GRADO EN ECONOMIA

CURSO ACADÉMICO: 2016 / 2017

TRABAJO FIN DE GRADO

ANÁLISIS DE RIESGO PARA LA TARIFICACIÓN DE SEGUROS DE AUTOMÓVIL MEDIANTE MODELOS LINEALES GENERALIZADOS

RISK ANALYSIS AND PRICING OF CAR INSURANCE WITH GENERALIZED LINEAR MODELS

AUTOR: DAVID VILLARINO GONZÁLEZ

DIRECTORES: VANESA JORDÁ GIL

JOSÉ MARÍA SARABIA ALEGRÍA

FECHA: 30 / JUNIO / 2017

1. RESUMEN

La finalidad de este trabajo fin de grado es la de combinar los conceptos aprendidos durante el Grado en Economía, en asignaturas como Estadística I y II, Métodos Estadísticos en Economía y Empresa, Análisis Multivariante de Datos y otras de carácter cuantitativo, con otros conceptos que acerquen al autor al estudio de las ciencias actuariales y financieras.

En este trabajo, vamos a recurrir a todos estos conocimientos para ajustar, comparar y seleccionar la mejor de tres distribuciones de tipo continuo y con grandes asimetrías de cola derecha, las cuales son utilizadas como modelizadoras del coste por siniestro en una base de datos del sector automovilístico. Por otra parte, también trataremos de ajustar un modelo lineal generalizado con la misma variable dependiente (coste del siniestro), a la que añadiremos otras variables explicativas como son el coste del vehículo, el género y el área desde el que se ha reportado el siniestro, entre otras, aportadas por nuestra base de datos, para de esta forma tratar de aportar ciertas conclusiones prácticas al estudio.

Palabras clave: Gamma, Log-normal, Pareto, AIC, BIC, VaR, TVaR, MLG, R.

2. ABSTRACT

This dissertation aims to combine all the concepts learnt in the Bacherlor´s Degree in Economics, during subjects like Statistics I and II, Statistical Methods in Economics and Business, Multivariate Data Analysis and other subjects on quantitative matters, with the concepts that would familiarize the author with the study of the actuarial and financial sciences.

During this working paper, we will make use of this knowledge to fit, to compare and to choose the best out of the three continuous right skewed distributions used in the modelling process of the claim size to a car insurance database. Furthermore, we are also going to fit a generalized linear model based on the same dependent variable (claim size), against many explanatory variables such as those about the vehicle value, the gender and the area of the claim, among others, all of them included in our database. We will fit the model to provide conclusive evidence based on a practical application.

Keywords: Gamma, Log-normal, Pareto, AIC, BIC, VaR, TVaR, GLM, R.

3. INTRODUCCIÓN

Ajustar una distribución a unos datos determinados puede parecer una tarea sencilla, no obstante, muchas son las consecuencias que acarrearía el utilizar modelos erróneos, desde malos cálculos que nos llevarían a tomar malas decisiones, desencadenando en resultados negativos como pérdidas sustanciales de tiempo y dinero, pudiendo llegar en algunas áreas incluso hasta la quiebra. Es por esto que durante este trabajo vamos a centrarnos en analizar esta poderosa herramienta, el ajuste de distribuciones, para tratar de controlar y cuantificar los riesgos financieros en el sector seguros. Además, este trabajo fin de grado surge debido al interés del autor por profundizar en la ciencia actuarial. Es por esto que hemos consultado varios trabajos fin de Máster de esta rama, en los que se tratan muchos de los conceptos analizados en el presente texto. Estos trabajos nos han servido de guía a la hora de estructurar este TFG y nos han aportado ideas sobre la realización de alguno de los apartados (Cachán 2016; Martín 2016; Plaza 2016).

La estructura de este estudio constará de cuatro partes, comenzando por esta introducción, tras la que daremos paso a la metodología, el análisis de resultados y finalmente las conclusiones.

En la metodología vamos a empezar definiendo tres distribuciones modelizadoras de la severidad siniestral, la gamma, log-normal y Pareto, y detallaremos cómo estimar sus parámetros por máxima verosimilitud. Tras esto, marcaremos una serie de pautas a seguir para seleccionar la que mejor se ajuste a unos datos, mediante comparación gráfica de QQ plot y PP plot, de los criterios de bondad de ajuste de Akaike y bayesiano, y de los estadísticos de bondad de ajuste Kolmogorov-Smirnov, Anderson Darlin y Cramér-von Mises. Además, propondremos la utilización de las medidas de riesgo VaR y TVaR como otro medio más de comparación de distribuciones.
Tras esta primera parte, pasaremos a definir la base de los modelos lineales generalizados, sus componentes y criterios de bondad de ajuste.

En la tercera parte, vamos a utilizar una base de datos de pólizas de seguro en la que encontraremos la variable dependiente coste siniestral y una serie de variables explicativas. Tras una exploración previa de los datos, pasaremos a utilizar los métodos estadísticos enumerados anteriormente para ajustar y seleccionar la mejor de las tres distribuciones. Analizaremos los gráficos y los valores de los estadísticos generados mediante el programa R.

Esta base de datos proviene de Jong y Heller (2008). En su libro Generalized Linear Models for Insurance Data, detallan las innumerables herramientas de uso actuarial para el análisis de bases de datos sobre seguros. En este libro hemos contrastado la mayor parte de la teoría que aquí utilizaremos sobre las distintas distribuciones modelizadoras de la severidad siniestral, y de los modelos lineales generalizados de posteriores apartados. Además, Jong y Heller dedican los últimos capítulos del libro a la realización de ejercicios prácticos. Entre ellos, proponen modelizar un GLM mediante la distribución gamma con link log-normal sobre unos datos de una aseguradora. En este trabajo tomaremos como base ese modelo, trataremos de darle forma, lo analizaremos y aportaremos una solución al mal ajuste que presenta.

Otros trabajos (Sarabia et al. 2016; Heller et at. 2007) proponen modelos combinados de severidad y frecuencia siniestral para la modelización de estos mismos datos, Pareto-Poisson en el primer caso e Inversa Gaussiana-Binomial Negativa en el último. Estos modelos mixtos avanzados se salen del objetivo de este trabajo y solo serán analizados

modelos de tipo continuo. El documento de Sarabia et al. ha servido también como precedente en la realización de este trabajo, a partir del cual ha sido propuesto el estudio de las medidas de riesgo que allí se utilizan y el caso práctico sobre la misma base de datos.

4. METODOLOGÍA

En este capítulo vamos a presentar la metodología que será utilizada a lo largo del estudio. Así pues, lo dividiremos en dos partes, donde se explicará en detalle el ajuste de las distribuciones modelizadoras de la severidad siniestral y las bases del modelo lineal generalizado (GLM).

4.1. DISTRIBUCIONES MODELIZADORAS DE LA SEVERIDAD SINIESTRAL

En este apartado desarrollamos la metodología que será utilizada para la modelización de las distribuciones de severidad siniestral. De este modo, la estructura se basará en desarrollar cada uno de los pasos que se darán en el posterior estudio. Selección de la distribución, estimación de los parámetros por máxima verosimilitud y elección de la distribución a ajustar en base a criterios gráficos, de bondad de ajuste y medidas de riesgo.

4.1.1 Selección de la distribución

Como paso previo en el análisis de este apartado, vamos a comenzar por la selección de la distribución a ajustar. Debido a que no tenemos identificado el modelo acorde a nuestros datos, vamos a proponer el estudio de tres de las distribuciones más utilizadas para modelizar la severidad siniestral, que tiene como particularidad una gran asimetría de cola derecha. Es por tanto importante recalcar, que aunque la elección de estas tres distribuciones pueda parecerle subjetiva y aleatoria al lector, nos hemos basado en dos aspectos fundamentales para dicha elección. Por un lado, que puedan ser ajustadas por el programa informático del que disponemos para este trabajo (R). Por otro lado, hemos querido seleccionar tres distribuciones con características distintas para que cada una de ellas pueda aportar un punto de vista en la región crítica de nuestro análisis, esto es, que la longitud y peso de sus colas nos aporten diferentes informaciones dependiendo del coste del siniestro. La distribución gamma que es la de colas más cortas, seguida por la log-normal y finalmente la distribución de Pareto que es la más frecuentemente utilizada para modelizar datos con valores catastróficos.

4.1.1.1 Distribución gamma

La distribución gamma es en la estadística actuarial una de las más importantes cuando se trata de ajustar un conjunto de datos positivos, unimodales y con asimetría positiva (Sarabia, Gómez y Vázquez 2007). Entre las tres distribuciones seleccionadas, es la de colas más cortas por lo que será interesante utilizarla en casos de ajustar datos con valores no extremos.

Una variable aleatoria *X* sigue una distribución gamma con parámetros *α* y *σ*, y será representada por *X ~ G (α, σ)*, si su función de densidad viene dada por:

$$f(x) \begin{cases} \dfrac{x^{\alpha-1}e^{-\frac{x}{\sigma}}}{\sigma^{\alpha}} & si\ x \geq 0\,, \\ 0 & si\ x < 0\,. \end{cases}$$

Siendo sus momentos:

$$E(X^r) = \frac{\Gamma(r+\alpha)\sigma^r}{\Gamma(\alpha)}.$$

De donde su media y varianza:

$$E(X) = \alpha\sigma\,;$$

$$Var(X) = \alpha\sigma^2\,.$$

4.1.1.2 Distribución log-normal

La distribución log-normal por su parte, modelizará mejor aquellos datos con colas más largas, por lo que es una buena candidata de cara a complementar las carencias de la anterior. Así pues, en este caso una variable aleatoria *X* seguirá una distribución log-normal con parámetros μ y σ^2, y será representada por $X\sim LN(\mu, \sigma^2)$, si su función de densidad viene dada por:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{logx-\mu}{\sigma}\right)^2\right\}, \qquad x > 0\,.$$

Siendo sus momentos:

$$E\left(X^k\right) = \exp\left(k\mu + \frac{k^2\sigma^2}{2}\right)\ , k = 1, 2, \ldots$$

De donde su media y varianza:

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right);$$

$$Var(X) = \exp(2\mu+\sigma^2)\,(\exp(\sigma^2)-1)\,.$$

4.1.1.3 Distribución de Pareto

Finalmente, la tercera distribución seleccionada es la de Pareto, que es útil en el caso de querer ajustar costes siniestrales muy elevados, ya que converge a cero de una forma mucho más lenta que las anteriores.

