es evaluada de acuerdo con el voltaje. Específicamente, en el segundo nivel de temperatura la dispersión del tiempo de resistencia de las muestras de vidrio puede ser considerada alta(63 %). Ahora, cuando el tiempo es evaluado en consideración al voltaje, es en el nivel 2 (250kV) que ocurre la mayor variabilidad.

Tabla 4.2: Medidas de resumen de la variable tiempo de resistencia del vidrio de acuerdo con los niveles de voltaje y temperatura.

	Voltaje				Temperatura	
	1 (200kV)	2 (250kV)	3 (300kV)	4 (350kV)	1 (170^oC)	2 (180^oC)
Mínimo	439	216	241	241	258	216
Q1	945,2	420	315	253,8	339	296,5
Mediana	1065	522,5	323,5	302,5	580	407,5
Media	964,5	589,4	370,6	352,9	621,5	517,2
Q3	1088	743,5	394,8	440	904	594,5
Máximo	1105	1090	628	588	1105	1087
CV	23,22	49,94	32,02	36,41	49,79	**62,97**

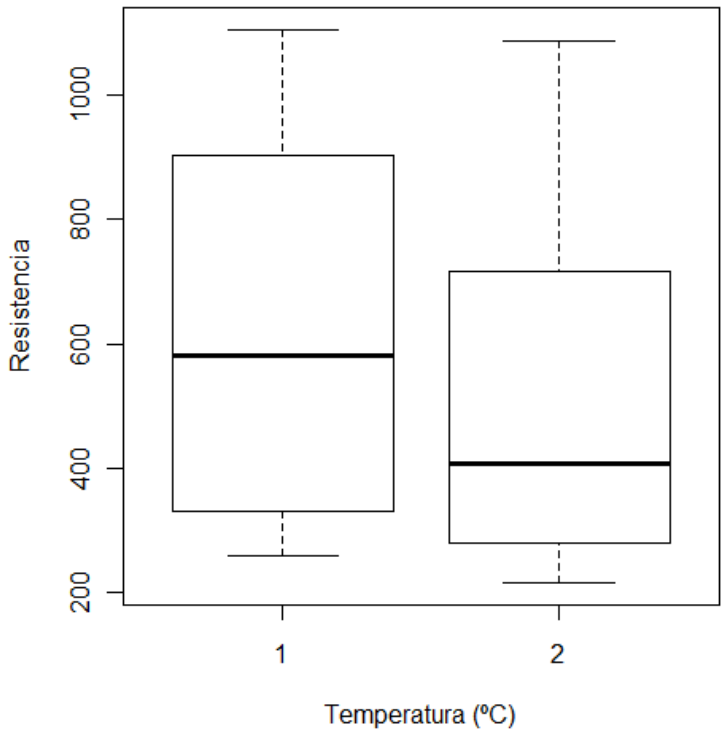
Por los boxplots de la Figura 4.2 se nota un mayor tiempo de resistencia de las muestras de vidrios cuando son evaluadas en el nivel 1 del voltaje (200kV). Además, se observa que la varianzas de los cuatro niveles del voltaje no parecen homogéneas pues en este caso, a nivel descriptivo, la variabilidad central y total son diferentes en los cuatro grupos. También puede ser observado que los niveles 1 (200kV) y 3 (300kV) del voltaje hay puntos marcados como potencialmente atípicos y que corresponden al caso #1 y #15 respectivamente.

Figura 4.2: Box-plots de la resistencia según niveles de voltaje.



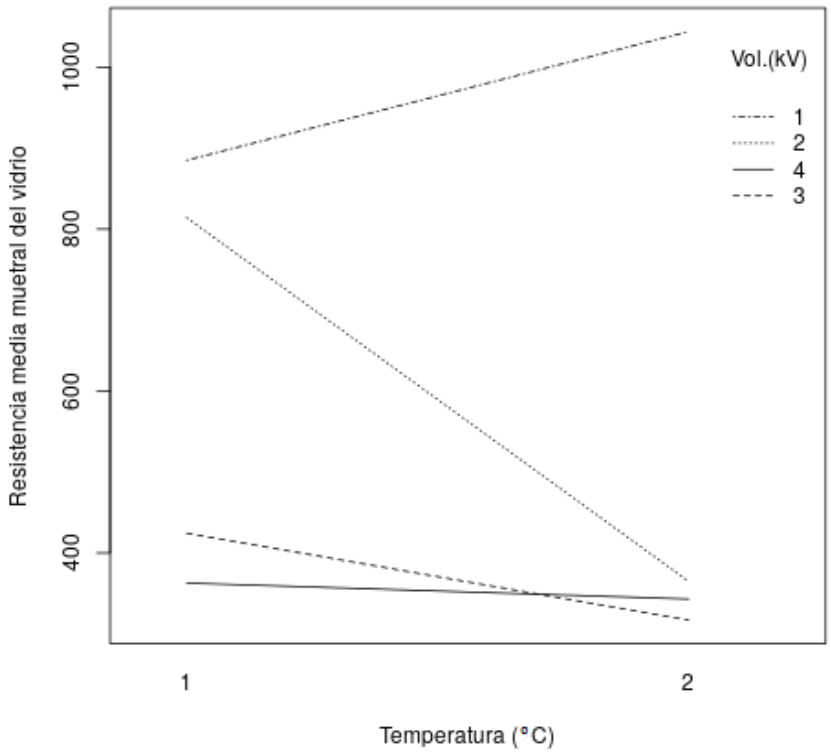
De acuerdo con los boxplots de la Figura 4.3 se nota que la tendencia central del tiempo de resistencia de las muestras de vidrios parece ser levemente mayor en el primer nivel de temperatura (170°C) con relación al segundo nivel. No son observadas grandes diferencias respecto a la variabilidad entre los dos niveles de temperatura.

