Universidad del País Vasco Euskal Herriko Unibertsitatea

EKONOMIA ETA ENPRESA FAKULTATEA
FACULTAD DE ECONOMÍA Y EMPRESA

Grado en Finanzas y Seguros

Curso 2016 – 2017

Técnicas estadísticas de la tarificación del seguro del automóvil

Autor: Daniel Del Castillo Llano

Directora: María Araceli Garín Martín

Bilbao, a 17 de febrero de 2017

Índice general

Resumen v

Abstract v

Introducción 1

Metodología 2

1. Tarificación *a priori* del seguro del automóvil 3
1.1. Características de los Modelos Lineales Generalizados . . . 4
1.1.1. El componente aleatorio . . . 4
1.1.2. El modelo lineal . . . 5
1.1.3. La función de enlace . . . 5
1.1.4. El parámetro de exposición . . . 6
1.2. Estimación de los Modelos Lineales Generalizados . . . 7
1.3. Tipos de Modelos Lineales Generalizados . . . 7
1.3.1. Modelo de Regresión de Poisson . . . 7
1.3.2. Modelo de Regresión Binomial Negativo . . . 9
1.4. Medidas de la bondad del ajuste de los MLG . . . 11
1.4.1. Sobredispersión . . . 11
1.4.2. Error cuadrático medio . . . 12
1.4.3. Desviación . . . 12
1.4.4. Criterio Informativo de Akaike . . . 12
1.4.5. Validación Cruzada . . . 13
1.5. Métodos de selección de variables . . . 13
1.5.1. Selección Progresiva . . . 13
1.5.2. Selección Regresiva . . . 13
1.5.3. Selección Escalonada . . . 14
1.6. Cálculo de la prima *a priori* . . . 14

2. Tarificación a posteriori, sistemas de ajuste de la tarifa 15
2.1. Sistema de Tarificación Bonus-Malus . . . 15
2.1.1. Prima relativa . . . 16
2.1.2. Clase inicial . . . 18
2.1.3. Reglas de transición . . . 18
2.2. Sistema de Tarificación de Franquicia . . . 18
2.2.1. Sistema combinado de Gran Franquicia y Crédito . . . 18
2.3. Sistema de Tarificación Basados en el Uso . . . 19

3. Un caso práctico de tarificación de responsabilidad civil en automóviles **20**
3.1. Datos . . . 20
3.1.1. Identificación . . . 20
3.1.2. Depuración y transformación . . . 22
3.1.3. Análisis descriptivo univariante y adaptación . . . 23
3.1.4. Análisis multivariante . . . 30
3.2. Tarificación *a priori* . . . 31
3.2.1. Selección de modelo y variables . . . 31
3.2.2. Estimación y validación . . . 34
3.2.3. Tarificación . . . 35
3.3. Tarificación *a posteriori* . . . 36
3.3.1. Escenarios de simulación . . . 37
3.3.2. Tabla Bonus–Malus . . . 37
3.3.3. Análisis de la simulación de la tarificación *a posteriori* . . . 38

4. Conclusiones **41**

Anexo 1 **42**

Anexo 2 **43**

Anexo 3 **53**

Índice de figuras

1. Frecuencia del número de reclamaciones 24
2. Distribución del número medio de reclamaciones por área de residencia 25
3. Distribución del número medio de reclamaciones por potencia del vehícul 26
4. Distribución del número medio de reclamaciones por edad del vehículo 27
5. Distribución del número medio de reclamaciones por edad del asegurado 28
6. Distribución del número medio de reclamaciones por marca del vehículo 29
7. Frecuencia del número de reclamaciones por densidad de habitantes 30
8. Validación cruzada del modelo de tarificación a priori 35
9. Frecuencia de las tarifas a priori 36
10. Simulación de la cartera 39
11. Frecuencia de las tarifas en el décimo año 39
12. Simulación del comportamiento individual de las pólizas 40

Índice de cuadros

1. Relación distribución–función de enlace canónica 6
2. Diccionario de datos de la BBDD freMTPL2freq 21
3. Diccionario de la BBDD freMTPL2sev 21
4. Frecuencia del número de reclamaciones (ClaimNb) 22
5. Estadísticos descriptivos del número de reclamaciones 24
6. Reclasificación de variable “Area” 26
7. Reclasificación de variable “VehPower” 26
8. Reclasificación de variable “VehAge” 27
9. Reclasificación de variable “DrivAge” 28
10. Reclasificación de variable “VehBrand” 29
11. Reclasificación de variable “VehGas” 29
12. Reclasificación de variable “Density” 30
13. Matriz de correlaciones 31
14. Parámetros y errores de los modelos analizados 33
15. Parámetros 33
16. Frecuencia del número de siniestros (Predicción) 33
17. Medidas de la bondad del ajuste 34
18. Tabla Bonus–Malus 38
19. Guía tabla de tarifas 53
20. Tabla de tarifas 54

Resumen

El seguro del automóvil constituye, debido a su obligatoriedad en la mayoría de países, uno de los principales ramos del sector asegurador en el mundo. Su tarificación consta generalmente de una parte *a priori*, para lo cual se utilizan las características observables de la póliza y una parte *a posteriori*, para la cual se utiliza la siniestralidad experimentada. En la tarificación *a priori* es común la aplicación de modelos lineales generalizados, concretamente de modelos de datos de conteo. Para la tarificación *a posteriori* se utiliza, entre otros, el sistema Bonus–Malus. El objetivo del presente trabajo es tanto realizar un análisis teórico de estas técnicas, como realizar un proceso de tarificación completo con los datos de una cartera real.

Palabras clave: Tarificación del seguro de automóvil, sistema Bonus–Malus, modelos de datos de conteo.

Abstract

Vehicle insurance is, due to its compulsory nature in most of the countries, one of the main insurance types in the world. Its pricing generally consists of an *a priori* part, for which the observable characteristics of the policy are used and an *a posteriori* part, for which the claims history is used. In the *a priori* pricing it is common the application of generalized linear models, specially count data models. For the *a posteriori* pricing, Bonus–Malus systems are used, among others. The objective of the present thesis is to perform a theoretical analysis of these techniques, as well as to perform a complete pricing process with the data of a real portfolio.

Keywords: Vehicle insurance pricing, Bonus–Malus system, count data models.

Introducción

El sector asegurador constituye una de los sectores principales en las economías desarrolladas, tanto el grupo incluido dentro del llamado ramo de vida como el resto de ramos englobados en no-vida. Con una cuota del 38 % de las primas totales del negocio de seguros de no-vida en Europa, el ramo del seguro de automóviles se muestra como líder indiscutible en esta tipología de seguros, mostrando un crecimiento del 1 % en el año 2015. El presente trabajo se va a centrar precisamente en el ramo del seguro de automóviles, mostrando las técnicas estadísticas de tarificación empleadas habitualmente en el mismo y su aplicación.

La confluencia de varías áreas de estudio presentes en el Grado en Finanzas y Seguros en el proceso de la tarificación, lo convierten en un objetivo de estudio de notable interés. Concretamente las aportaciones de la estadística actuarial en su vertiente estocástica, con las distribuciones de probabilidad de las variables aleatorias, las aportaciones de la econometría con los modelos de regresión, así como las aportaciones de las matemáticas actuariales de no-vida con los modelos de riesgo.

Se trata de un tema que ha sido tratado por numerosos autores, entre ellos, (Lemaire, 1995) y (Kaas, 2009). Por norma general la tarificación se trata en dos fases, en la primera, llamada *a priori*, se realiza una predicción de la siniestralidad en base únicamente a las características observables de la póliza, mediante modelos lineales generalizados o de datos de conteo. La aplicación de estos modelos no es exclusiva del ámbito actuarial, cabe destacar el estudio de (Nelder y Wedderburn, 1972) en el ámbito general. Y una segunda fase de tarificación, llamada *a posteriori*, donde la tarifa se ajusta a la siniestralidad individual del asegurado, cabe destacar el trabajo de (Lemaire, 1995) acerca del sistema Bonus–Malus óptimo.