Una variable dependiente *X* sigue una distribución de Pareto de tipo II con parámetros σ y α, y será representada por $X\sim Pa(\sigma, \alpha)$, si su función de densidad es la siguiente:

$$f(x) = \frac{\alpha/\sigma}{(1+\frac{x}{\sigma})^{\alpha+1}}.$$

Siendo sus momentos:

$$E(X^r) = \frac{\Gamma(\alpha-r)\Gamma(r+1)\sigma^r}{\Gamma(\alpha)}, \qquad \alpha > r\,.$$

De donde su media y varianza:

$$E(X) = \frac{\sigma}{\alpha - 1}, \alpha > 1 ;$$

$$Var(X) = \frac{\alpha\sigma^2}{(\alpha - 1)^2(\alpha - 2)}, \alpha > 2 .$$

4.1.2. Estimación de parámetros por máxima verosimilitud

La estimación por momentos es la más sencilla para un caso general en el que quisiéramos ajustar, como ejemplo, una distribución normal, ya que solo necesitaríamos la media y varianza muestrales para estimar las poblaciones. Sin embargo, esta simplificación nos hace perder una parte importante de la información necesaria en el análisis actuarial[1], esto es, las colas pesadas de los datos. Debido a la importancia de tener en cuenta estos valores extremos, necesitamos utilizar un método de estimación que haga máxima la probabilidad de la muestra observada (Sarabia, Gómez y Vázquez 2007).

El método de máxima verosimilitud cumple además, que genera estimadores insesgados y consistentes, cumpliendo la cota de Crámer-Rao (Crámer 1946), y estos son asintóticamente normales e invariantes (Klugman et al. 2012).

Definimos la función de verosimilitud a partir de unos datos con *n* variables aleatorias e independientes entre sí, que van desde $x_1 \ldots x_n$, que forman parte de una misma distribución y que dependen de un mismo vector de parámetros θ. Su forma viene dada por:

$$L(\theta) = \prod_{j=1}^{n} f(x_i|\theta) .$$

A partir de aquí, encontraremos el estimador de máxima verosimilitud $\hat{\theta}$ donde se maximice esta función[2], esto es:

$$l(\theta) = \sum_{i=n}^{n} \log f(x_i|\theta) .$$

Y que para ser máximo cumplirá:

$$\frac{\partial l(\theta)}{\partial \theta_i} = 0, i = 1, \ldots, k .$$

[1] Otro problema muy común es la existencia de varios deducibles para grupos de individuos, lo que nos daría problemas a la hora de estimar sus medias. Klugman (Klugman et al. 2012) propone la estimación de los momentos a través del estimador Kaplan-Meier para solventarlo.

[2] En nuestro caso, la estimación de los parámetros de las tres distribuciones seleccionadas fue llevada a cabo a través del paquete *fitdistrplus* de R no siendo necesaria la computación manual, no obstante, nos ha parecido importante detallar la base teórica.

4.1.3. Elección de la distribución a ajustar

Como tercer paso, partiremos de la afirmación del estadístico George Box, quien decía que todos los modelos son erróneos de partida, ya que son aproximaciones de la realidad, pero algunos de ellos podrán ayudarnos a inferir resultados, por lo que nuestro objetivo será encontrar el modelo que mejor ajuste nuestros datos.

"…essentially, all models are wrong, but some are useful." (Box y Draper 1987, p. 424)

Tras haber estimado los parámetros por máxima verosimilitud, ajustaremos las tres distribuciones y seleccionaremos aquella que se comporte mejor de acuerdo a la comparación gráfica por un lado, y la comparación de estadísticos y de criterios de bondad de ajuste por el otro, tal y como explicaremos a continuación.

4.1.3.1. Comparación gráfica QQ-plot

El *QQ plot* o *quantile-quantile plot* es una herramienta gráfica que nos ayuda a señalar de una forma muy intuitiva si nuestros datos se pueden ajustar mediante cada una de nuestras distribuciones. El gráfico cuanti-cuantil representa los cuantiles teóricos contra los empíricos. En el caso de que nuestros datos se ajustasen perfectamente a la distribución seleccionada, veríamos una línea recta de 45º, en todos los demás casos, podremos ver asimetrías en la distribución de la muestra, si las colas son más o menos pesadas, e incluso valores atípicos.

Esta herramienta no será capaz de confirmar qué distribución siguen nuestros datos, pero sí nos dará pistas sobre cuál de las tres se aproxima más a la forma de estos. Puede parecer por tanto una herramienta subjetiva, pero su uso es muy fácil e intuitivo, motivo por el cual lo introducimos como primera aproximación en nuestro análisis.

4.1.3.2. Comparación gráfica PP-plot

A diferencia de la herramienta gráfica del caso anterior, el gráfico probabilidad-probabilidad compara las distribuciones empíricas y teóricas. Una ventaja de utilizar esta técnica al lado del *QQ plot* es que genera mejores interpretaciones en la zona donde se dan mayores densidades de probabilidad, ya que en es en estas zonas donde se pueden apreciar mayores variaciones de probabilidad en comparación con las colas donde la densidad es menor (Gnanadesikan 1997).

4.1.3.3 Criterios de información de Akaike y bayesiano

Debido a la necesidad de generar un criterio estadístico al que acudir para contrastar qué modelo se ajusta más a unos datos concretos, primero Akaike y después Schwarz propusieron los criterios de información de Akaike (Akaike 1973) y bayesiano (Schwarz 1978) respectivamente, cuyas expresiones son las siguientes:

$$AIC = -2\ln(\hat{\theta}) + 2K \quad ; \quad BIC = -2\ln(\hat{\theta}) + K\,ln(n)\,.$$

Donde *K* es el número de parámetros del modelo y $\hat{\theta}$ es de nuevo el estimador de máxima verosimilitud.

Menores estadísticos determinarán un mejor modelo y podemos apreciar cómo en ambos casos se introduce el número de parámetros del modelo como penalización. Esta penalización es superior en el caso del criterio bayesiano, por lo que sería interesante

en una comparación de modelos con heterogeneidad en el número de parámetros, sin embargo, en nuestro caso las interpretaciones van a ser las mismas para los dos estadísticos al tener las tres distribuciones el mismo número. Estos criterios serán también estudiados en la implementación del modelo lineal generalizado que detallaremos en próximos apartados.

4.1.3.4. Estadísticos de bondad de ajuste

Siguiendo con la tónica de este apartado, en este caso vamos a presentar los tres estadísticos de bondad de ajuste más utilizados: Kolmogorov-Smirnov (Massey 1951), Cramér-Von Mises (Cramér 1928) y Anderson-Darlin (Anderson y Darlin 1952).

En los tres casos, su uso habitual se daría contrastando el par de hipótesis tradicional, llevándonos su rechazo a la alternativa, esto es, que los datos no sigan la distribución de estudio, y la hipótesis nula nos llevaría a la aceptación de la distribución como una candidata al ajuste de estos. Sin embargo, en este trabajo, como ya hemos comentado anteriormente, partimos de que las tres distribuciones seleccionadas son candidatas a ajustar nuestros datos y utilizaremos estos estadísticos a modo de comparación, por lo que nos basaremos en el criterio de que, a menor estadístico, mejor ajustará la distribución, como se pasará a explicar a continuación.
Señalar que la principal crítica a KS y CVM es que no consiguen seleccionar correctamente la mejor distribución para tamaños muestrales pequeños en cuyo caso la mejor alternativa es el test AD. Pero para muestras grandes, como es nuestro caso, sí tienen una buena actuación (Flowers-Cano, Flowers y Rivera-Tejo 2014). Una vez aceptada la validez de los tres criterios, veremos que cada uno de ellos tiene un mayor poder estadístico para diferentes situaciones.

4.1.3.4.1. Kolmogorov-Smirnov y modificación de Lilliefors

El estadístico KS está basado en la comparación de las distancias verticales máximas entre la distribución teórica ($F^*(x)$) y empírica ($F_n(x)$). Es por esto, que podemos prever que este test sea de mayor utilidad sobre la zona central de la distribución, que sobre las colas, ya que las diferencias serán mayores en esta primera región.

La definición general de este estadístico es la siguiente[3]:

$$T = sup_x |F^*(x) - F_n(x)| \,.$$

Donde “sup” (supremum) significa mayor.

Dieciséis años después de la publicación de este estadístico, Lilliefors (1969) presentaría su modificación como alternativa a KS en los casos en que no se conozca la verdadera distribución (exponencial[4]) que sigue la muestra de datos a estudiar. En este caso, los parámetros serán estimados directamente a través de esta. Además, Lilliefors aportaría una tabla de valores críticos distinta a la de KS, por lo que mismos valores arrojados por los dos estadísticos nos aportarían conclusiones distintas. Sin embargo, este test será utilizado como una herramienta de comparación entre datos (como se ha comentado anteriormente), por lo que en este trabajo hablar de KS y de LF

[3] En la literatura encontramos multitud de definiciones para este estadístico. En nuestro caso vamos a utilizar la definición de Conover (1999), a través de Razali y Wah (2011).
[4] En 1967 Lilliefors publicaría su modificación al test de normalidad de KS (Lilliefors 1967). Dos años más tarde extendería esta modificación para también contrastar los casos exponenciales.

será equivalente al calcularse de la misma manera, como se puede ver en la siguiente definición del estadístico Lilliefors:

$$D = máx_x |F^*(x) - S_n(x)| .$$

Donde ahora $S_n(x)$ hace referencia a la distribución empírica, estimada con los parámetros de la muestra.