Figura 4.3: Box-plots de la resistencia según niveles de temperatura.



Con el objetivo de visualizar una posible interacción entre los factores voltaje y temperatura fue construido el grafico de perfiles medios muestrales el cual está representado en la Figura 4.4. Se nota que no hay un paralelismo entre las rectas, o sea hay indicios de interacción entre los factores voltaje y temperatura. Lo anterior, quiere decir que existe indicio, a nivel descriptivo, que el coportamiento de la resistencia media del vidrio, según voltaje, no es el mismo para los dos niveles de temperatura lo que sugiere efecto de interacción entre los factores. Por ejemplo, se puede apreciar que cuando se pasa de la temperatura 1 (170 °C) para temperatura 2 (180 °C) en nivel 2 de voltaje (250 kV) ocurre una disminución en el tiempo de resistencia medio del vidro muy superior a la disminución correpondiente dentro de los niveles 3 (300 kV) y 4 (350 kV) de voltaje.

Figura 4.4: Perfiles muestrales de la resistencia media según niveles de kV y °C.



4.3. Análisis inferencial

Sea y_{ijk} el tiempo de resistencia de la k-ésima muestra de vidro sometida a la i-ésima temperatura y al j-ésimo voltaje ($k = 1, 2, 3, 4$, $i = 1, 2$, $j = 1, 2, 3, 4$). Vamos a suponer inicialmente que $y_{ijk} \sim G(\mu, \phi)$, o sea ignorar los efectos voltaje y temperatura. Se obtienen las siguientes estimaciones:

$$\widehat{\mu} = 569,34(56,03) \quad e \quad \widehat{\phi} = 3,54(0,85).$$

Por lo tanto se confirma por la estimación de ϕ (valor no tan pequeño) una cierta asimetría en la distribuición empírica del tiempo de resistencia de los vidros. Los números entre paréntesis corresponden a los errores estándar de las estimaciones

4.3.1. Ajuste de modelos.

En este caso los modelos que consideran la variable respuesta como gamma (con función enlace identidad) son ajustados para explicar el tiempo de resistencia (en horas) de los vidrios por medio de los fatores voltaje y temperatura. Como estamos trabajando solo con fatores la función de enlace no interfiere ni en las estimacones de los parámetros ni en los resultados inferenciales (Codeiro, 2000)

A continuación se realiza el ajuste de dos modelos, el primero considera los factores voljate y temperatura, mientras que el segundo modelo incluye ambos factores pero a la vez considera una interacción entre ellos. Luego se evalua la calidad del ajuste y las estimaciones de cada uno de los modelos para posteriormente ser comparados y realizar la selección de modelos.

(Modelo 1: M1) Vamos a suponer que $y_{ijk} \sim G(\mu_{ij}, \phi)$ con parte sistemática dada por

$$\mu_{ij} = \alpha + \beta_j + \gamma_i, \qquad \textbf{M1}$$

en que β_j (con $j = 1, 2, 3, 4$) y γ_i (con $i = 1, 2$) denotan, respectivamente, los efectos de voltaje y temperatura y que lo llamaremos **Modelo 1**

Fue adoptado el modelo de casilla de referencia (Salgado, Nääs, Pereira y Moura, 2007) con las siguientes restricciones:

$$\beta_1 = 0 \quad e \quad \gamma_1 = 0.$$

El ajuste de **M1** (ejecutado en paquetee estadístico R) es presentado en la Tabla 4.3.