El primer objetivo de este trabajo es realizar una revisión teórica de las técnicas estadísticas y econométricas de tarificación. Para ello se va a analizar en primer lugar varios modelos de datos de conteo utilizados en la tarificación *a priori*, concretamente los modelos de regresión de Poisson y Binomial Negativo. En segundo lugar se van a analizar los sistemas de tarificación *a posteriori*, en especial el sistema Bonus-Malus.

El segundo objetivo es examinar tanto la efectividad en el ajuste de la prima al riesgo de los métodos anteriores, como su viabilidad financiera mediante su puesta en práctica en un caso real de tarificación del seguro de responsabilidad civil del automóvil.

El objetivo de este trabajo se sitúa en la parte estadística del proceso de tarificación, sin profundizar en temas relacionados como el marketing, la regulación o el reaseguro. Se asumirá también que no hay contagio en los siniestros, es decir, que son independientes unos de otros. La fijación de las tarifas se establecerá únicamente en función de la estimación de la siniestralidad observada, sin hacer referencia a la severidad de los siniestros, ni a la ocurrencia de siniestros no comunicados.

Tras esta breve introducción, en el Capítulo 1 se presenta el desarrollo teórico de la tarificación *a priori*. En el Capítulo 2 se presenta el desarrollo también teórico de varios tipos de sistemas de tarificación *a posteriori*. Por último, en el Capítulo 3 se desarrolla el proceso de tarificación en sí, realizando posteriormente un proceso de simulación donde poder observar su comportamiento

a lo largo del tiempo, todo ello mediante el apoyo del software estadístico R. El trabajo termina con las principales conclusiones obtenidas en su desarrollo, las tablas de tarificación obtenidas y una relación de la bibliografía consultada en su elaboración.

Metodología

La metodología llevada a cabo para la revisión teórica de las técnicas estadísticas de tarificación del seguro del automóvil ha sido principalmente la obtención de información a través de fuentes secundarias o indirectas, generalmente libros y artículos, tomando como punto de partida los conocimientos adquiridos en diversas asignaturas del Grado.

Por otro lado, la metodología llevada a cabo en el caso práctico de tarificación consta de un proceso con varias partes:

1. Se establece la población como el conjunto total de pólizas del automóvil.
2. Se selecciona como muestra una cartera de pólizas de un asegurador.
3. Se realiza un análisis descriptivo, selección y transformación de las variables disponibles de las pólizas, tarificación *a priori*, validación del modelo de tarificación, tarificación *a posteriori* y una simulación para observar su comportamiento a lo largo del tiempo.

Capítulo 1

Tarificación *a priori* del seguro del automóvil

El seguro del automóvil es un seguro mediante el cual el tomador del seguro desea mitigar los costes financieros que pudieran surgir tras verse involucrado en un siniestro con un coche, moto, camión, etc. El tomador del seguro, mediante el pago de la prima, satisface los gastos que el asegurador estima que le ocasionará durante el periodo de pago al que se refiera, generalmente en el espacio temporal de un año. La naturaleza fortuita de los sucesos a los que se compromete el asegurador a hacer frente, deriva en que se dé el caso de que un cierto número de asegurados no realice ninguna reclamación, pero que uno de ellos realice una reclamación por una importante cantidad económica. Este método de mancomunación de los riesgos es en el que se basa el principio de solidaridad humana de los seguros.

Debido al gran tamaño de las carteras de asegurados, un primer planteamiento de simplificación será segmentarlas en grupos homogéneos, con el objetivo de que las pólizas incluidas en un mismo grupo que compartan las mismas características observables paguen la misma prima (Kaas, 2009). Esta técnica permitirá tanto facilitar la estimación de los modelos, como comunicar las tarifas de un modo muy simple.

La necesidad de explicar la relación de dependencia entre la variable que se desea estimar, "número de siniestros" y las variables de que se dispone acerca de las características de las pólizas, lleva a realizar un primer planteamiento acerca de la idoneidad de utilizar un modelo de regresión lineal, debido principalmente a su sencillez de aplicación y potencia de predicción. Por desgracia, se presentan varios problemas a la hora de utilizar este método en la siniestralidad del automóvil en relación al cumplimiento de sus hipótesis básicas. En primer lugar, la variable dependiente "número de siniestros" no tiene una distribución continua, se trata de un recuento, por lo tanto, será siempre no-negativa y contendrá números enteros. En segundo lugar, no se puede afirmar que esta variable se distribuya con una distribución normal, por las mismas razones. En tercer lugar, la relación entre los predictores y la variable dependiente puede no ser lineal. En cuarto lugar, la varianza de las perturbaciones puede no ser constante a lo largo de las observaciones y presentarse heterocedasticidad.

Para sobrepasar la rigidez impuesta por estos modelos, Nelder y Wedderburn propusieron una extensión de estos, donde la variable dependiente pueda estar distribuida como una distribución de la familia exponencial y además pueda ser discreta, les llamaron Modelos Lineales Generalizados (Nelder y Wedderburn, 1972).

1.1. Características de los Modelos Lineales Generalizados

La estructura de los MLG puede describirse como un conjunto de tres componentes fundamentales (McCullagh y Nelder, 1989), el componente aleatorio o distribución de probabilidad de ocurrencia del suceso de conteo, el modelo lineal y la función de enlace.

1.1.1. El componente aleatorio

Los MLG tratan de describir la dependencia de una variable de respuesta escalar y_i respecto a un vector de regresores, x_i, tal que:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1.1}$$

Donde el valor observado de la variable y_i será una realización de las variables aleatorias y mutuamente independientes pero no idénticamente distribuidas $Y_1, Y_2, ..., Y_n$. La distribución condicionada de $y_i|x_i$ que estime el valor de la variable aleatoria Y_i, será una distribución de la familia exponencial con la siguiente función de densidad de probabilidad:

$$f_{Y_i}(y_i; \theta_i, \phi, \omega_i) = exp\left[\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi, \omega_i)} + c(y_i, \phi)\right] \tag{1.2}$$

Sea θ_i el parámetro canónico que depende de los regresores, que variará de acuerdo a las características de éstos y ϕ el parámetro de dispersión, mediante la selección de las funciones $a(\phi, \omega_i)$, $b(\theta_i)$ y $c(y_i, \phi)$ se determinará qué miembro de la familia de distribuciones se utilizará, pero siempre deberán garantizar las siguiente condiciones:

- $a(\phi, \omega_i)$ será en todo caso positiva y continua;
- $b(\theta_i)$ debe ser una función continuamente diferenciable (sus derivadas deben ser siempre continuas); y
- $c(y_i, \phi)$ será independiente de θ_i

Se establece la función de dispersión $a(\phi, \omega_i)$ como

$$a(\phi, \omega_i) = \frac{\phi}{\omega_i}, \tag{1.3}$$

donde ω_i es un valor conocido de ponderación de cada observación, como puede ser la credibilidad de ésta o el tiempo de exposición. Estos valores son utilizados para corregir el efecto que tiene el hecho de que las distintas observaciones tengan distinta varianza y que, por lo tanto, se pueda mantener la hipótesis de que la dispersión ϕ es constante. Con carácter general y excepto que sea necesario recurrir a él, se fijará el valor de ϕ a uno, eliminando su efecto.

En resumen, las funciones $a(\phi, \omega_i)$, $b(\theta_i)$ y $c(y_i, \phi)$ serán iguales para todas las realizaciones de la variable aleatoria Y_i y será el parámetro θ_i (siempre que no se introduzcan efectos de exposición o ponderación) el que marque la diferencia entre cada estimación de la misma.