4.1.3.4.2. Anderson-Darlin

El test estadístico AD computa los cuadrados de las diferencias para cada uno de los puntos, a diferencia de KS que solo computaba la diferencia máxima. Es por esto que se dice que AD da un mayor peso a las colas de la distribución que el test KS. Además, pondera por un parámetro ψ que definimos a continuación

$$\psi = \left[F^*(x)\left(1 - F^*(x)\right)\right]^{-1}.$$

De este modo, el estadístico es definido como:

$$W_n^2 = nn \int_{-\infty}^{\infty} [F^*(x) - F_n(x)]^2 \psi[F(x)] dF(x) .$$

4.1.3.4.3. Cramér-von Mises

Este estadístico está recogido como un caso especial del anterior cuando $\psi = 1$.

$$CVM = n \int_{-\infty}^{\infty} [F^*(x) - F_n(x)]^2 [F(x)] dF(x) .$$

O equivalentemente:

$$\frac{1}{12n} + \sum_{i=1}^{n} \left[F_0(x_i) - \frac{2i-1}{2n}\right]^2 .$$

4.1.3.5. Medidas de riesgo

Hemos decidido incluir este apartado como una herramienta más de decisión, ya que una de las bases del análisis actuarial es la evaluación de riesgos, por lo que es de suma importancia tener en cuenta el comportamiento de las distribuciones en la zona que define los mayores riesgos, sus colas.

Utilizaremos por un lado el valor en riesgo o VaR, y dado que esta no es una medida de riesgo "coherente" (cuyo significado explicaremos en detalle en este apartado), utilizaremos también el valor de la cola en riesgo o TVaR.

4.1.3.5.1. Valor en Riesgo (VaR)

También llamado *value at risk*, el VaR es como toda medida de riesgo, una herramienta de determinación de la exposición al riesgo, siniestral en nuestro caso, pero también de un gran uso en el caso financiero. Es así pues, un cuantil que denota la pérdida máxima que una empresa aseguradora podría soportar a un cierto nivel de confianza dado.

Sea *X* una variable aleatoria independiente, el VaR de *X* a un nivel *p%*, que podemos definir como $Var_p(X)$ o π_p, y representa el *p* cuantil de la distribución de *X* (Klugman, Panjer y Willmot 2012). Donde π_p para una variable continua satisface:

$$\Pr(X > \pi) = 1 - p\,.$$

El VaR es definido como:

$$Var(X) = \inf\{x \in \mathbb{R}, F(x) \geq p\}\,.$$

Sin embargo, como avanzábamos en la introducción de este apartado, el VaR no es una medida de riesgo coherente (Artzner et al. 1999) debido a que falla en el cumplimiento de la propiedad de subaditividad. Desarrollamos a continuación los cuatro supuestos que toda medida de riesgo coherente ha de cumplir.

- **Monótona:** Si $X \leq Y$ entonces, $\rho(X) \leq \rho(Y)$, donde $\rho(X)$ y $\rho(Y)$ es como denotaremos a la medida de riesgo objetivo.
- **Positiva homogénea:** Para cualquier constante positiva *c*, $\rho(Xc) = c\rho(X)$
- **Invariante a la traslación:** Para cualquier *c* positiva, $\rho(X + c) = c + \rho(X)$
- **Subaditividad:** $\rho(X + Y) \leq \rho(X) + \rho(Y)$

El supuesto de subaditividad nos señala que la diversificación ayudaría a reducir el riesgo. Sin embargo, esta última condición no se da para el VaR en la mayoría de los casos[5], por lo que utilizaremos el TVaR que sí es coherente como explicaremos a continuación.

4.1.3.5.2. Valor de la cola en riesgo (TVaR)

El valor de la cola en riesgo, es la pérdida esperada dado que esta supere el cuantil del valor en riesgo, que habíamos definido como la pérdida máxima soportada a un nivel de confianza dado. Esto no es otra cosa que la esperanza matemática en la cola, lo cual ya nos está dando información sobre una región completa de riesgo, más que el VaR que solo hace referencia a un punto. Además, como hemos adelantado, el TVaR sí es una medida de riesgo coherente, al cumplir también la subaditividad (Wirch 1999).

Matemáticamente el TVaR del cuantil *p* es definido como:

$$TVaR_p(X) = E_p[X|X \geq VaR_p]\,.$$

Como podemos comprobar en la fórmula, el TVaR siempre será mayor al VaR al computar la media de los riesgos superiores para ese percentil. La diferencia entre ambas medidas puede ser contemplada en el siguiente gráfico (figura 4.1) donde se aprecia de una forma más intuitiva lo que representa cada una de ellas. El nivel de confianza $1 - \beta$ hace referencia a la probabilidad de que *X* sea mayor que el VaR *(1-p).*

[5] Wirch (1999) muestra como el VaR no es una medida coherente para distribuciones que no sean la normal (como es nuestro caso) mediante un ejemplo en el que tras combinar dos activos financieros en uno nuevo, el riesgo del nuevo portafolio es superior al de la suma de los dos anteriores. Para solucionar este problema, propone el uso del TVaR.

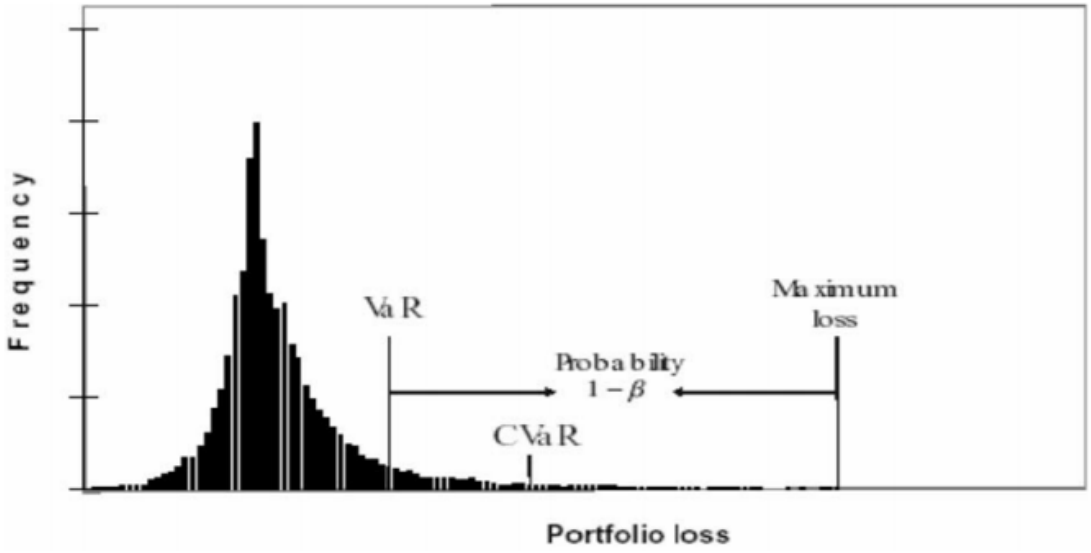


Figura 4.1: VaR y TVaR.

Fuente: Szegö 2002.

4.2. MODELO LINEAL GENERALIZADO

El modelo lineal generalizado, también MLG o GLM (Nelder y Wedderburn 1972), es como su propio nombre indica, una generalización del modelo lineal simple. Es por esto, que conviene dedicar unas líneas a este último para entender el porqué del uso de los MLG.

El modelo de regresión simple, se basa en los siguientes tres supuestos fundamentales:

- **Linealidad en los parámetros.** Su forma es la siguiente:

$$y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \, .$$

- **Término de error distribuido de forma normal.**

$$\varepsilon_i \sim N(0, \sigma^2) \quad ; \quad E[\varepsilon_i] = 0 \quad ; \quad Var[\varepsilon_i] = \sigma^2 \, .$$

- Esto implica que **la varianza del error es homocedástica.**

El principal problema del modelo de regresión simple es que, a pesar de su amplio uso en la literatura, estos supuestos son a menudo infringidos. Por ejemplo, en el caso de la varianza, esta suele aumentar con la media ya que esperamos que aquellas pólizas con costes siniestrales más altos, tengan también mayores varianzas en el número de siniestros (Goldburd, Khare y Tevet 2016). El no cumplimiento de todos estos supuestos se puede solucionar mediante una transformación logarítmica de los datos, pero en muchos casos puede no ser suficiente.

Es aquí donde aparece el modelo lineal generalizado a partir de la necesidad de aportar un modelo que solucione la no linealidad en los parámetros y la no normalidad y heterocedasticidad del error. El MLG tratará de modelizar la relación entre la variable respuesta y el resto de variables explicativas, al igual que hacía el modelo anterior.

4.2.1. Componentes del MLG

La predicción de la variable respuesta en un MLG se dará gracias a la influencia de una componente sistemática y una componente aleatoria. La componente sistemática hace referencia a la influencia de las variables del modelo sobre la variable dependiente y la componente aleatoria hace referencia a todo aquello que nuestras variables no explican.

4.2.1.1. Componente aleatoria

En el modelo lineal generalizado, la variable respuesta es una serie de datos independientes que siguen una distribución de la familia exponencial (gamma, log-normal, Pareto…). Esta trata de modelizar de una forma no-normal la distribución del término de error de nuestro modelo.