Tabla 4.3: Resumen del ajuste del **Modelo 1**

Parámetro	Estimación	E. Estándar	Z-valor	P valor
α	1039,94	122,50	8,489	$< 0,0001$
β_2	-426,49	135,61	-3,145	0,00402
β_3	-608,81	126,21	-4,824	$< 0,0001$
β_4	-612,89	126,04	-4,863	$< 0,0001$
γ_2	-117,77	56,43	-2,087	0,04644
ϕ	9,49	2,33	4,073	$< 0,0001$

La devianza del **M1** está dado por

$$D_1^* = \widehat{\phi} D_1 = (9,49)3,43 = 32,55 \;\; (27 g.l.)$$

P-valor dado por $P = 0,21$ (no rechazamos el modelo **M1**). Además, se observa que el valor de Akaike es $AIC_1 = 428,53$.

Como todos los parámetros del **M1** son significativos, podemos tener las siguientes interpretaciones:

- La intersección indica la hora resistencia media cuando el voltaje es de 200kV y la temperatura es de 170°C;
- Para un nivel de temperatura fijo tenemos que: La disminución promedio en el tiempo de resistencia de 200kV a 250 kV es de 426,49 horas; La disminución promedio en el tiempo de resistencia de 200kV a 300kV es 608,81 horas; La disminución promedio en el tiempo de resistencia de 200kV a 350 kV es de 612,89 horas.
- Para un nivel de voltaje fijo, la disminución promedio en el tiempo de resistencia cuando pasamos de 170°C a 180°C es de 117.77 horas;
- El alto valor de la estimación ϕ confirma la asimetría correcta para la distribución del tiempo de resistencia del vidrio.

(Modelo 2: M2) Considerando ahora que $y_{ijk} \sim G(\mu_{ij}, \phi)$ con parte sistemática dada por

$$\mu_{ij} = \alpha + \beta_j + \gamma_i + \delta_{ij}, \qquad \textbf{M2}$$

en que β_j (con $j = 1, 2, 3, 4$), γ_i (con $i = 1, 2$) e δ_{ij} denotan, respectivamente, los efectos de voltaje, temperatura e interacción, llamado **Modelo 2**

El modelo de casilla de referencia se adoptó con las siguientes restricciones:

$$\beta_1 = 0, \quad \gamma_1 = 0, \quad \delta_{11} = \delta_{12} = \delta_{13} = \delta_{14} = \delta_{21} = 0.$$

El ajuste **M2** (realizado en el paquete estadístico R) se presenta en la tabla 4.4
La devianza de **M2** esta dado por

$$D_2^* = \widehat{\phi} D_2 = (13, 35)2, 43 = 32, 44 \ \ (24 g.l.)$$

P-valor dado por $P = 0{,}12$ (no rechazamos el modelo **M2**). Además, se observa que el valor de Akaike es $AIC_2 = 423.29$.

4.3.2. Selección del modelo.

Una forma de elegir un modelo apropiado es realizar la prueba de razón de probabilidad. En este caso, utilizaremos la función " **rv.gama (y, fit1, fit2)** " del paquete estadístico R. Esta función

Tabla 4.4: Resumen del ajuste del **Modelo 2**.

Parámetro	Estimación	E. Estándar	Z-valor	P valor
α	885,0	137,8	6,424	$< 0,0001$
β_2	-71,0	187,2	-0,379	0,70780
β_3	-460,8	152,8	-3,016	0,00598
β_4	-522,3	148,9	-3,508	0,00181
γ_2	159,0	213,1	0,746	0,46275
δ_{22}	-608,3	254,3	-2,392	0,02496
δ_{23}	-266,3	228,4	-1,165	0,25529
δ_{24}	-178,8	226,8	-0,788	0,43831
ϕ	13,35	3,3	4,045	$< 0,0001$

calcula el valor estadístico de la razón de probabilidad para probar dos modelos anidados gamma. Las hipótesis estadísticas para contrastar son:

$H_0 : \delta_{22} = \delta_{23} = \delta_{24} = 0$ (no hay efecto de interacción)

$$v/s$$

H_1 : Al menos uno $\delta_{2j} \neq 0, j = 2, 3, 4$.

En nuestro caso para usar la función de paquete estadístico R, "**rv.gama (y, fit1, fit2)**", dado por "y"denota la variable de respuesta (**resistencia**), "fit1.$^{\text{el}}$ modelo se ajusta bajo la hipótesis nula (**M1**) y "fit2.$^{\text{el}}$ modelo se ajusta a la hipótesis alternativa **M2**

El valor estadístico de la razón de verosimilitud es

$$RV = 11,22995 \ \ (3g.l.).$$

P-valor dado por $P = 0{,}01054527$ (rechazamos H_0). Por lo tanto, el El modelo elegido es el que tiene interacción, es decir, **M2**