El valor esperado de la variable aleatoria Y_i se obtendrá simplemente como la primera derivada de $b(\theta_i)$, es decir:

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{1.4}$$

Y su varianza, como un valor proporcional a la dispersión, es decir, a la derivada segunda de $b(\theta_i)$:

$$Var(Y_i) = a(\phi, \omega_i) \cdot b''(\theta_i) \tag{1.5}$$

1.1.2. El modelo lineal

Este componente del modelo, también llamado *componente sistemático* relaciona las variables independientes de la forma

$$\beta_0 + X_1\beta_1 + X_2\beta_2 + ... + X_p\beta_p \tag{1.6}$$

y por lo tanto atribuye a cada observación el predictor lineal

$$\eta_i = \sum_{j=0}^{p} x_{ij}\beta_j = X_i^T\beta, \quad (i = 1, 2, ..., n) \tag{1.7}$$

cuyos coeficientes son representados por la fila iésima, X_i, de la matriz de variables independientes X y sus parámetros por el vector de parámetros

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \tag{1.8}$$

Como puede observarse, será necesario introducir una fila constante e igual a 1 en X, mediante la cual, β_0 recogerá los efectos en la variable de respuesta cuando todas las variables independientes sean igual a 0.

1.1.3. La función de enlace

En un modelo de regresión lineal la variable de respuesta puede fluctuar en un rango $(-\infty, \infty)$ y se espera que esté normalmente distribuida. No sucede lo mismo en los modelos de recuento analizados en este trabajo, donde la variable dependiente se espera que tome un valor en el rango $[0, +\infty)$. Sea $E(Y_i) = \mu_i$, donde, se establece una transformación no-lineal de la media μ_i de la variable Y_i tal que

$$g(\mu_i) = \eta_i, \quad (i = 1, 2, ..., n) \tag{1.9}$$

En este caso, la función $g(\cdot)$ es una función monótona y diferenciable llamada función de enlace. Estableciendo una relación no-lineal entre la variable de respuesta y las variables explicativas. Será de mayor utilidad considerar la media de la variable de respuesta como una función del predictor lineal, para ello habitualmente se utiliza la inversa de la función $g(\cdot)$ del siguiente modo:

$$\mu_i = g^{-1}(\eta_i), \quad (i = 1, 2, ..., n) \tag{1.10}$$

Introduciendo la relación (1.7) se podrá observar la relación entre la media μ_i y la combinación lineal de regresores $X_i = (x_{i1}, x_{i2}, ...x_{ip})$:

$$\mu_i = g^{-1}(X_i^T \beta) = g^{-1}(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p), \quad (i = 1, 2, ..., n) \tag{1.11}$$

Por último, para relacionar el valor del parámetro canónico θ_i con los regresores X_i se partirá de la relación (1.4)

$$\mu_i = b'(\theta_i) \iff \theta_i = (b')^{-1}(\mu_i), \quad (i = 1, 2, ..., n) \tag{1.12}$$

donde, de (1.11),

$$\theta_i = (b')^{-1}(g^{-1}(\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p)), \quad (i = 1, 2, ..., n) \tag{1.13}$$

Para facilitar el cálculo del modelo, entre otras ventajas, se propone seleccionar una función de enlace g que establezca una relación lineal entre el parámetro canónico θ_i y el predictor lineal η_i, tal que

$$\theta_i = (b')^{-1}(g^{-1}(\eta_i)) = \eta_i, \quad (i = 1, 2, ..., n) \tag{1.14}$$

A esta función de enlace, tal que su inversa, g^{-1}, es la inversa de la primera derivada de la función, $b(\cdot)$, se le denomina función de enlace canónica. Una pequeña relación de las funciones de distribución analizadas en este trabajo y sus funciones de enlace canónicas se muestra en el Cuadro 1.

Cuadro 1

Relación distribución–función de enlace canónica

Distribución de Y_i	Función de enlace canónica	$g(x)$	$g^{-1}(x)$	Forma
Poisson	Logarítmica	$\ln(x)$	e^x	$\eta_i = \ln(\mu_i)$
Binomial Negativa	Logarítmica	$\ln(x)$	e^x	$\eta_i = \ln(\mu_i)$

1.1.4. El parámetro de exposición

La inclusión de un parámetro de exposición o desplazamiento en los MLG de datos de conteo es un procedimiento habitual, ya que generalmente el conteo de los datos no está especificado para periodos iguales en las distintas observaciones de las muestras de datos. En el ámbito asegurador entre otros, el periodo de grabación de los datos coincide con el año natural, mientras que los periodos de vigencia de las pólizas anuales se encuentran desfasados respecto a éste. Por lo que la información grabada sobre las reclamaciones reflejara únicamente una porción de la vigencia total del contrato. Además de poderse introducir la exposición como un valor de ponderación ω_i, también puede incorporarse como un efecto conocido al predictor lineal η_i, mediante ξ_i, del siguiente modo:

$$\eta_i = X_i^T \beta + \xi_i, \quad (i = 1, 2, ..., n) \tag{1.15}$$

de donde se establece la siguiente relación con el valor esperado de la variable aleatoria Y_i:

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(X_i^T \beta + \xi_i), \quad (i = 1, 2, ..., n) \tag{1.16}$$

Este efecto también se puede aprovechar para simplificar el número de observaciones mediante la agrupación de observaciones con las mismas características, acumulando el cómputo del valor de la exposición de las mismas.
Bajo la hipótesis de que el número de siniestros crezca linealmente con la exposición, se establecerá la siguiente relación:

$$\xi_i = \ln(\delta_i), \quad (i = 1, 2, ..., n) \tag{1.17}$$

donde δ_i es el tiempo de exposición de la i-ésima observación.

1.2. Estimación de los Modelos Lineales Generalizados

La estimación de los parámetros del modelo será el camino a poder realizar una inferencia sobre los datos, para ello, en (McCullagh y Nelder, 1989) se propone en primer lugar determinar una medida de la bondad del ajuste entre los valores estimados por el modelo y los datos. Los valores máximo-verosímiles de los parámetros β del modelo serán aquellos que minimicen la desviación, que será analizada en el apartado 1.4.2. Para conseguirlo, en segundo lugar, se propone utilizar un algoritmo de mínimos cuadrados ponderados iterativo y debido a que en general no se dispone de una ecuación cerrada que cumpla esas características óptimas, será necesario realizar la estimación mediante métodos numéricos.

1.3. Tipos de Modelos Lineales Generalizados

Tal y como se explicó con anterioridad, la configuración de MLG con diferentes funciones de distribución llevará a que éstos se ajusten con mayor o menor precisión a los datos y que por lo tanto las conclusiones extraídas a partir de éstos tengan una distinta validez. Se presentan a continuación varios ejemplos.

1.3.1. Modelo de Regresión de Poisson

La distribución más simple utilizada para modelos de datos de conteo es la distribución de Poisson, puesto que se trata de un caso especial en el que el parámetro de dispersión $\phi = 1$. Las propiedades subyacentes a una distribución de Poisson (Kupper, 1963) son:

- La población estudiada es homogénea.
- La ocurrencia de un siniestro es un evento raro, o lo que es lo mismo, la probabilidad de ocurrencia es muy pequeña.
- No existe contagio, es decir, la ocurrencia de un siniestro posterior no está influenciada por los anteriores.