Se expresa:

$$y_i \sim Familia\ exponencial(\mu_i, \emptyset)\ .$$

Donde μ es la media de la distribución y $\emptyset$ el parámetro de dispersión que está relacionado con la varianza, y a su vez esta está relacionada con la media, tal que:

$$Var(y) = \emptyset V(\mu)\ .$$

En el siguiente gráfico (figura 4.2) podemos ver cómo ahora la varianza ya no es constante para todo *X* y depende de la media determinada por la distribución seleccionada.

Distribución	$\boldsymbol{V(\mu)}$	$\boldsymbol{\emptyset V(\mu)}$
Normal	1	$\emptyset$
Poisson	μ	$\emptyset\mu$
Gamma	μ^2	$\emptyset\mu^2$
Inversa gaussiana	μ^3	$\emptyset\mu^3$

Figura 4.2: Función de la varianza para distribuciones exponenciales.

Fuente: Goldburd, Khare y Tevet 2016.

4.2.1.2. Componente sistemática

La componente sistemática de un MLG es una combinación lineal de las variables explicativas, tal que:

$$\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\ .$$

También llamado predictor lineal y expresado de la siguiente manera:

$$\eta_i = \sum_j \beta_j x_{ij}\ .$$

Donde x_{ij} es el valor *j-ésimo* predictor en el *i-ésimo* individuo, e $i = 1, \ldots, N$. Los valores de los predictores serán generados mediante el programa R.

4.2.1.3. Función Link

La función link es la transformación de η_i representada como $g(\mu_i)$, donde:

$$\mu_i = E(Y_i)\ ;\ \ \eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}\ .$$

En la siguiente tabla (figura 4.3) podemos ver las diferentes funciones link, donde por ejemplo la identidad daría paso a la regresión lineal simple.

Función Link	$\boldsymbol{\eta_i = g(\mu_i)}$	$\boldsymbol{\mu_i = g^{-1}(\eta_i)}$
Identidad	μ_i	η_i
Log	$log_e \mu_i$	$e^{-\eta_i}$
Inversa	μ_i^{-1}	η_i^{-1}
Inversa cuadrado	μ_i^{-2}	$\eta_i^{-\frac{1}{2}}$
Logit	$log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\phi^{-1}(\mu_i)$	$\phi\eta_i$

Figura 4.3: Funciones link más comunes.

Fuente: Elaboración propia.

Como podemos ver, tras aplicar una función link a nuestros datos, nos quedaría el valor de $g(\mu_i)$, el cual carece de interés. Sí nos interesaría por el contrario calcular la inversa de la transformación para obtener la predicción. A continuación, vemos el ejemplo del caso de un link logarítmico:

$$ln\mu = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\ .$$

El valor del $ln\mu$ no es de interés como comentábamos anteriormente, pero si realizamos su inversa al predictor lineal, obtendremos:

$$\mu_i = \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}) = e^{\beta_0} e^{\beta_1 X_{1i}} \ldots e^{\beta_k X_{ki}}\ .$$

Donde tenemos la predicción para cada valor. Además, esta forma de μ_i es muy interesante dentro de los MLG ya que tiene forma de modelo multiplicativo, muy utilizada en modelos de *insurance pricing* o primas de seguros, aunque no haremos mayor referencia a estos modelos en el presente documento.

4.2.2. Criterios de bondad de ajuste

Además de los criterios AIC y BIC anteriormente comentados, en la comparación de modelos GLM vamos a utilizar la "deviance". La deviance se basa en la comparación del modelo Al igual que los criterios de AIC y BIC, cuando el modelo mejora, el valor arrojado será más pequeño, tal y como podemos extraer de su forma:

$$D = -2\left[ln\left(p(y|\hat{\theta}_0)\right) - \ln\left(p(y|\hat{\theta}_S)\right)\right].$$

Este estadístico compara el modelo "nulo" con solo el intercepto, con el saturado, es decir, el modelo con variables explicativas.

4.2.3. PROCESAMIENTO DE LOS DATOS

En siguientes apartados, más concretamente en el que analizamos la base de datos y las variables que se incluyen en él y utilizaremos para el estudio práctico, veremos que el número de variables del que disponemos es muy reducido. Si bien esto plantea cierta dificultad debido a que nuestros modelos no tendrán una buena capacidad predictiva, sí que nos facilita el ejercicio de la selección de variables. Al no haber muchas combinaciones posibles, hemos podido contrastar de forma manual si nuestro modelo mejoraba o empeoraba, incluyendo o excluyendo aquellas variables que por no ser significativas o por no disminuir los AIC y deviance, no aportaban ninguna información.

No obstante, sí queremos mencionar algunos de los métodos más utilizados para la selección de variables, que podrán ser utilizados en futuros trabajos con mayores bases de datos y mayores números de variables explicativas.

La regresión stepwise es una de las técnicas más utilizadas en la literatura[6]. Está basada en un algoritmo en el que se van añadiendo (stepwise forward) o quitando (backward elimination) variables una a una hasta que finalmente se encuentra el mejor modelo. Sin embargo, esta es una técnica que está perdiendo valor en los últimos años ya que es una técnica *ad hoc* y nos puede llevar a distintos resultados (Jong y Heller 2008).

Otras de las técnicas más usuales son *ridge regression, lasso* y la generalización de ambas*, las eslastic nets*. Todas ellas son utilizadas para seleccionar variables cuando el número de estas es muy elevado y la introducción de todas ellas puede causar sobredispersión. Precisamente, el escaso número de variables en nuestro caso ha sido determinante para no utilizar ninguna de estas técnicas.

5. ANÁLISIS DE LOS DATOS Y RESULTADOS

Nuestro objetivo en este apartado es aplicar todos los conceptos comentados anteriormente, sobre un caso práctico para el coste agregado por siniestros reclamados a una compañía de seguros del automóvil australiana. Para ello disponemos de una base de datos (Jong y Heller 2008) con 67.856 pólizas de seguros para el año 2004-2005, donde encontramos nuestra variable dependiente (coste del siniestro), sobre la que basaremos la primera parte de este capítulo, y el resto de variables explicativas que utilizaremos a la hora de modelizar el GLM.

5.1. BASE DE DATOS Y DEFINICIÓN DE VARIABLES

En un análisis exploratorio previo, podemos ver (figura 5.1) la forma que toma nuestra base de datos. Explicamos a continuación el significado de cada una de las variables:

- **Veh_value:** Valor del vehículo en miles de dólares australianos.
- **Exposure:** Exposición al riesgo de la póliza durante el año. Este valor se encuentra acotado entre 0 y 1, donde el valor mínimo (0.00274) correspondería a una póliza que ha permanecido dada de alta un día y 1 haría referencia a una exposición del año completo.
- **Clm:** Variable binaria, que nos dice si la póliza ha reclamado (1) o no (0) un siniestro.
- **Numclaims:** Número de claims entre 0 y 4.

[6] Como ejemplo, en este paper (véase Whittingham et al. 2006) contrastan que estas técnicas fueron utilizadas en el 57% de los papers publicados en los que se utilizaba una regresión múltiple.

- **Claimcst0:** Coste total de la póliza por siniestros (dólares australianos).
- **Veh_body:** Los doce tipos de vehículo son: descapotable, deportivo, vehículo de 3 y 5 puertas, carrocería descapotable, caravana, minibús, furgoneta, deportivos de dos plazas, turismos, familiares, camión y coche de servicios.
- **Veh_age:** Antigüedad del vehículo siendo 1 el valor para el más nuevo.
- **Gender:** Género del propietario de la póliza.
- **Area:** Seis áreas en las que se divide Australia (A-F), formada por New South Wales, Queensland, South Australia, Tasmania, Victoria y Western Australia. No hay informaci a qué estado corresponde cada una de las letras.
- **Agecat:** Seis categorías de edad del propietario (donde 1 es el grupo más joven). No conocemos el rango de edad al que corresponde cada categoría.

	Veh_value	Exposure	Clm	Numclaims	Claimcst0	Veh_body	Veh_age	Gender	Area	Agecat
1.	1.060	0.3039	0	0	0.000	HBACK	3	female	C	2
2.	1.030	0.6489	0	0	0.000	HBACK	2	female	A	4
3.	3.260	0.5694	0	0	0.000	UTE	2	female	E	2
4.	4.140	0.3175	0	0	0.000	STNWG	2	female	E	2
5.	0.720	0.6488	0	0	0.000	HBACK	4	female	C	2
6.	2.010	0.8542	0	0	0.000	HDTOP	3	male	C	4
7.	1.600	0.8542	0	0	0.000	OANVN	3	male	A	4
8.	1.470	0.5557	0	0	0.000	HBACK	2	male	B	6
9.	0.520	0.3613	0	0	0.000	HBACK	4	female	A	3
10.	0.380	0.5202	0	0	0.000	HBACK	4	female	B	4
11.	1.380	0.8542	0	0	0.000	HBACK	2	male	A	2
12.	1.220	0.8542	0	0	0.000	HBACK	3	male	C	4
13.	1.000	0.4928	0	0	0.000	HBACK	2	female	C	4
14.	1.660	0.4845	0	0	0.000	STNWG	1	male	A	5
15.	1.660	0.4845	1	1	669.51	SEDAN	3	male	B	6

Figura 5.1: Primeras 15 pólizas de la base de datos.

Fuente: Jong y Heller 2008.