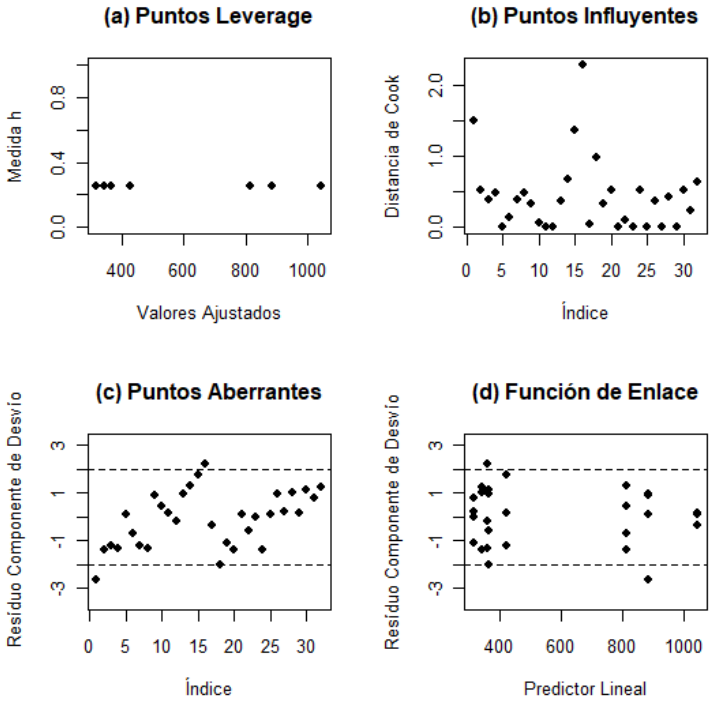
Además, el modelo elegido tiene un valor más bajo (algo deseable) de Akaike ($AIC_2 = 423{,}29$) en comparación con el modelo bajo H_0 ($AIC_1 = 428{,}53$)

4.4. Análisis de diagnóstico

En la Figura 4.5, se muestran algunos gráficos de diagnóstico con el objetivo de revelar en presencia de observaciones extremas con alguna interferencia desproporcionada Al ajustar el modo, se verificará cómo verificar los supuestos del modelo ajustado. En este caso le diremos lo siguiente:

- El gráfico 4.5 (a) muestra un valor constante para los puntos de apalancamiento en todos los casos (índices) debido a trabajar solo con factores.
- El gráfico 4.5 (b) muestra que los casos #1, #15 y #16 se destacan entre otros, siendo posibles puntos de influencia que pueden cambiar la inferencia
- El gráfico 4.5 (c) que los casos #1 y #16 están fuera del rango $(-2, 2)$ y, por lo tanto, se consideran puntos aberrantes o discrepantes.
- Se puede ver en el gráfico 4.5 (d) que la función de enlace utilizada es adecuada.

Figura 4.5: Gráficos de diagnóstico.



La siguiente es la identificación de casos comentados previamente:

1. Caso #1 (439 horas): Tiempo de resistencia de una muestra de vidrio cuando se somete a una temperatura de 170°C y un voltaje de 200kW. Es la observación que presenta la resistencia más baja dentro de esta categoría.

2. Caso #15 (628 horas): Tiempo de resistencia de una muestra de vidrio cuando se somete a una temperatura de 170°C y un voltaje de 300kW. Es la observación que presenta la mayor resistencia dentro de esta categoría.

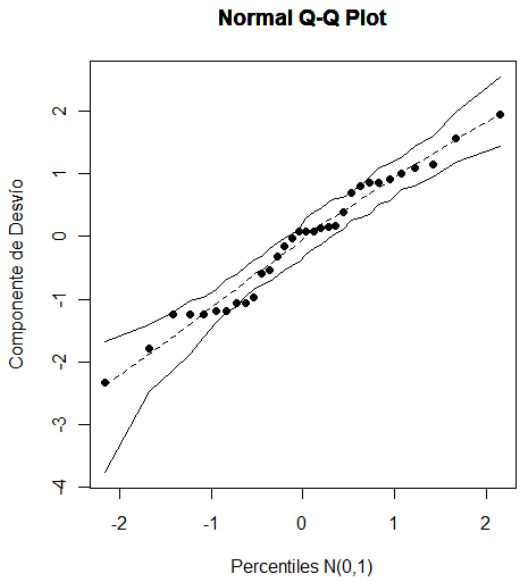
3. Caso #16 (588 horas): Tiempo de resistencia de una muestra de vidrio cuando se somete a una temperatura de 170°C y un voltaje de 350kW. Es la observación que presenta la mayor resistencia dentro de esta categoría.

Es de destacar que todos estos casos están dentro del nivel de temperatura de 170°C.