La función de cuantía o de probabilidad de la distribución de Poisson se obtendrá a partir de la función (1.2), fijando la siguiente configuración:

- $a(\phi, \omega_i) = 1$
- $b(\theta_i) = e^{\theta_i}$
- $c(y_i, \phi) = -\ln(y_i!)$

- $\theta_i = \ln(\mu_i)$

Sustituyendo en la ecuación (1.2) y operando, se obtiene la función de probabilidad de una variable aleatoria Y_i que sigue una distribución de Poisson de parámetro μ_i:

$$f_{Y_i}(y_i;\mu_i) = exp\left[\frac{y_i \cdot \ln(\mu_i) - e^{\ln(\mu_i)}}{1} + (-\ln(y_i!))\right] = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i \geq 0, \mu_i > 0 \tag{1.18}$$

Del mismo modo, sustituyendo en la ecuación (1.4) se obtiene la media de la distribución:

$$E(Y_i) = b'(\theta_i) = (e^{\theta_i})' = e^{\theta_i} = e^{\ln(\mu_i)} = \mu_i \tag{1.19}$$

Y de la ecuación (1.5) la varianza:

$$Var(Y_i) = a(\phi, \omega_i) \cdot b''(\theta_i) = 1 \cdot (e^{\theta_i})'' = e^{\theta_i} = e^{\ln(\mu_i)} = \mu_i \tag{1.20}$$

La estimación del modelo se hará mediante el método de máxima verosimilitud. La función de versimilitud obtenida a partir de la función (1.18) será:

$$\mathcal{L}(y;\mu) = \prod_{i=1}^{n} f_{Y_i}(y_i;\mu_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{1.21}$$

Para facilitar su operación se trabajará sobre la función de log-verosimilitud, ya que el resultado en su maximización es equivalente.

$$log\mathcal{L}(y;\mu) = \sum_{i=1}^{n} [y_i \cdot \ln(\mu_i) - \mu_i - \ln(y_i!)] \tag{1.22}$$

A continuación se introducirá la relación con los regresores mediante la función de enlace canónica, que en el caso de Poisson es logarítmica y de forma $\mu_i = exp(\sum_{j=1}^{p} x_{ij}\beta_j + \xi_i)$, por lo tanto:

$$log\mathcal{L}(y; e^{X\beta+\xi}) = \sum_{i=1}^{n} \left[y_i \cdot \sum_{j=1}^{p}(x_{ij}\beta_j + \xi_i) - exp\left[\sum_{j=1}^{p}(x_{ij}\beta_j + \xi_i)\right] - \ln(y_i!)\right] \tag{1.23}$$

El valor de la log-verosimilitud se maximizará cuando para cada valor de j, la derivada parcial de primer orden de la función de log-verosimilitud respecto de β_j se iguale a 0

$$\frac{\partial log\mathcal{L}}{\partial \beta_j} = 0, \qquad (j = 1, ..., p) \tag{1.24}$$

Obtener la estimación de máxima verosimilitud $\hat{\beta}$ requiere la resolución de un sistema de ecuaciones no lineales. Debido al gran número de observaciones utilizadas, se suelen utilizan métodos numéricos iterativos, para la obtención de dichas estimaciones.

A pesar de la utilidad en ciertos casos concretos de este modelo, en la práctica, como se verá posteriormente, el incumplimiento de algunas de las principales propiedades de la distribución de Poisson, en las carteras de seguros, producirá diferentes anomalías muestrales:

- Sobre-dispersión:
  Al hecho de que la varianza muestral sea superior a la media en los datos se le denomina "sobre-dispersión", ya que si la siniestralidad estuviera generada por una distribución de Poisson, la media debería ser igual a la varianza. Este efecto suele ser habitual en datos de carteras de seguros.

- Inflado de ceros:
  Esta anomalía se observa cuando los datos muestrales presentan una frecuencia más elevada para la ocurrencia de cero siniestros que la que se debería esperar si la muestra hubiera sido generada mediante una distribución de Poisson.

1.3.2. Modelo de Regresión Binomial Negativo

El problema de la sobre–dispersión lleva a buscar distribuciones que admitan este comportamiento, como es la Binomial Negativa (BN). Para ello se parte de un experimento con n repeticiones aleatorias e independientes que pueden resultar en éxito A o fracaso A^c. De cada una de las repeticiones se puede extraer una variable binaria, tal que $X_n = 1$ si se produce un éxito, con $p = P(X_n = 1)$ y $X_n = 0$ si se produce un fracaso, con $q = 1 - p = P(X_n = 0)$. Mediante la sucesiva repetición de estos ensayos quedará definida una sucesión de v.a. independientes $\{X_n\}_{n\in\mathcal{N}}$. La variable aleatoria Y_m representa el número de fracasos hasta obtener m éxitos y a la distribución de parámetros m y p que sigue se le denomina Binomial Negativa.

El soporte de Y_m será el conjunto de los enteros positivos, $\mathbf{Z}^+$. Para obtener su función de cuantía, hay que analizar el suceso $\{Y_m = y\}$, el cual indica que ha habido y fracasos, hasta obtener m éxitos.

$$\{Y_m = y\} = \{S_{m+y-1} = m-1, X_{m+y} = 1\} \tag{1.25}$$

Es decir, en el mismo ensayo en el que se ha obtenido el m-ésimo éxito se han realizado en total $m + y$ repeticiones del ensayo, momento en el que se concluyen las repeticiones. Por lo tanto, entre las $m + y - 1$ primeras repeticiones del mismo se han producido los $m - 1$ éxitos y los y fracasos.

Sea $S_{m+y-1} = X_1 + X_2 + ... + X_{m+y-1}$ la suma de las primeras $m+y-1$ v.a. independientes de la sucesión $\{X_n\}$, seguirá una distribución binomial de parámetros $(m + y - 1, p)$. Por otro lado, esta v.a. es independiente de X_{m+y}, por lo que la probabilidad del suceso anterior, es

$$\begin{aligned} P(Y_m = y) &= P(S_{m+y-1} = m-1)P(X_{m+y} = 1) = \\ &= \binom{m+y-1}{m-1} p^{m-1}q^y p = \\ &= \binom{m+y-1}{y} p^m q^y = \\ &= \binom{-m}{y} p^m(-q)^y. \end{aligned}$$

Este último binomio negativo es al que debe la distribución el nombre de binomial negativa.
La media y la varianza son respectivamente:

$$E(Y) = \frac{mq}{p} \quad y \quad V(Y) = \frac{mq}{p^2} \tag{1.26}$$

La distribución BN se introduce habitualmente en la literatura actuarial como una mixtura Poisson-Gamma. Fue derivada por primera vez como mixtura de una distribución de Poisson y una Gamma en el trabajo de (Greenwood y Yule, 1920). Las mixturas o funciones compuestas se fundamentan en que una variable aleatoria Y con función de probabilidad dependiente del parámetro ψ_0, tal que $P(y;\psi_0)$, el parámetro ψ sea a su vez una variable aleatoria con función de probabilidad $P(\psi)$, tal que $P(y|\psi = \psi_0)$. Esta distribución de Y incondicionada al valor de ψ, se obtiene directamente del teorema de la partición:

$$P(Y = y) = \sum_{\psi} P(y|\psi)P(\psi) \tag{1.27}$$

si Y y ψ son v.a. discretas y

$$f(y) = \int_{\psi} f(y|\psi) f(\psi) d\psi \tag{1.28}$$

si tanto Y como θ son v.a. continuas. Aunque en el caso actual, como se verá a continuación, se da una combinación v.a. discretas y continuas.

Sea Y una v.a. con distribución de Poisson de parámetro μ_i, donde a su vez este parámetro tiene una función de distribución de probabilidad $\gamma(\tau, \theta)$.

Es decir:

$$P(Y_i = y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{1.29}$$

y

$$f(\mu_i) = \frac{\tau^{\theta}}{\Gamma(\theta)}\mu_i^{\theta-1}e^{-\tau\mu_i}, \quad \mu_i \geq 0 \tag{1.30}$$

y $f(\mu_i) = 0$, en otro caso. Donde

$$\Gamma(\theta) = \int_0^{\infty} x^{\theta-1}e^{-x}dx. \tag{1.31}$$

Por lo tanto la distribución de Y incondicionada a μ_i es:

$$P(y_i) = \int_0^{\infty} P(y_i|\mu_i)f(\mu_i)d\mu = \frac{\tau^{\theta}}{\Gamma(\theta)y_i!}\int_0^{\infty} e^{-\mu_i(\tau+1)}\mu_i^{y_i+\theta-1}d\mu \tag{1.32}$$

Sabiendo que

$$\frac{(\tau+1)^{y_i+\theta}}{\Gamma(y_i+\theta)}\int_0^{\infty} \mu_i^{y_i+\theta-1}e^{-(\tau+1)\mu_i}d\mu = 1 \tag{1.33}$$

puesto que es la función de densidad de una v.a. $\gamma(\tau+1, y_i+\theta)$. En la expresión (1.32), si $y_i + \theta \in \mathbf{Z}^+$, :

$$P(y_i) = \frac{\tau^{\theta}}{\Gamma(\theta)y_i!}\frac{\Gamma(y_i+\theta)}{(\tau+1)^{y_i+\theta}} = \begin{pmatrix} \theta + y_i - 1 \\ y_i \end{pmatrix}\left(\frac{\tau}{\tau+1}\right)^{\theta}\left(\frac{1}{\tau+1}\right)^{y_i} \tag{1.34}$$

que corresponde a una función de cuantía de una v.a. BN de parámetros $m = \theta$, $p = \frac{\tau}{(\tau+1)}$.