5.1.2. Selección de la frecuencia siniestral

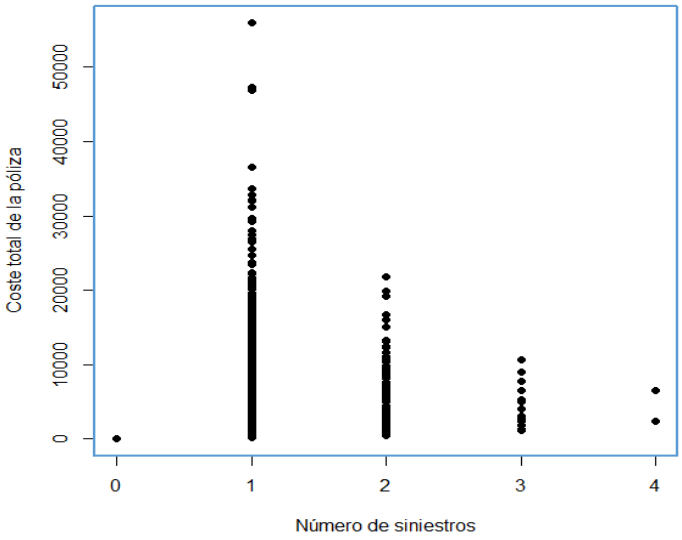
Las bases de datos de pólizas de aseguradoras pueden estar organizadas de múltiples formas. La más habitual es la que sigue un año natural completo, en cuyo caso hay que ser conscientes de que la exposición juega un papel importante ya que no es lo mismo que la póliza esté expuesta al riesgo durante medio año (exposición de 0.5), que durante el año completo (exposición de 1). Además, en nuestro caso, la agregación de los costes por siniestro de cada uno de los clientes se da en una misma póliza, de este modo un coste agregado de 1.890$ (primer registro con dos siniestros de nuestra base de datos) podría hacer referencia a dos siniestros de 945$ para ese cliente, a uno de 630$ y otro de 1.260$, o cualquier otra de las infinitas combinaciones posibles.

El lector podrá observar que dependiendo de la combinación seleccionada nuestros resultados variarán sobremanera, además de que esta se daría sin criterio alguno, de forma aleatoria y subjetiva. Debido a cómo está recopilada nuestra base de datos, aquellas pólizas con más de un siniestro declarado restan validez a nuestros resultados al no poder ser descompuestas en siniestros individuales con características individuales. Por esto serán descartadas las pólizas con más de un siniestro declarado.

Por otro lado, las tres distribuciones seleccionadas requieren de valores positivos en sus funciones de distribución acumuladas, por lo que implícitamente nos hará descartar aquellas frecuencias siniestrales iguales a cero.

No hemos de olvidar que el objetivo de este trabajo es el de centrarse en la modelización de la severidad siniestral, dejando para futuros estudios la modelización de la frecuencia mediante la cual podríamos resolver estos problemas comentados. Esto se daría en modelos en los que se combine severidad y frecuencia[7], o se modelicen conjuntamente en un modelo *premium*.

Podemos ver a continuación (figura 5.2), como a pesar de descartar los valores de siniestros mayores y menores que uno, seguimos conservando el mayor número de pólizas que han generado pérdidas para la empresa. Además, es para el caso de un solo siniestro donde podemos apreciar las mayores pérdidas, por lo que nuestro modelo no perderá esta importante información sobre la zona de las colas que es la que más nos interesa contrastar.



Siniestros	Tota	Proporción
0.	63.232	93.18%
1.	4.333	6.38%
2.	271	0.39%
3.	18	0.03%
4.	2	0.00%

Figura 5.2: Número de siniestros y sus estadísticos.

Fuente: Elaboración propia mediante R.

[7] En la introducción mencionábamos dos trabajos (Sarabia et al. 2016) y (Heller et al. 2007) sobre la base de datos utilizada en el presente documento. En el primer caso se utilizaba un modelo colectivo de Pareto-Poisson. En el segundo caso se proponía Poisson, Poisson inflada de ceros y la distribución negativa binomial para la frecuencia, junto con la distribución gamma para la severidad. Por otro lado, este último proponía la distribución Tweedie para el modelo premium.

5.1.3. MODELIZACIÓN DE LA DISTRIBUCIÓN COSTE TOTAL

Tras un primer filtro de los datos a utilizar, vamos a ver los principales estadísticos de nuestra muestra y su histograma, tras lo que procederemos a la estimación de parámetros por máxima verosimilitud y finalmente compararemos las distintas distribuciones.

5.1.3.1. Principales estadísticos e histograma

Tal y como se puede apreciar (figuras 5.3 y 5.4) nuestros datos tienen una alta asimetría de cola derecha, por lo que se justifica que comparemos nuestros datos con las tres distribuciones seleccionadas.

N	Media	Mediana	Mín.	Máx.	Desv. típ.	Asimetría	Curtosis
4.333	1.946,74	695,96	200	55.922,13	3547,02	5.23	45.77

Figura 5.3: Principales estadísticos para las pólizas con un solo siniestro.

Fuente: Elaboración propia mediante R.

El histograma por su parte ha sido truncado a un siniestro máximo de 15.000$AU para una mejor visualización. Podemos observar que hay un gran número de siniestros cercanos a cero, que hacen referencia a las distintas franquicias que podemos observar en los datos, siendo la de 200$AU la más habitual con un total del 16% para las pólizas con un solo siniestro reclamado.

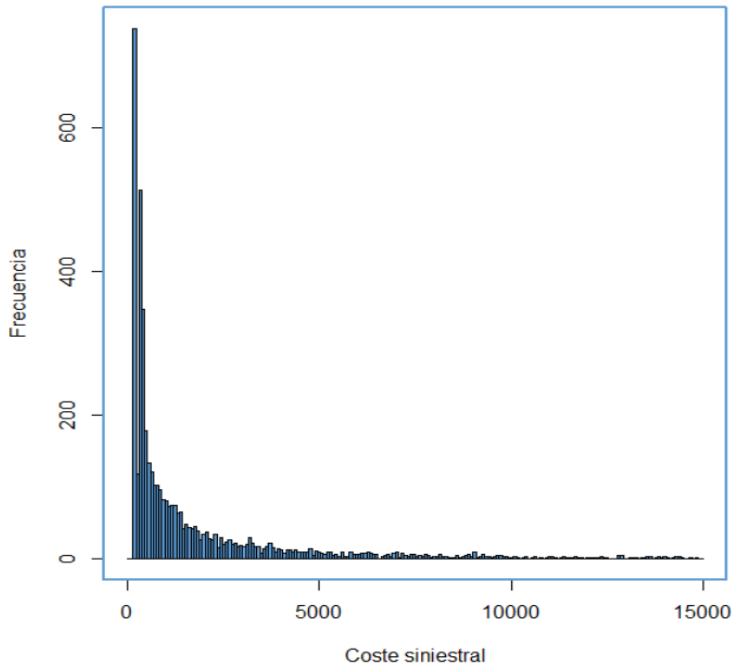


Figura 5.4: Histograma del coste siniestral truncado a 15.000$AU.

Fuente: Elaboración propia mediante R.

5.1.3.2. Parámetros estimados por máxima verosimilitud

Como adelantábamos en la metodología, los parámetros han sido estimados por máxima verosimilitud utilizando el paquete *fitdistrplus* de R, mediante el comando *fitdist* para cada una de las distribuciones. En el caso de Pareto, es necesario definir la función de distribución acumulada y sus momentos, que hemos conseguido mediante el paquete *actuar*. Las estimaciones por máxima verosimilitud y sus intervalos de confianza al 95% para cada uno de los parámetros (obtenidos por simulación) se encuentran representados a continuación (figura 5.6).

	Máx.ver.	Forma	I.C. Forma	Escala	I.C. Escala
Gamma	-36.999	0.7359	0.7079 - 0.7649	2646	2502 - 2773
Log-normal	-36.181	6.8574	6.8214 - 6.8895	1.887	1.863 - 1.912
Pareto	-36.488	1.9583	1.8052 - 2.1607	1964	1742 - 2246

Figura 5.6: Estimaciones de máxima verosimilitud e intervalos de confianza al 95%.

Fuente: Valores estimados mediante R.

De estos resultados podemos extraer que el valor máximo de verosimilitud viene dado por la log-normal (-36.181) entre las tres distribuciones ajustadas. Esto nos dice que es la distribución que mejor ha ajustado los datos ya que lo que buscamos es maximizar la función de verosimilitud, por lo que mayores valores denotan un mejor ajuste.

5.1.3.3. Comparación y elección de la distribución

Comenzaremos este apartado por la comparación gráfica de las distribuciones y continuaremos analizando la evidencia que arrojan los diferentes estadísticos que hemos adelantado en la metodología. La idea de este apartado es seleccionar aquella distribución que mejor ajuste nuestros datos, y que después nos ayude a sacar conclusiones sobre nuestros datos.

5.1.3.3.1. Comparación gráfica

En los siguientes cuatro gráficos (figura 5.7) tenemos el histograma, la función de distribución acumulada, el QQ plot y el PP plot de las tres distribuciones, que hemos incluido en el mismo gráfico para facilitar la interpretación de los resultados.

A priori, los dos primeros gráficos (histograma y FDA) no arrojan ninguna conclusión a tener en cuenta. Ninguna de las tres distribuciones parece tener un ajuste perfecto. Sin embargo, sí podemos observar el truncamiento en cero y las largas colas derechas esperadas. Si acudimos a los coeficientes de asimetría podríamos ver que Pareto (19.10) ajusta una cola mucho más larga que la log-normal (9.09) y esta a su vez mayor que la gamma (2.32).[8]

[8] Coeficientes de asimetría generados para las tres distribuciones mediante simulación en R.