Tenga en cuenta que los casos #1 y #15 también fueron resaltados por los gráficos de los gráficos de caja (Figura 4.2). Sin embargo, el caso #16 solo se descubrió en el análisis de diagnóstico.

La Figura 4.6 muestra el gráfico de Q-Q Plot. Observamos en el gráfico que no hay evidencia de desviaciones graves en base a la suposición de que el tiempo de resistencia del vidrio tiene una distribución gamma. Sin embargo, hay ocurrencias de " barrigas " dentro de las bandas de confianza, lo que indica que puede existir heterocedasticidad.

Figura 4.6: Gráfico de Q-Q Plot.



4.5. Análisis confirmatorio

A continuación, se realiza el análisis confirmatorio para evaluar si el impacto de la eliminación de los casos descritos anteriormente, caracterizados entre influyentes y aberrantes, tiene algún grado de influencia en el modelo ajustado.

Las Tablas 4.5, 4.6 y 4.7 presentan las estimaciones del modelo ajustado y las inferencias resultantes cuando se toman los casos #1, #15 y #16 respectivamente. La Tabla 4.8 presenta las estimaciones del modelo ajustado y las inferencias resultantes cuando los tres casos se toman juntos. Se observa en cada situación que el resultado inferencial no tiene cambios (a un nivel de significancia del 10 %) en comparación con los resultados cuando se incluyen todas las observaciones (Tabla 4.4)

Tabla 4.5: Resumen del ajuste del **Modelo 2** sin el caso #1.

Parámetro	Estimación	E. Estándar	Z-valor	P valor
α	1033,67	175,07	5,904	$< 0,0001$
β_2	-219,67	211,90	-1,037	0,31068
β_3	-609,42	185,80	-3,280	0,00328
β_4	-670,92	182,97	-3,667	0,00128
γ_2	10,33	232,59	0,044	0,96495
δ_{22}	-459,58	266,86	-1,722	0,09846
δ_{23}	-117,58	245,21	-0,480	0,63610
δ_{24}	-30,08	243,84	-0,123	0,90288
ϕ	16,00	4,02	3,980	$< 0,0001$

Tabla 4.6: Resumen del ajuste del **Modelo 2** sin el caso #15.

Parámetro	Estimación	E. Estándar	Z-valor	P valor
α	885,0	131,9	6,711	$< 0,0001$
β_2	-71,0	179,2	-0,396	0,69556
β_3	-528,7	145,4	-3,635	0,00139
β_4	-522,3	142,5	-3,664	0,00129
γ_2	159,0	203,9	0,780	0,44355
δ_{22}	-608,3	243,4	-2,499	0,02005
δ_{23}	-198,3	218,1	-0,909	0,37265
δ_{24}	-178,8	217,1	-0,823	0,41872
ϕ	14,5	3,6	4,028	$< 0,0001$

Tabla 4.7: Resumen del ajuste del **Modelo 2** sin el caso #16.

Parámetro	Estimatición	E. Estándar	Z-valor	P valor
α	885,0	125,0	7,079	$< 0,0001$
β_2	-71,0	169,9	-0,418	0,679840
β_3	-460,8	138,7	-3,323	0,002960
β_4	-597,3	133,5	-4,473	0,000173
γ_2	159,0	193,3	0,822	0,419317
δ_{22}	-608,3	230,8	-2,636	0,014781
δ_{23}	-266,3	207,3	-1,284	0,211835
δ_{24}	-103,7	204,8	-0,506	0,617501
ϕ	15,6	3,9	4,000	$< 0,0001$

Tabla 4.8: Resumen del ajuste del **Modelo 2** sin el caso #1, #15 y #16.

Parámetro	Estimación	E. estándar	Z-valor	P valor
α	1033,67	142,96	7,231	$< 0,0001$
β_2	-219,67	173,04	-1,269	0,218166
β_3	-677,33	151,22	-4,479	0,000207
β_4	-746,00	148,39	-5,027	$< 0,0001$
γ_2	10,33	189,93	0,054	0,957126
δ_{22}	-459,58	217,92	-2,109	0,047124
δ_{23}	-49,67	199,86	-0,249	0,806155
δ_{24}	45,00	198,35	0,227	0,822720
ϕ	23,01	6,00	3,875	0,000107

La Tabla 4.9 presenta los valores de los *valor-P* excluyendo los casos citados en el análisis de diagnóstico. Tenga en cuenta en esta tabla que el nivel de significancia del 10 % no hubo cambio inferencial ya que los parámetros no significativos siguen siendo no significativos y, por lo tanto, los parámetros significativos siguen siendo significativos incluso cuando se eliminan los casos #1, #15 y #16 individualmente o juntos.

Tabla 4.9: *P valores* de las estimaciones de los parámetros excluyendo los casos citados.