Para facilitar la estimación del MLG se realiza una parametrización de (1.34) en función de la media μ_i de la v.a. Y_i.

$$q = 1 - p = 1 - \frac{\tau}{\tau+1} = \frac{\tau+1-\tau}{\tau+1} = \frac{1}{\tau+1} \tag{1.35}$$

$$E(Y_i) = \frac{mq}{p} = \frac{\theta \cdot \frac{1}{\tau+1}}{\frac{\tau}{\tau+1}} = \frac{\theta}{\tau} = \mu_i \Rightarrow \tau = \frac{\theta}{\mu_i} \tag{1.36}$$

$$\frac{\tau}{\tau+1} = \frac{\frac{\theta}{\mu_i}}{\frac{\theta}{\mu_i}+1} = \frac{\frac{\theta}{\mu_i}}{\frac{\theta+\mu_i}{\mu_i}} = \frac{\theta}{\theta+\mu_i} \tag{1.37}$$

$$\frac{1}{\tau+1} = \frac{1}{\frac{\theta}{\mu_i}+1} = \frac{\mu_i}{\theta+\mu_i} \tag{1.38}$$

$$f_{Y_i}(y_i; \mu_i, \theta) = \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y_i} \tag{1.39}$$

Se recurrirá para la estimación de sus parámetros de nuevo al método de máxima verosimilitud. La función de versimilitud obtenida a partir de la función (1.39) será:

$$\mathcal{L}(y;\mu,\theta)=\prod_{i=1}^{n} f_{Y_i}(y_i;\mu_i,\theta)=\prod_{i=1}^{n}\left[\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}\left(\frac{\mu_i}{\mu_i+\theta}\right)^{y_i}\right] \tag{1.40}$$

Tomando logaritmos se obtiene la función de log-verosimilitud, equivalente a la anterior:

$$log\mathcal{L}(y;\mu,\theta)=\sum_{i=1}^{n}\left[ln\Gamma(y_i+\theta)-\ln\Gamma(y_i+1)-\ln\Gamma(\theta)+\theta\ln\theta+y_i\ln\mu_i-(\theta+y_i)\ln(\mu_i+\theta)\right] \tag{1.41}$$

Tras obtener la función de log-verosimilitud, se introducirá la relación con los regresores mediante la función de enlace canónica, que en el caso de la Binomial Negativa es logarítmica y de forma $\mu_i = exp(\sum_{j=1}^{p} x_{ij}\beta_j + \xi_i)$, por lo tanto:

$$\begin{aligned} log\mathcal{L}(y;e^{X\beta+\xi},\theta)=&\sum_{i=1}^{n}\left[\ln\Gamma(y_i+\theta)-\ln\Gamma(y_i+1)-\ln\Gamma(\theta)+\theta\ln\theta+y_i\ln e^{X\beta+\xi}\right.\\ &\left.-(\theta+y_i)\ln(e^{X\beta+\xi}+\theta)\right] \end{aligned} \tag{1.42}$$

El valor de la log-verosimilitud se maximizará cuando para cada valor de j, las derivadas parciales de primer orden de la función de log-verosimilitud respecto de β_j y θ se igualen a 0

$$\frac{\partial log\mathcal{L}}{\partial\beta_j}=0, \qquad (j=1,...,p) \tag{1.43}$$

$$\frac{\partial log\mathcal{L}}{\partial\theta}=0 \tag{1.44}$$

Obtener las estimaciones de máxima verosimilitud $\hat{\beta}$ y $\hat{\theta}$ explícitamente requiere resolver un sistema de ecuaciones no lineales. En este caso se obtienen soluciones aproximadas mediante métodos numéricos.

1.4. Medidas de la bondad del ajuste de los MLG

Las medidas de la bondad del ajuste cuantifican en qué grado se ajusta el modelo estimado al conjunto de observaciones. Estas medidas podrán utilizarse tanto para estimar los parámetros del modelo en sí, como para comparar la precisión de distintos modelos y verificar que el modelo cumpla unas determinadas reglas.

1.4.1. Sobredispersión

Como ya se ha comentado con anterioridad, un fenómeno común en las distribuciones de siniestros de las carteras del seguro del automóvil es la sobre–dispersión, que recordamos aparece cuando el valor de la varianza es mayor que el valor de la media. Para comprobar su existencia se puede utilizar un test como el propuesto por (Cameron, 1985):

$$\begin{aligned} H_0:&\quad E(Y)=\mu \quad \text{y} \quad Var(Y)=\mu,\\ H_a:&\quad E(Y)=\mu \quad \text{y} \quad Var(Y)=\mu+\alpha\cdot f(\mu) \quad \forall \quad \alpha>0 \end{aligned} \tag{1.45}$$

Como puede observarse, el test contrasta la hipótesis nula de que la media es igual a la varianza frente a la hipótesis alternativa de que la varianza es la media más un múltiplo no negativo de la media.

1.4.2. Error cuadrático medio

El error cuadrático medio (ECM), es una medida de la bondad del ajuste que muestra la desviación de la variable estimada sobre la variable observada como el promedio de los errores al cuadrado:

$$ECM = \frac{\sum_{i=1}^{n}(Y - \hat{Y})^2}{n} \tag{1.46}$$

Por lo tanto, un menor valor del ECM indica un mejor ajuste del modelo a los datos.

1.4.3. Desviación

La Desviación, también llamada Devianza o por su nombre en inglés Deviance es un estadístico de la bondad del ajuste, aunque indica justo lo contrario, ya que un valor superior indica un peor ajuste. Se trata de una generalización de la Suma de Cuadrados Residuales de la estimación por Mínimos Cuadrados Ordinarios a los modelos estimados por el método de Máxima Verosimilitud.

Esta medida parte de la base de que para observar la eficacia del modelo se debe comparar con un modelo más amplio, que contenga el máximo número de parámetros sin repetición posible, llamado modelo saturado. Este modelo $M_{saturado}$ será muy similar al modelo que se trata de medir $M_{propuesto}$, ya que contendrá la misma distribución y la misma función de enlace. Además, el valor de su función de verosimilitud $\mathcal{L}(M_{saturado})$ será mayor que ninguna otra función para estos datos, ya que proporcionará la más completa y ajustada descripción de los datos. Al compararlo con el valor de la función de verosimilitud del modelo propuesto $\mathcal{L}(M_{propuesto})$ se obtendrá el siguiente ratio de verosimilitud:

$$\lambda = \frac{\mathcal{L}(M_{saturado})}{\mathcal{L}(M_{propuesto})} \tag{1.47}$$

En la práctica, se propone en (Nelder y Wedderburn, 1972) utilizar el doble del logaritmo de su diferencia, a lo que llaman Desviación.

$$ln(\lambda) = \mathcal{L}(M_{saturado}) - \mathcal{L}(M_{propuesto}) \tag{1.48}$$

$$D = 2[\mathcal{L}(M_{saturado}) - \mathcal{L}(M_{propuesto})] \tag{1.49}$$

1.4.4. Criterio Informativo de Akaike

El Criterio Informativo de Akaike (AIC) es otra medida de la bondad del ajuste de gran utilidad a la hora de comparar distintos modelos, ya que no sólo tiene en cuenta el ajuste del modelo a los datos, si no que introduce además un componente de penalización sobre el incremento de parámetros del modelo. Esta penalización trata de evitar el sobreajuste que podría producirse al introducir un exceso de parámetros en el modelo.