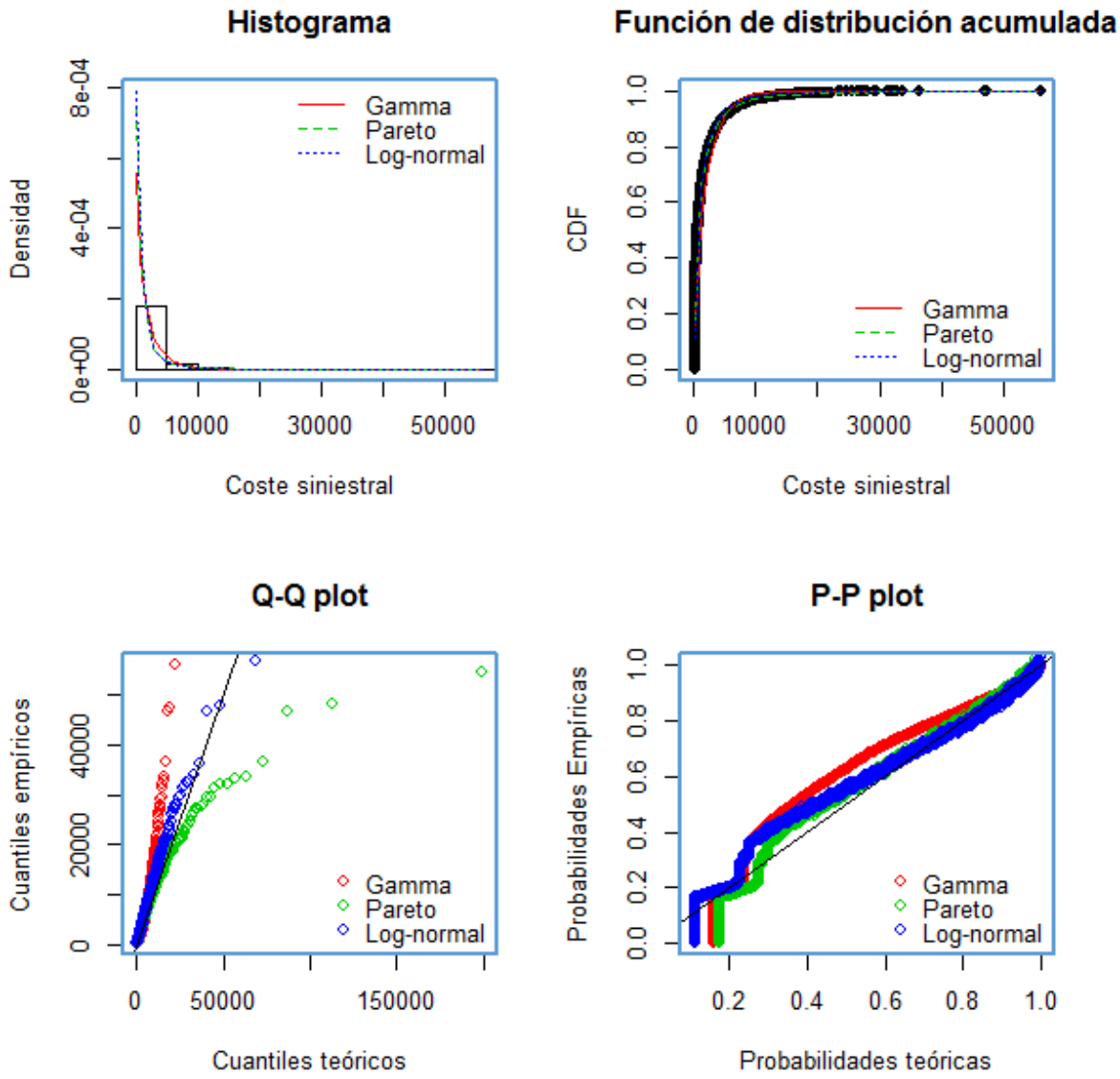


Figura 5.7: Comparación gráfica de las distribuciones.

Fuente: Elaboración propia mediante R.

En cuanto a los gráficos QQ y PP, vemos como en ninguno de los casos el ajuste de la gamma hace un buen trabajo, ya que el PP plot por una parte dibuja un mal ajuste en la zona central de la distribución y el QQ plot nos dice que los datos tienen una cola mayor de lo que deduce la gamma.

Sí que nos pueden generar ciertas dudas las dos distribuciones restantes. Pareto parece tener el mejor ajuste según el PP plot ya que se ve como en su zona central las probabilidades teóricas y empíricas se comportan de una manera más similar que en el caso de la log-normal. Sin embargo, si hacemos referencia al QQ plot, vemos que los cuantiles teóricos de Pareto se alejan mucho de los empíricos, lo que nos hace pensar que nuestros datos no tienen una cola tan larga como deduce esta distribución. La log-normal, que como comentábamos en el capítulo de la metodología tiene una cola a media distancia entre la gamma y Pareto, es sin duda la que mejor ajusta nuestros datos. La cola de nuestros datos es la zona más importante a tener en cuenta en nuestro

estudio, por lo que en esta primera comparación gráfica, nos decantaríamos por la elección de la distribución log-normal.

5.1.3.3.2. Comparación de estadísticos

Hemos visto que mediante la comparación gráfica no podemos seleccionar una distribución con total seguridad, sin bien la log-normal es la que mejor parece actuar. En el siguiente cuadro (figura 5.8) vemos los principales estadísticos discutidos en la metodología, donde detallamos el cálculo de cada uno de ellos.

Distribución	AIC	BIC	KS	AD	CVM
Gamma	74.002,46	74.015,21	0,1581	201,204	36,4253
Pareto	72.980,86	72.993,61	0,1730	94,7386	12,1437
Log-normal	72.366,96	72.379,71	0,1097	80,4486	12,1623

Figura 5.8: Criterios y estadísticos de bondad de ajuste.

Fuente: Valores estimados mediante R.

Comparando los criterios de Akaike y bayesiano, vemos como el valor más pequeño se da para la distribución lognormal. En la metodología adelantábamos que el BIC penalizaba el número de parámetros en mayor medida que el AIC, y vemos que efectivamente su valor es un poco superior para las tres distribuciones. Sin embargo, también comentábamos que al tener todas ellas el mismo número de parámetros, las conclusiones iban a ser las mismas para ambos criterios. La log-normal ajusta mejor nuestros datos según estos estadísticos.

Por otro lado, vemos los estadísticos de bondad de ajuste. Dijimos que **Kolmogorov-Smirnov** tenía un mayor poder estadístico sobre la zona central y en nuestro caso particular nos dice que Pareto sería la peor ajustando esta zona y log-normal la que mejor ajustaría.
Continuando por el estadístico **Anderson-Darlin,** que es el que mayor peso da a las colas, vemos como de nuevo la distribución log-normal es la que mejor se comporta. En este caso ahora es la gamma la que peores resultados ha obtenido, como ya vimos en la comparación gráfica de las colas.
Finalmente, **Cramer-von Mises** nos sorprende ya que por su computación esperábamos una conclusión a medio camino entre los otros dos estadísticos, pero vemos que en este caso la distribución de Pareto es la que mejores resultados ha obtenido. Ahora bien, si el óptimo hubiera sido que todos nuestros estadísticos estuvieran conformes en cuanto a la distribución a ajustar, parecen discernir en los resultados para solo el CMV, que como ya dijimos se ve relegado a un segundo plano en la mayoría de los análisis que hemos contrastado en la literatura.

Es por todo esto que parece justificable la elección de la distribución log-normal para el análisis de nuestros datos, tras ver que se comporta mejor que el resto de distribuciones según la máxima verosimilitud, la comparación gráfica, la comparación de criterios de bondad de ajuste, y finalmente, según dos de los tres estadísticos de bondad de ajuste.

5.1.3.3.3. Medidas de riesgo

Vamos ahora a presentar los resultados para el cálculo del valor en riesgo y valor de la cola en riesgo de cada distribución. Estos han sido estimados mediante una simulación para cada uno de los tres casos, utilizando los parámetros estimados por máxima verosimilitud y generando muestras de tamaño 10.000. Tras la simulación de las muestras, hemos calculado el VaR (figura 5.9) para los siguientes niveles; *p=0.95, 0.97, 0.99, 0.995* y *0.999* con la intención de visualizar cómo se comporta la cola de cada distribución y de este modo también utilizar este estadístico como una herramienta de comparación de distribuciones.

Distribución	**95%**	**97.5%**	**99%**	**99.5%**	**99.9%**
Gamma	6.532	8.266	10.535	12.483	15.818
Log-normal	6.157	9.169	13.273	16.807	30.329
Pareto	7.231	11.398	19.131	29.566	62.743
Datos	7.805	11.648	18.079	22.361	35.546

Figura 5.9: VaR en dólares australianos para las tres distribuciones y los datos reales al p%.

Fuente: Elaboración propia mediante R.

Si comenzamos por el nivel máximo seleccionado (99.9%), podemos observar cómo la distribución de Pareto predice el mayor valor en riesgo siendo 3.96 veces superior al de la gamma y 2.06 veces superior al de la log-normal. Esto justifica que en la definición de las distribuciones dijéramos que Pareto tiene las colas más largas ya que para la simulación a partir de parámetros estimados por unos mismos datos, está siendo mucho más pesimista que sus contrincantes. Esto lo podemos ver desde el primer nivel seleccionado (95%) que ya marcaba una diferencia respecto a las otras dos distribuciones, diferencia que va creciendo a la vez que incrementamos el cuantil de estudio.

Comparando por separado la gamma y log-normal, vemos que sucede lo mismo en este caso. Parten de niveles muy similares y la log-normal llegará a casi doblar la predicción de la gamma para el nivel del 99.9%, esto de nuevo da pie a justificar que la log-normal tiene unas colas más largas. Como curiosidad, se puede apreciar además, que la log-normal no solo tiene colas más largas que la gamma, sino también más pesadas, ya que como vemos, en el nivel del 95% su VaR está por debajo, ya que converge más rápido hacia cero que la gamma[9].

Hemos querido también comparar los resultados de las tres distribuciones con los VaR reales de nuestros datos y nos sorprende ver que van a la par de la distribución Pareto hasta el percentil 99, donde comienzan a divergir y ya en el último tramo el VaR de la Pareto llega a casi doblar el VaR 99.9% real. Esto nos deja ver que las colas reales son menos cortas que las de Pareto, y que comienzan a converger a la de la log-normal en este último tramo.