	P valor								
Casos	α	β_2	β_3	β_4	γ_2	δ_{22}	δ_{23}	δ_{24}	ϕ
Todos	$< 0,0001$	0,70780	0,00598	0,00181	0,46275	0,02496	0,25529	0,43831	$< 0,0001$
-#1	$< 0,0001$	0,31068	0,00328	0,00128	0,96495	0,09846	0,63610	0,90288	$< 0,0001$
-#15	$< 0,0001$	0,69556	0,00139	0,00129	0,44355	0,02005	0,37265	0,41872	$< 0,0001$
-#16	$< 0,0001$	0,67984	0,00296	0,00017	0,41932	0,01478	0,21184	0,61750	$< 0,0001$
-#1#15#16	$< 0,0001$	0,21817	0,00021	$< 0,0001$	0,95713	0,04712	0,80616	0,82272	0,00011

En general, podemos concluir que el tiempo de resistencia del vidrio está influenciado por la temperatura y el voltaje, con interacciones entre ellos y, por lo tanto, se puede inferir que el tiempo de resistencia del vidrio no es el mismo para todas las combinaciones de los dos factores.

Los casos #1, #15 y #16 se destacan en los análisis realizados y cambian ligeramente las estimaciones del modelo elegido, pero no producen cambios significativos en las inferencias del modelo.

Capítulo 5

Conclusión

5.1. Consideraciones finales

En este trabajo de título se estudiaron las técnicas de diagnóstico de influencia en el modelo lineal generalizado con respuesta gamma. Uno de los principales aspectos abordados fue el desarrollo e implementación de técnicas de estimación clásicas y también el estudio de la influencia global. El Capítulo 3 ofrece varias herramientas y técnicas de diagnóstico para aplicar sobre el modelo señalado en el Capítulo 2. El capítulo 4 es la aplicación de datos reales en donde existen 3 casos que destacan en los análisis de diagnóstico de influencia realizados y cambian ligeramente las estimaciones del modelo elegido.

5.2. Sugerencias para futuras investigaciones

- Extender el modelo presentado en el Capítulo 2 a la distribución Birnbaum Saunders.

- Externder las técnicas propuestas en el Capítulo 3 incluyendo el diagnóstico de influencia local.

Capítulo 6

Referencias

1 Billor, N., y Loynes, R. M. (1993). Local influence: a new approach. Communications in Statistics-Theory and Methods, 22(6), 1595-1611.

2 Bringi, V. N., Huang, G. J., Chandrasekar, V., y Gorgucci, E. (2002). *A methodology for estimating the parameters of a gamma raindrop size distribution model from polarimetric radar data: Application to a squall-line event from the TRMM/Brazil campaign.* Journal of Atmospheric and Oceanic Technology, 19(5), 633-645.

3 Cook, R. D. (1977). *Detection of influential observation in linear regression.* Technometrics, 19 (1),15-18.

4 Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression.* Chapman and Hall, London.

5 Cook,R.D.(1986). *Assessment of local influence(with discussion).*Journal of the Royal Statistical Society, Series B 48, 133-169.

6 del Carmen García, M., Castellana, N., Rapelli, C., Koegel, L., y Catalano, M. (2014). Criterios de información y predictivos para la selección de un modelo lineal mixto. SaberEs, (6), 61-76.

7 Dennis, B., y Costantino, R. F. (1988). *Analysis of steady-state populations with the gamma abundance model: application to Tribolium.* Ecology, 69(4), 1200-1213.

8 Grover, G., Sabharwal, A. S. A., y Mittal, J. (2013). *An application of gamma generalized linear model for estimation of survival function of diabetic nephropathy patients.* International Journal of Statistics in Medical Research, 2(3), 209-219.

9 Guo, F., Wang, G., Innes, J. L., Ma, X., Sun, L., y Hu, H. (2015). *Gamma generalized linear model to investigate the effects of climate variables on the area burned by forest fire in northeast China.* Journal of forestry research, 26(3), 545-555.

10 Hawkins, D. H. (1980). *Identification of outliers.* Chapman and Hall, London.

11 Lawless, J., y Crowder, M. (2004). *Covariates and random effects in a gamma process model with application to degradation and failure.* Lifetime Data Analysis, 10(3), 213-227.

12 Lozano, C., y Arturo, B. (2018). Regresión Gamma generalizada: Extensiones y aplicaciones al análisis de datos espaciales (Doctoral dissertation, Universidad Nacional de Colombia-Sede Bogotá).

13 McCullagh, P. y Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman and Hall, London.

14 Ng, V. K., y Cribbie, R. A. (2017). *Using the Gamma Generalized Linear Model for modeling continuous, skewed and heteroscedastic outcomes in psychology.* Current Psychology, 36(2), 225-235.