Sea $\mathcal{L}(M_{propuesto})$ el máximo valor de la función de verosimilitud del modelo propuesto y p el número de parámetros del mismo modelo, el estadístico AIC será:

$$AIC = -2ln(\mathcal{L}(M_{propuesto})) + 2p \tag{1.50}$$

Por lo tanto, un valor menor del AIC indicará en principio que se trata de un modelo de mayor calidad que otro con un AIC mayor. Esta selección de modelos óptimos no solo es válida cuando se trata de modelos anidados, sino que puede ser extendida a cualesquiera modelos, incluso no anidados, ver (Lewis *et al.*, 2011).

1.4.5. Validación Cruzada

La Validación Cruzada tiene como objetivo la validación y comparación de modelos juzgando la independencia de éstos sobre los datos y así evitar las negativas consecuencias del sobre-ajuste. Existen diversos métodos para implementar este sistema, aunque todos ellos se fundamentan en la idea de hacer varias escisiones entre datos de entrenamiento del modelo y datos de validación de un modo sistemático o aleatorio.

En el presente trabajo se utilizará por su buen equilibrio entre tiempo de computación y efectividad la Validación Cruzada de $k - iteraciones$, concretamente de $k = 100$ iteraciones. En cada una de las 100 iteraciones se separa la muestra de datos aleatoriamente pero sin repetición en 99 % datos de entrenamiento y 1 % datos de validación, por lo que al final todos los registros habrán pasado por los dos subconjuntos. Además, se calculará en cada una de ellas el Error Cuadrático Medio.

Una distribución aproximadamente normal y sin fuertes oscilaciones de los Errores Cuadráticos Medios indicarán una buena independencia del modelo sobre los datos.

1.5. Métodos de selección de variables

El objetivo de este proceso es incluir en el modelo el mínimo número de variables necesario para realizar una buena explicación de los datos, ya que esto aporta muchos beneficios, entre otros: previene el sobre-ajuste, reduce la dificultad de estimación del modelo,... Los métodos analizados se basan en el AIC para contrastar las distintas hipótesis de alteración de las variables de modelo.

1.5.1. Selección Progresiva

La selección Progresiva o Forward parte del modelo sin ninguna variable y contrasta la hipótesis nula de que el modelo no puede mejorar su AIC frente a la hipótesis alternativa de que añadir una variable mejorará el AIC. En caso de que más de una variable mejore el modelo se añadirá en primer lugar la que más valor añada, para luego repetir el proceso hasta que ninguna variable mejore el modelo.

1.5.2. Selección Regresiva

La selección Regresiva o Backward es el caso contrario, parte del modelo con todas las variables y contrasta la hipótesis nula de que el modelo no puede mejorar su AIC frente a la hipótesis alternativa de que eliminar una variable mejorará el AIC. En caso de que más de una variable al eliminarla mejore el modelo se eliminará en primer lugar la que más valor añada, para luego repetir el proceso hasta que eliminando variables no mejore el modelo.

1.5.3. Selección Escalonada

La selección Escalonada o Stepwise es una combinación de ambos modelos, parte del modelo con todas las variables y contrasta la hipótesis nula de que el modelo no puede mejorar su AIC frente a la hipótesis alternativa de que eliminar o añadir una variable mejorará el AIC. En caso de que más de una variable al eliminarla o añadirla mejore el modelo se eliminará o añadirá en primer lugar la que más valor añada, para luego repetir el proceso hasta que eliminando o añadiendo variables no mejore el AIC del modelo.

1.6. Cálculo de la prima *a priori*

Una vez estimado el valor de la media del número de reclamaciones μ_i de cada póliza, se obtiene el valor del riesgo relativo que representan cada una de ellas sobre el conjunto de la cartera.

$$r_i = \frac{\mu_i}{\sum_{i=1}^{n} \mu_i} \tag{1.51}$$

Por último, se obtiene el valor de la prima *a priori* de la póliza i como la relación entre el valor del riesgo relativo, r_i, por la cuantía total de las reclamaciones realizadas Q en el conjunto de la cartera:

$$P_{0,i} = r_i \cdot Q \tag{1.52}$$

Capítulo 2

Tarificación a posteriori, sistemas de ajuste de la tarifa

Las compañías de seguros sufren una gran incertidumbre acerca de sus nuevos asegurados, ya que desconocen muchas variables difícilmente cuantificables que pueden afectar a la siniestralidad de estos. Recientemente se han puesto en marcha bases de datos comunes para tratar de minimizar la incertidumbre, al menos, entre los que no sean asegurados noveles, donde se comunican de unas compañías a otras las siniestralidades sufridas por los asegurados. La tarificación *a posteriori* trata de adaptar al máximo la tarifa del seguro a los riesgos sobre los que se ofrece la cobertura. Para ello deberá tratar de conocer el comportamiento del conductor.

En el presente estudio se analizarán tres de las técnicas de tarificación *a posteriori* más comunes, la utilización de cada una de ellas en cada momento responde a las necesidades históricas de las compañías aseguradoras, tanto de abarcar la mayor parte del mercado posible sin un drástico aumento de los costes, como de poder ofrecer unas tarifas competitivas adaptadas a los riesgos individuales que atraigan y consoliden a los clientes.

Será también una práctica habitual en el mercado la utilización conjunta de varias de estas técnicas, entre las que se encuentran:

- Sistema Bonus–Malus
- Franquicia
- Sistema Basado en el Uso

2.1. Sistema de Tarificación Bonus-Malus

El *sistema de tarificación Bonus-Malus*, en adelante *SBM*, como modelo de clasificación de riesgos, ha sido ampliamente utilizado en el mercado asegurador del automóvil en toda Europa y en gran parte del mundo a partir de los años 60 (Pérez Sánchez, 2011).

Debido a las particularidades regulatorias de cada país, el *SBM* se ha aplicado de un modo ligeramente distinto en la mayoría de ellos, aunque algunos países como España no ha tenido fuertes regulaciones en torno a el. En otros países como Bélgica y Alemania se encuentra fuertemente regulado.

En esencia, el *SBM* bonifica a los conductores que no hayan efectuado una reclamación en un determinado espacio de tiempo con un descuento en la prima (*Bonus*) y a su vez penaliza a los conductores que hayan efectuado reclamaciones de un modo proporcional al número de éstas con un sobrecoste de la prima (*Malus*). De este modo se trata de conseguir un sistema de

tarificación más justo, en el que los asegurados tendrán la tarifa de su prima adaptada al riesgo que comporten.

Siguiendo a (Lemaire, 1995), una compañía estará utilizando el *SBM* cuando:

- Los asegurados se reparten entre un número finito de clases $C_1, .., C_s$ (sea s el número máximo de clases), de modo que el monto de la prima anual depende solo de la clase en la que se encuentra.
- La clase a la que pertenece un asegurado en un determinado periodo de tiempo (habitualmente un año) depende exclusivamente de la clase en la que se encontraba en el periodo precedente y del número de reclamaciones hechas durante ese periodo.

Este sistema está determinado por tres elementos:

- La prima relativa.
- La clase inicial C_0 que les es asignada a los nuevos asegurados, la cual, determina su prima *a priori*.
- Las reglas de transición de una clase a otra cuando el número de reclamaciones sea conocido.