[9] Esta sentencia queda contrastada matemáticamente en varios trabajos (Halliwell 2013).

En cuanto al TVaR (figura 5.10), vemos como en este caso la diferencia de la Pareto con los datos reales es mucho mayor, ya que ahora cada valor toma en cuenta toda la cola por delante del percentil seleccionado, por lo que las diferencias en las longitudes de las colas quedan mucho más claras. Además, ahora vemos que la log-normal actúa mucho mejor, llegando a converger en los niveles del 99.9% y también vemos como la gamma definitivamente no es apropiada para sacar conclusiones sobre nuestros datos.

Distribución	95%	97.5%	99%	99.5%	99.9%
Gamma	9.042	10.786	12.947	14.464	18.565
Log-normal	11.095	14.779	20.591	26.384	45.923
Pareto	16.729	24.521	39.486	55.708	109.640
Datos	13.996	18.593	25.475	31.201	46.647

Figura 5.10: TVaR en dólares australianos para las tres distribuciones y los datos reales al p%.

Fuente: Elaboración propia mediante R.

Por esto, podemos concluir que en caso de querer realizar un análisis de probabilidad de quiebra utilizando una medida de riesgo coherente, deberíamos de seleccionar la distribución log-normal ya que la gamma produce estimaciones demasiado conservadoras y la Pareto se sitúa en el lado contrario, arrojando valores demasiado optimistas.

5.1.4. MODELO LINEAL GENERALIZADO

En este apartado trataremos de aplicar un modelo lineal generalizado a los datos expuestos anteriormente. Para ello partiremos del modelo propuesto por Jong y Heller (2008) mediante la distribución gamma y función link logarítmica. Veremos que tal y como hemos contrastado en la comparación de distribuciones previa, la distribución logarítmica nos dará un mejor ajuste para este mismo modelo. Finalmente, trataremos de mejorar las predicciones del modelo de Jong y Heller introduciendo aquellas variables explicativas que no han tenido en cuenta.

5.1.4.1. Modelo propuesto por Jong y Heller

Jong y Heller (2008) proponen la modelización del coste siniestral mediante la distribución gamma con función link logarítmica. Las variables explicativas seleccionadas son: la edad de los conductores con base en el tercer grupo, el género con base en el género masculino, el área donde está dada de alta la póliza con base en el área C, el tipo de vehículo siendo su base el tipo sedán (turismo) y finalmente también añade una interacción entre la edad y el género. En la siguiente tabla (5.12), podemos ver en la columna de la izquierda los valores que arroja este modelo, y en la columna de la derecha el modelo ajustado mediante log-normal.

Podemos ver cómo el modelo ajustado mediante la función gamma arroja un alto número de coeficientes no significativos para la mayoría de variables. Apoyamos estos resultados mediante un test de significatividad del modelo con cada una de las variables y de comparación de AIC y deviance mediante el comando *drop1* de R (figura 5.11).

En la tabla (figura 5.11) vemos como solo es significativa la inclusión de la variable área en el modelo. Además, se pueden ver los distintos AIC y deviances que se obtendrían

	Distribución Gamma		Distribución Log-normal	
	Estimación	**P-valor**	**Estimación**	**P-valor**
Intercepto	7.5046	**<0.01**	6.6910	**<0.01**
Grupos de edad				
1	0.4898	**<0.01**	0.3714	**<0.01**
2	0.2252	0.07	0.0653	0.45
4	0.1392	0.25	0.0870	0.29
5	-0.1085	0.42	-0.0027	0.98
6	0.1516	0.34	0.0327	0.76
Área				
A	-0.0884	0.23	-0.0492	0.32
B	-0.0980	0.19	-0.0336	0.51
D	-0.1106	0.24	0.0075	0.91
E	0.0847	0.42	0.1326	0.06
F	0.3412	0.06	0.2599	**<0.01**
Género				
Mujer	-0.0420	0.11	0.0056	0.94
Tipo de vehículo				
Descapotable	-0.4223	0.49	-0.0913	0.82
Deportivo	0.3936	0.70	0.1991	0.77
3 y 5 puertas	0.2882	0.21	0.2557	0.10
Carrocería	0.1167	0.09	0.0660	0.16
Descapotable	0.0489	0.77	-0.0295	0.79
Caravana	-1.0709	0.03	-0.4935	0.14
Minibús	0.4419	0.11	0.4081	0.03
Furgoneta	0.2289	0.33	0.0997	0.53
Deportivos	-1.9999	0.25	-1.3564	0.25
Familiares	-0.0233	0.75	0.0125	0.80
Camiones	0.1699	0.33	0.1139	0.34
Servicios	0.0342	0.78	0.0419	0.62
Género:Edad				
1	-0.3942	**0.05**	-0.3022	**0.02**
2	-0.1802	0.27	-0.0946	0.39
4	-0.1673	0.28	-0.0622	0.62
5	0.0722	0.19	-0.0127	0.93
6				
AIC	**74.001**		**13.798**	
Deviance residual	**6.869 (4.304 g.l.)**		**6.043 (4.304 g.l.)**	

Figura 5.12: Estimaciones y p-valores para GLM gamma y log-normal

Fuente: Elaboración propia mediante R

tras eliminar del modelo la variable seleccionada. Por ejemplo, vemos en el caso de eliminar la variable área como el AIC resultante sería 74.008, el mayor de todos ya que estaríamos eliminando una variable significativa. Como contraste, en caso de eliminar la variable del tipo de vehículo nos quedaríamos con el mejor valor de AIC. Esto nos llevará a eliminar tanto la interacción de género y edad, como la variable del tipo de vehículo de nuestro modelo, ya que no son significativas estadísticamente y empeoran nuestro modelo.

	Deviance	AIC	Pr(>Chi)
Área	6921	74.008	**0.004496**
Tipo veh.	6913	73.992	0.275065
Género:Edad	6890	73.998	0.234927

Figura 5.12: Chi cuadrado y comparación de AIC

Fuente: Elaboración propia mediante R

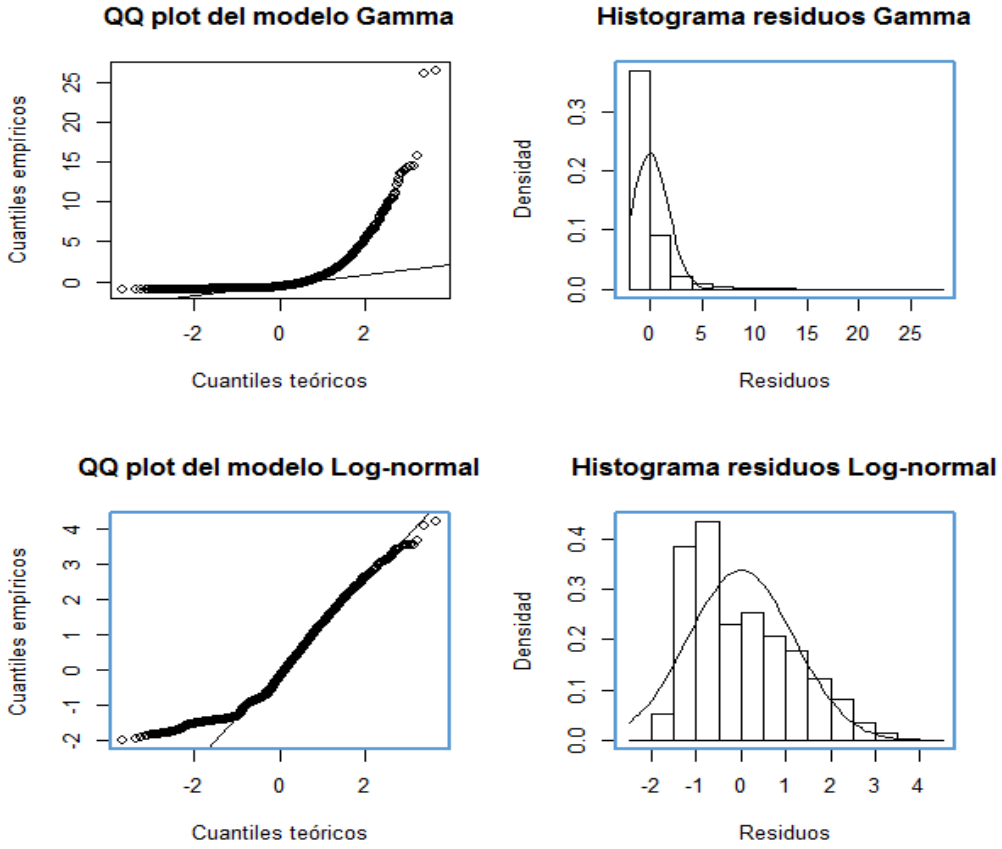


Figura 5.13: Comparación gráfica de la normalidad de los residuos

Fuente: Elaboración propia mediante R

Ahora bien, nuestro objetivo era mejorar las predicciones del modelo de Jong y Heller. Esto lo hacemos mediante la distribución log-normal ya que hemos visto que ajusta mejor nuestros datos que la gamma. Para comparar los resultados de ambos GLMs

vamos a utilizar el criterio de normalidad de los residuos (McCullagh y Nelder 1989), que nos dirá qué distribución se ha comportado mejor (figura 5.13).

Vemos para la gamma que tanto el QQ plot como el histograma distan de distribuirse como una normal. A pesar de que por la zona central se da un buen ajuste, seguimos viendo asimetría hacia las colas, lo que nos da pistas de que el ajuste del GLM mediante gamma no es de buena calidad.