15 Paula, G. A. (2004). *Modelos de regressão*: com apoio computacional (pp. 28-55). São Paulo: IME-USP.

16 Pregibon, D. (1981). Logistic regression diagnostics. The Annals of Statistics, 9(4), 705-724.

17 Rivas, L., y Galea, M. (2019). *Influence measures for the Waring regression model.* Brazilian Journal of Probability and Statistics, 33(2), 402-424.

18 Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outliers detection.* John Wiley, New York.

19 Salgado, D. D., Naas, I. D. A., Pereira, D. F., y Moura, D. J. D. (2007). *Modelos estatísticos indicadores de comportamentos associados a bem-estar térmico para matrizes pesadas.* Engenharia Agrícola, 619-629.

20 Semeraro, P. (2008). *A multivariate variance gamma model for financial applications.* International journal of theoretical and applied finance, 11(01), 1-18.

21 Williams, D. A. (1984). *Residuals in generalized linear models.* In: Proceedings of the 12th. International Biometrics Conference, Tokyo, pp. 59-68.

22 Zhu, H. y Lee, S. (2001). *Local influence for incomplete-data models.* Journal of the Royal Statistical Society, Series B, 63, 111-126.

Capítulo 7

Apéndice

Código R

- Lectura de datos
  $vidros <- scan(\text{“}C:/Users/croja/OneDrive/Escritorio/Ej\_Vidrios\_Camila/vidros\_datos.txt\text{”},$
  $list(resist = 0, voltagem = 0, tempera = 0))\ attach(vidros)$

- Gráfico densidad cruda de la variable repuesta omitiendo las covariables

  $plot1 <- plot(density(resist), xlab = \text{”}Resistencia(horas)\text{”}, ylab = \text{”}Densidad\text{”}, \text{””})$

  $savePlot(\text{”}C:/Users/croja/Downloads/tesis1/1\text{”}, type = \text{”}ps\text{”})$

  $voltaje <- factor(voltagem)$
  $voltaje <- C(voltagem, treatment)$
  $temperatura <- factor(tempera)$
  $temperatura <- C(tempera, treatment)$

- Gráficos Box plots y perfiles muestrales

  $plot2 <- boxplot(split(resist, voltaje), ylab = \text{“}Resistencia\text{”}, xlab = \text{“}Voltaje(kV)\text{”})$
  $title(\text{“”})$

  $savePlot(\text{“}C:/Users/croja/Downloads/tesis1/2\text{”}, type = \text{“}png\text{”})$

  $plot3 <- boxplot(split(resist, temperatura), ylab = \text{“}Resistencia\text{”}, xlab = \text{“}Temperatura(^{o}C)\text{”})$
  $title(\text{“”})$

  $savePlot(\text{“}C:/Users/croja/Downloads/tesis1/3\text{”}, type = \text{“}png\text{”})$

  $plot4 <- interaction.plot(temperatura, voltaje, resist, trace.label = deparse(substitute(Vol.(kV))), xlab =$
  $\text{“}Temperatura(^{o}C)\text{”}, ylab = \text{“}Resistenciamediamuestraldelvidrio\text{”})$

$savePlot(\text{“}C:/Users/croja/Downloads/tesis1/4\text{”}, type = \text{”}png\text{”})$

- Ajustar modelo ignorando los fatores
  $fit0.vidros <- glm(resist \sim 1, family = Gamma(link = identity))$
  $summary(fit0.vidros)$

- Estimación del parámetro de dispersión
  $library(MASS)$
  $gamma.shape(fit0.vidros)$

- Ajustar modelo considerando los fatores sin interacción
  $fit1.vidros <- glm(resist \sim voltaje + temperatura, family = Gamma(link = identity))$
  $summary(fit1.vidros)$

- Estimación del parámetro de dispersión
  $library(MASS)$
  $gamma.shape(fit1.vidros)$

- Ajustar modelo considerendo los fatores con interacción
  $fit2.vidros <- glm(resist \sim voltaje+temperatura+voltaje^{*}temperatura, family = Gamma(link = identity))$
  $summary(fit2.vidros)$

- Estimacion del parámetro de dispersión
  $library(MASS)$
  $gamma.shape(fit2.vidros)$

  $source(\text{“}C:/Users/croja/OneDrive/Escritorio/Ej\_Vidrios\_Camila/rv\_gama.R\text{”})$
  $rv.gama(resist, fit1.vidros, fit2.vidros)$

  $pvalor <- -1 - pchisq(11{,}22995, 3)$
  $pvalor$

  $fit.model <- -fit2.vidros$
  $attach(vidros)$
  $source(\text{“}C:/Users/croja/OneDrive/Escritorio/Ej_Vidrios_Camila/diag\_gama.R\text{”})$
  $savePlot(\text{“}C:/Users/croja/Downloads/tesis1/diag\_gama\text{”}, type = \text{“}png\text{”})$