2.1.1. Prima relativa

El objetivo de este apartado es establecer una prima relativa a la prima *a priori* mediante un sistema de bonus–malus óptimo, para ello se utilizará el análisis Bayesiano. Partiendo de la hipótesis de que el número de reclamaciones se distribuye como una distribución binomial negativa (Lemaire, 1995, Bermudez *et al.*, 2001), se puede aprovechar la ventaja de la estabilidad de su función estructural, que como se vio en el Capítulo 1, se trata de una distribución Gamma. Esta propiedad se cumple gracias a que la distribución Gamma es una distribución conjugada de la distribución de Poisson, lo que garantiza que la distribución *a posteriori* será de la misma familia que la distribución *a priori* tras aplicar el análisis Bayesiano (Lemaire, 1995), como se muestra a continuación.

Sea el número de reclamaciones realizadas por un asegurado en los últimos t años,

$$t\bar{y} = \sum_{j=1}^{t} y_j \,, \tag{2.1}$$

se obtiene la función de verosimilitud de un proceso de Poisson repetido durante esos mismos periodos, tal que,

$$f(y_1, y_2, ..., y_t|\mu) = f(y_1|\mu) \cdot f(y_2|\mu) \cdots f(y_t|\mu)$$

$$= \frac{e^{-\mu}\mu^{y_1}}{y_1!} \cdot \frac{e^{-\mu}\mu^{y_2}}{y_2!} \cdots \frac{e^{-\mu}\mu^{y_t}}{y_t!} = \frac{e^{-t\mu}\mu^{t\bar{y}}}{\prod_{j=1}^{t}(y_j!)}. \tag{2.2}$$

Por el teorema de Bayes, multiplicando la función de distribución de probabilidad *a priori* γ por la función de verosimilitud de la distribución de Poisson y dividiendo por la función de normalización, que garantiza que la probabilidad total sea igual a la unidad, se obtiene la distribución *a posteriori* de μ:

$$u(\mu|y_1, y_2, ..., y_t) = \frac{f(y_1, y_2, ..., y_t|\mu)u(\mu)}{\displaystyle\int_0^\infty [f(y_1, y_2, ..., y_t|\mu)u(\mu)]\, d\mu}$$

$$= \frac{\left(\frac{e^{-t\mu}\mu^k}{\prod_{j=1}^{t}(y_j!)}\right)\cdot\left(\frac{\tau^\theta}{\Gamma(\theta)}\mu^{\theta-1}e^{-\tau\mu}\right)}{\int_0^\infty\left[\left(\frac{e^{-t\mu}\mu^k}{\prod_{j=1}^{t}(y_j!)}\right)\cdot\left(\frac{\tau^\theta}{\Gamma(\theta)}\mu^{\theta-1}e^{-\tau\mu}\right)\right]d\mu} = \frac{\mu^{k+\theta-1}e^{-(t+\tau)\mu}}{\int_0^\infty\left[\mu^{k+\theta-1}e^{-(t+\tau)\mu}\right]d\mu} \tag{2.3}$$

$$= \frac{(\tau+t)^{\theta+k}\mu^{k+\theta-1}e^{-(t+\tau)\mu}}{\int_0^\infty\left[[\mu(\theta+t)]^{k+\theta-1}e^{-(t+\tau)\mu}\right]d[\mu(\theta+t)]} = \frac{(\tau+t)^{\theta+k}\mu^{k+\theta-1}e^{-(t+\tau)\mu}}{\Gamma(\theta+k)}$$

Como vemos, se trata de la función de densidad de una distribución Gamma de parámetros $\gamma(\theta+t\bar{y},\tau+t)$. Por lo tanto, el valor de la media del número de reclamaciones que realizará un asegurado con un historial de reclamaciones $(y_1,y_2,...,y_t)$ en el periodo $t+1$, μ_{t+1}, será:

$$\mu_{t+1}(y_1,y_2,...,y_t) = \frac{\theta+t\bar{y}}{\tau+t} \tag{2.4}$$

Como puede observase, la actualización del valor, únicamente necesita de la suma de los valores $t\bar{y}$ y t en los parámetros de una distribución γ, respectivamente.

La obtención de la prima relativa, se realiza habitualmente mediante el principio del valor esperado de la prima, tal que,

$$P_{t+1}(y_1,y_2,...,y_t) = (1+\alpha)\cdot\mu_{t+1}(y_1,y_2,...,y_t) = (1+\alpha)\frac{\theta+t\bar{y}}{\tau+t}\,, \tag{2.5}$$

donde α es el recargo de seguridad introducido por el asegurador. En el caso de que no se efectúe este recargo, se tratará del principio de prima neta.

La prima Bonus–Malus (en adelante BM) *a posteriori* se define como la proporción de la prima neta en el periodo $t+1$ entre la prima neta en el periodo t:

$$P_{BM}(y_1,y_2,...,y_t) = \frac{\mu_{t+1}}{\mu_t} = \frac{\frac{\theta+t\bar{y}}{\tau+t}}{\frac{\theta+0}{\tau+0}} = \frac{\tau(\theta+t\bar{y})}{\theta(\tau+t)} \tag{2.6}$$

Además, se puede observar que la prima BM *a posteriori* cuando el número de siniestros crece, ésto es, cuando $t\bar{y}\to\infty$ tiende a ∞, para un periodo de tiempo fijo, tiende a infinito. Por el contrario, cuando $t\to\infty$ para un número de siniestros fijo, la prima BM tiende a 0.

Por último, se obtiene *a posteriori* la prima neta a pagar por un a póliza i, como el producto de su prima calculada *a priori* en $t=0$, $P_{0,i}$, obtenida en la ecuación (1.52) por la prima BM $P_{BM,i}(y_{1i},y_{2i},...,y_{ti})$:

$$P_{t+1,i}(y_{1i},y_{2i},...,y_{ti}) = P_{0,i}\cdot\frac{\tau(\theta+t\bar{y}_i)}{\theta(\tau+t)} = P_{0,i}\cdot P_{BM,i}(y_{1i},y_{2i},...,y_{ti}) \tag{2.7}$$

Algunas propiedades de este sistema bonus–malus óptimo propuesto en (Lemaire, 1995) son las siguientes:

1. Se trata de un sistema "justo", en sentido Bayesiano. Esto es, la prima a pagar por cada asegurado, se obtiene mediante el teorema de Bayes, en función de "todas" las reclamaciones pasadas observadas en la cartera de asegurados.

2. Es un sistema equilibrado financieramente, es decir, la suma de las primas pagadas por el conjunto de la cartera de asegurados se mantiene constante en todos los periodos. Aunque esta propiedad nunca se cumple en la práctica.

3. La prima BM únicamente depende del número de reclamaciones realizadas.

4. En el periodo inicial $t=0$, todos los asegurados pagan únicamente su prima *a priori*.

2.1.2. Clase inicial

La elección de la clase inicial depende principalmente de la existencia de registros de siniestralidad del asegurado, ya sea por parte de la compañía o por parte de algún ente que aglutine los registros de siniestralidad de los tomadores de seguros de distintas aseguradoras, como ocurre en España con el Fichero Histórico de Seguros de Automóviles (SINCO) o en Reino Unido el Claims and Underwriting Exchange (CUE). En caso de no disponerse de datos históricos del asegurado se le asignará la clase "neutra", es decir, no se le aplicará ni *bonus* ni *malus*, únicamente la prima *a priori*.

2.1.3. Reglas de transición

Las reglas de transición pueden ser muy fácilmente reconocidas en el sistema BM propuesto en (Lemaire, 1995) en el que se basa el presente trabajo, ya que el coeficiente de ajuste de la prima únicamente dependerá del horizonte temporal disponible de la siniestralidad del asegurado, del horizonte temporal que abarque el sistema BM en su diseño, así como del número de siniestros reclamados en ese horizonte temporal. El establecimiento de un número de años máximo en la tabla del sistema BM ejercerá un efecto de "recuperación" u "olvido" de la siniestralidad reclamada.