En cuanto a la log-normal, vemos que el ajuste es definitivamente mucho mejor y se acerca más a la normalidad de los residuos. El histograma se parece más al de una normal aunque parece haber cierta tendencia hacia la cola izquierda, quizás por la abundancia de datos con costes siniestrales muy bajos. El QQ plot nos dice lo mismo, en general podemos aceptar la normalidad, salvo esa cola izquierda en la que los cuantiles empíricos están por encima de los teóricos. En cualquier caso, podemos concluir que el modelo GLM mediante log-normal se comparta mejor.

5.1.4.2. Extensión del modelo GLM log-normal

A partir de los datos anteriores y teniendo en cuenta el escaso número de variables que aparecen en nuestro modelo, vamos a excluir las que no aportan nada a nivel interpretativo y a añadir aquellas otras que sí lo hacen.

El modelo resultante tiene como variables explicativas el valor del vehículo tanto en su forma lineal como cuadrática como proponen Heller et al. (2007), el género, la edad del conductor, el área, la antigüedad del vehículo y finalmente la variable de exposición al riesgo introducida en dos tramos (0,0.698,1) para recoger las dos pendientes que dibuja. Esta es una variable con mayor uso en modelos de frecuencia, ya que esperamos ver que a mayor exposición se dé un mayor número de siniestros, pero con la cuantía del siniestro también parece tener cierta relación tal y como veremos a continuación. No hemos introducido ninguna interacción al no ser estas significativas.

En la tabla (figura 5.14) tenemos los coeficientes arrojados por este modelo. Vemos así pues, que todas las variables introducidas son significativas y que el modelo ha mejorado en AIC y deviance.

El valor exponencial del intercepto $\exp(\beta_0) = \exp(6.84) = 934.49\$AU$ nos da el coste medio de una póliza que declare un siniestro, para un hombre, del grupo de edad 3, que vive en el área C, que tiene un coche nuevo (categoría 1), y que ha estado asegurado con esta póliza menos de 255 días (70% de un año).

A partir de esa póliza, veríamos que una mujer promedio generaría un coste menor, de solo el 91.53% del de un hombre. También veríamos como aquellos clientes más jóvenes (categoría 1) provocarían un coste un 22% mayor que los de la categoría 3, no siendo significativa la diferencia con el resto de grupos. Los conductores del área F generarían un coste 1.31 veces superior a los del área base. Vemos además como a partir del tercer grupo de antigüedad, el coste promedio también aumenta, siendo un 13% en el caso del grupo 3 y llegando hasta un 17% para los coches más antiguos (grupo 4).

	Estimación	P-valor
Intercepto	6.8410	<0.01
Valor veh.		
Lineal	-0.0811	0.05
Cuadrático	0.0123	0.04
Edad		
Grupo 1	0.2013	<0.01
Género		
Mujer	-0.0884	0.01
Área		
F	0.2693	<0.01
Exposición		
>0.698	-0.1647	<0.01
Antig. veh.		
Grupo 3	0.1235	0.03
Grupo 4	0.1593	0.02
AIC	**13.755**	
Deviance	**6.014 (4.315 g.l.)**	

Figura 5.14: Extensión del GLM log-normal

Fuente: Elaboración propia mediante R.

6. CONCLUSIONES

Durante la realización de este trabajo hemos estudiado las diferentes técnicas estadísticas para el ajuste de distribuciones de la familia exponencial: gamma, log-normal y Pareto. Hemos propuesto varios medios de comparación de distribuciones con el objetivo de seleccionar aquella que mejor se ajuste a unos datos. Hemos concluido que la mejor distribución que mejor se ha comportado respecto a nuestros datos ha sido la log-normal, ya que ha arrojado los mejores valores de máxima verosimilitud, AIC, BIC, Kolmogorov-Smirnov y Anderson-Darlin. También ha sido la que más se ha aproximado en su ajuste a los cuantiles teóricos de nuestros datos. Finalmente, en esta primera parte hemos comparado los valores en riesgo (VaR) y de la cola en riesgo (TVaR) viendo que la log-normal es de nuevo la que mejor predice el riesgo de quiebra, muy importante en nuestro análisis.

En la última parte de este trabajo, tras ver que la log-normal ajustaba mejor los datos, hemos estimado el GLM gamma con link logarítmica propuesto por Jong y Heller (2008) y lo hemos comparado con un GLM log-normal mediante dos gráficos de diagnosis de normalidad de los residuos. Hemos llegado a la conclusión de que efectivamente la predicción es más válida si consideramos que los errores siguen una distribución log-normal. Tras esto, hemos mejorado el nuevo modelo mediante la eliminación de la variable del tipo de vehículo y de la interacción entre género y edad, y del mismo modo hemos añadido la variable del valor del vehículo en forma lineal y cuadrática, el grupo de edad, el género, el área, la exposición al riesgo y la antigüedad del vehículo, las cuales son estadísticamente significativas y del mismo modo han mejorado el AIC y deviance, resultando en un modelo mejor. Finalmente, hemos podido generar interpretaciones para una póliza tipo, hombre, del grupo de edad 3, que vive en el área C, que tiene un coche nuevo (categoría 1), y que ha estado asegurado con esta póliza menos de 255 días (70% de un año).

BIBLIOGRAFÍA

Todas las citas y referencias a otros autores han sido formateadas de acuerdo a las reglas publicadas por la Universidad de Cantabria en Universidad de Cantabria, Biblioteca (2013).

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Aut. Control 19:716–723.

Artzner, P.; Delbaen, F.; Eber, J.; Heath, D. 1999. Coherent Measures of Risk, Mathematical Finance, Vol. 9, No.3. pp. 203-228.

Anderson, T. W.; Darling, D. A. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics. 23: 193–212.

Box, G; Draper, N. 1987. Empirical Model-Building and Response Surfaces. p. 424, New York: Wiley Series in Probability and Statistics.

Cachán, A. 2016. Tarificación en espacios de alta dimensionalidad a través del aprendizaje automático. Madrid: Trabajo fin de Máster. Universidad Carlos III de Madrid.

Conover, W.J. 1999. Practical Nonparametric Statistics. Tercera edición, New York: John Wiley & Sons, Inc., pp.428-433.

Cramér, H. 1928. On the Composition of Elementary Errors. Scandinavian Actuarial Journal. Pp. 13–74.

Cramér, H, 1946. Mathematical Methods of Statistics. Princeton: Princeton University Press.

Flowers-Cano, R.S; Flowers, R.J.; Rivera-Trejo, F. 2014. Evaluación de criterios de selección de modelos probabilísticos: validación con series de valores máximos simulados. Universidad Juárez Autónoma de Tabasco. División Académica de Ingeniería y Arquitectura. México.

Gnanadesikan, R., 1997. Methods for Statistical Data Analysis of Multivariate Observations. Segunda Edición. New York: John Wiley & Sons, Inc.

Golburd, M, Khare, A, Tevet, D, 2016. Generalized Linear Models for Insurance Rating. Virginia: Casualty Actuarial Society. CAS Monograph Series. Number 5.

Heller, G.Z.; Stasinopoulos, D.M.; Rigby, R.A.; Jong, P. 2007. Mean and Dispersion modelling for policy claims costs. Scandinavian actuarial journal, Vol. 2007, Issue 4, p.281-292

Halliwell, L.J. 2013. Classifying the Tails of Loss Distributions. Casualty Actuarial Society E-Forum, Spring 2013-Volume 2.

Jong, P; Heller, G.Z. 2008. Generalized Linear Models for Insurance Data. New York: Cambridge University Press.

Klugman, S.A.; Panjer, H.H.; Willmont, G.E. 2012, Loss Models, From Data to Decisions. 4ª edición. New Jersey: John Wiley & Sons, Inc.

Lilliefors, H. 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, Vol. 62. pp. 399–402.

Lilliefors, H. 1969. On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association, Vol. 64. pp. 387–389.

Martín, J.A. 2016. Análisis e inclusión de variables exógenas en la tarificación de autos mediante modelización por GLM. Madrid: Trabajo fin de Máster. Universidad Carlos III de Madrid.

Massey, F. J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association. Vol. 46, No. 253, pp. 68–78.

Nelder, J.A.; Wedderburn, R.W.M, 1972. Generalized Linear Models. London: Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3. Pp. 370-384.

Plaza, L. 2016. Sistemas de Geolocalización (GIS) en el Pricing GLM del Seguro Multirriesgo del Hogar. Madrid: Trabajo fin de Máster. Universidad Carlos III de Madrid.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Razali, N. M.; Wah, Y.B. 2011. Power comparisons of Shapiro-Wilks, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistics Modeling and Analytics. Vol. 2 No.1, pp. 21-33, 2011.

Sarabia, J.M.; Gómez, E.; Vázquez, F.J., 2007. Estadística actuarial. Teoría y aplicaciones. Madrid: Pearson Educación, S.A. Madrid, 2007.

Sarabia, J.M.; Gómez-Déniz, E.; Prieto, F.; Jordá, V. 2016. Risk aggregation in multivariate dependent Pareto distributions. Insurance: Mathematics and Economics 71 (2016) 154-163.

Schwarz, Gideon E. 1978. Estimating the dimension of a model. The Annals of Statistics. Vol. 6, No. 2, p. 461-464.

Szegö, G. 2002. Measures of Risk. Roma: Journal of Banking & Finance 26 (2002) 1253-1272.

Universidad de Cantabria, Biblioteca, 2013. Tutorial de autoformación sobre cómo Citar en Trabajos y Artículos con Referencias [sitio web]. [Santander]: la Biblioteca. [Consulta: 26 junio 2017]. Disponible en:
http://www.buc.unican.es/sites/default/files/tutoriales/CITAR/PAG0.html

Whittingham, M.J.; Stephens, P.A.; Bradbury, R.B.; Freckelton, R.P. 2006. Why do we still use stepwise modelling in ecology and behaviour? Bristol: Journal of Animal Ecology 2006. 75, 1182-1189.

Wirch J. 1999. Raising Value at Risk. North American Actuarial Journal, 3, 106-115.