  $fit.model <- -fit2.vidros$
  $attach(vidros)$
  $source(\text{“}C:/Users/croja/OneDrive/Escritorio/Ej\_Vidrios\_Camila/sobre\_gama.R\text{”})$
  $savePlot(\text{“}C:/Users/croja/Downloads/tesis1/sobre\_gama\text{”}, type = \text{“}png\text{”})$

Gráficos de diagnóstico

$X <- model.matrix(fit.model)$
$n <- nrow(X)$
$p <- ncol(X)$
$w <- fit.model\$weights$
$W <- diag(w)$
$H <- solve(t(X) \%* \%W \%* \%X)$
$H <- sqrt(W) \%* \%X \%* \%H \%* \%t(X) \%* \%sqrt(W)$
$h <- diag(H)$
$library(MASS)$
$fi <- gamma.shape(fit.model)\$alpha$
$ts <- resid(fit.model, type = \text{“}pearson\text{”})*sqrt(fi/(1-h))$
$td <- resid(fit.model, type = \text{“}deviance\text{”})*sqrt(fi/(1-h))$
$di <- (h/(1-h))*(ts^2)$
$par(mfrow = c(2,2))$
$a <- max(td)$
$b <- min(td)$
$plot(fitted(fit.model), h, xlab = \text{“}Valores\ Ajustados\text{”}, ylab = \text{“}Medida\ h\text{”}, pch = 16,$
$ylim = c(0,1))$
$title(sub = \text{“”})$
$title(\text{“}(a)Puntos\ de\ Leverage\text{”})$

$identify(fitted(fit.model), h, n = 1)$

$plot(di, xlab = \text{“}Indice\text{”}, ylab = \text{“}Distancia\ de\ Cook\text{”}, pch = 16)$
$title(sub = \text{“”})$
$title(\text{“}(b)Puntos\ Influyentes\text{”})$
$identify(di, n = 3)$

$plot(td, xlab = \text{“}Indice\text{”}, ylab = \text{“}Residuo\ Componente\ de\ Desvo\text{”},$
$ylim = c(b-1, a+1), pch = 16)$
$title(sub = \text{“”})$
$title(\text{“}(c)Puntos\ Aberrantes\text{”})$
$abline(2, 0, lty = 2)$
$abline(-2, 0, lty = 2)$
$identify(td, n = 2)$
$tsi, xlab = \text{“}Indice\text{”},$

$plot(predict(fit.model), td, xlab = \text{“}Predictor\ Lineal\text{”},$
$ylab = \text{“}Residuo\ Componente\ de\ Desvio\text{”}, ylim = c(b-1, a+1), pch = 16)$
$title(sub\text{“”})$
$title(\text{“}(d)Funcion\ de\ Ligacion\text{”})$
$abline(2, 0, lty = 2)$
$abline(-2, 0, lty = 2)$
$lines(smooth.spline(predict(fit.model), td, df = 2))$

$identify(predict(fit.model), td, n = 2)$

Gráfico de sobre

$phi < -(n-p)/sum((ro/(fitted(fit)))^2)$
$e[, i] < -sort(resid(fit, type = \text{“}deviance\text{”})^{*}sqrt(phi/(1-h)))$

$e1 < -numeric(n)$
$e2 < -numeric(n)$

$for(iin1 : n)\{$
$eo < -sort(e[i,])$
$e1[i] < -(eo[2] + eo[3])/2$
$e2[i] < -(eo[97] + eo[98])/2\}$

$med < -apply(e, 1, mean)$
$rango < -range(td, e1, e2)$
$par(pty = \text{“}s\text{”})$
$qqnorm\ td, xlab = \text{“}PercentilesN(0,1)\text{”},$
$ylab = \text{“}Componente\ de\ Desvio\text{”}, ylim = rango, pch = 16$
$par(new = T)$

$qqnorm(e1, axes = F, xlab = \text{“”}, ylab = \text{“”}, type = \text{“}l\text{”}, ylim = rango, lty = 1)$
$par(new = T)$
$qqnorm(e2, axes = F, xlab = \text{“”}, ylab = \text{“”}, type = \text{“}l\text{”}, ylim = rango, lty = 1)$
$par(new = T)$
$qqnorm(med, axes = F, xlab = \text{“”}, ylab = \text{“”}, type = \text{“}l\text{”}, ylim = rango, lty = 2)$
$qqnorm(med, axes = F, xlab = \text{“”}, ylab = \text{“”}, type = \text{“}l\text{”}, ylim = rango, lty = 2)$