2.2. Sistema de Tarificación de Franquicia

Los sistemas de franquicia deducible han formado parte del mercado asegurador durante mucho tiempo, pero generalmente como complementario a otros sistemas de tarificación y especialmente en las garantías de daños propios. El concepto principal es el de compartir el riesgo entre el asegurado y el asegurador. La franquicia será el monto económico cuya pérdida será únicamente soportada por el patrimonio del asegurado. Por lo tanto, si un siniestro no alcanza la cuantía de la franquicia, será el asegurado el que haga frente en su totalidad a las pérdidas originadas. Por otro lado, en caso de que valor del siniestro supere el valor de la franquicia, el asegurado se hará cargo de una cuantía económica igual a la franquicia y el asegurador del resto.

La contratación de este sistema es de gran interés para el asegurador ya que de este modo el asegurado será corresponsable de hacer frente a las pérdidas económicas y por lo tanto, el número de siniestros declarados descenderá, además reducirá los gastos de gestión de siniestros de la compañía al reducir su número. Su aplicación es de especial interés en modelos estadísticos de predicción de la severidad de las reclamaciones de siniestros.

2.2.1. Sistema combinado de Gran Franquicia y Crédito

El Sistema combinado de Gran Franquicia y Crédito surge como alternativa al Sistema Bonus–Malus de la mano de (Holtan, 1994) y como crítica a dos de sus características:

- El monto de las reclamaciones no es utilizado como variable en la retarificación.
- El asegurado podrá abandonar en cualquier momento la compañía consiguiendo bajo ciertas circunstancias evadir la penalización económica en la tarifa.

Se trata de una idea fácilmente asimilable por la cada vez más habitual figura de "bancaseguros", entendiendo este concepto como una combinación de entidades bancarias y aseguradoras.

La idea es fijar una gran franquicia (en torno a 3000€) y en caso de siniestro concederle al asegurado un crédito para hacer frente a esa franquicia, de ese modo, la cuota del crédito hará la función de Malus del sistema anterior.

2.3. Sistema de Tarificación Basados en el Uso

El seguro basado en el uso (SBU), ha sido una idea que ha despertado el interés de las compañías aseguradoras desde hace mucho tiempo, pero su aplicación no ha sido completamente satisfactoria puesto que la supervisión y rigurosidad de la supervisión recae prácticamente por completo sobre el asegurado. La entrada en el mercado de los sistemas telemáticos, sin embargo, no se ha extendido hasta los últimos años, especialmente en EEUU. De acuerdo a (Alcañiz Zanón, 2014) se trata de una reciente innovación de las aseguradoras de automóviles que trata de ajustar al máximo el comportamiento en la conducción con la tarifa aplicada. Se empieza a comercializar como una opción para jóvenes conductores con el objetivo de que puedan demostrar una conducción de baja siniestralidad y así reducir las altas tarifas que son comunes en este colectivo. Se pueden distinguir distintos tipos de SBU:

- Sistemas de kilometraje estimado por el asegurado (utilizado como parte de la tarificación *a priori).*
- Sistemas de captación telemática del comportamiento.

En los sistemas de captación telemática del comportamiento el kilometraje recorrido y/o el comportamiento en la conducción son captados y enviados a la compañía aseguradora, idealmente en tiempo real. Estos sistemas captan las mediciones de interés como pueden ser: distancia recorrida, franja horaria, localización del vehículo, estilo de conducción relajado/agresivo,... El nivel de datos recogidos depende tanto de la tecnología utilizada como de la disposición del tomador del seguro a compartir información.

La obtención de la tarifa seguirá el mismo proceso de la tarificación *a priori*, generalmente mediante técnicas de MLG.

Capítulo 3

Un caso práctico de tarificación de responsabilidad civil en automóviles

En el presente caso práctico se realizará la tarificación de la garantía de responsabilidad civil del seguro de automóviles para una cartera real de pólizas. La tarificación constará de dos fases, una *a priori* mediante MLG y otra *a posteriori* mediante el Sistema Bonus–Malus.

Debido al gran tamaño de la muestra con la que se trabaja se requerirá la utilización de un software para la realización de las diferentes operaciones estadísticas, así como para el tratamiento de los datos. Existen diversas opciones tanto comerciales (SAS, Matlab, SPSS, Stata,...), como de software libre (R, Octave,...). En esta ocasión se ha seleccionado R tanto por tratarse de software libre y multiplataforma, como por la amplia documentación disponible, con el objetivo, además, de que el estudio pueda ser fácilmente reproducible. Ver el Anexo 1 para las instrucciones de instalación y de descarga de paquetes necesarios de R, así como una guía básica. En el Anexo 2 se encuentran las líneas de código utilizadas para la realización del caso práctico.

3.1. Datos

En este apartado, en primer lugar, se identificarán los datos, a continuación se depurarán y se transformarán, y por último se analizarán.

3.1.1. Identificación

Los datos utilizados en este caso práctico son una muestra de 677.991 pólizas de seguro de responsabilidad civil de automóvil en un espacio temporal de un año, llamada *freMTPL2, French Motor Third-Part Liability datasets*, proporcionadas por un asegurador francés desconocido para el libro "Computational Actuarial Science with R" editado por Arthur Charpentier e incluidos ahora dentro del paquete de R "CASdatasets".

La Base de Datos (en adelante, BBDD) *freMTPL2freq* consta de 11 variables cuyo campo clave o identificador único es la variable "IDpol", código de identificación de cada póliza.

La última parte del nombre de la BBDD *freMTPL2**freq*** indica que el objetivo con el que se ha creado es el análisis de la frecuencia de las reclamaciones de los siniestros por parte de los asegurados, por lo que es de esperar que se hayan incluido variables de interés de acuerdo a la experiencia del asegurador. A continuación, se muestra en el Cuadro 2 una relación de las variables incluidas junto con una breve descripción de las mismas.

Cuadro 2

Diccionario de datos de la BBDD freMTPL2freq

Variable	Descripción
IDpol	Identificador de la póliza
ClaimNb	Número de reclamaciones durante el periodo de exposición
Exposure	Tiempo de exposición
Area	Código de área
VehPower	Potencia del vehículo (por categorías)
VehAge	Edad del vehículo (en años)
DrivAge	Edad del conductor
BonusMalus	Bonus–Malus entre 50 y 350, $<$100 es bonus
VehBrand	Marca del vehículo
VehGas	Tipo de combustible
Density	Densidad de habitantes en lugar de residencia
Region	Región de procedencia de la póliza (clasificación estándar francesa)

Dado el anterior conjunto de variables se focalizará el análisis en la inferencia del número de siniestros a partir de las observaciones de las variables "ClaimNb" y "Exposure" y su relación con el resto de variables a excepción de "BonusMalus" y "Region".

La variable "BonusMalus" será omitida para evitar una doble penalización al realizar la tarificación *a posteri*. En un caso real se optaría por obtener la información de la siniestralidad anterior de un ente como SINCO tal y como se comentó en el Capítulo 2 o simplemente de la propia experiencia del asegurador, para así aplicar la retarificación mediante Bonus–Malus teniendo en cuenta el número de reclamaciones hechas en los años previos a la puesta en marcha del modelo actual.

La variable "Region" será también descartada debido a su gran similitud con la variable "Area", siendo la primera mucho más específica al contar con 22 regiones frente a 6 áreas y aportar una mayor complejidad a la segmentación en grupos homogéneos de riesgo.

La BBDD *freMTPL2sev* contiene 26.639 reclamaciones realizadas por asegurados incluidos en la BBDD *freMTPL2freq*, en este caso consta únicamente de 2 variables. La última parte del nombre de la BBDD *freMTPL2**sev*** indica que el objetivo con el que se ha creado es el análisis de la severidad o coste económico de las reclamaciones hechas por parte de los asegurados relacionándola con la BBDD de anterior. En este caso el campo "IDpol" podrá aparecer en varios registros en caso de que una misma póliza haya hecho más de una reclamación en el año de estudio. A continuación, se muestra en el Cuadro 3 las variables incluidas junto con una breve descripción de estas.

Cuadro 3

Diccionario de la BBDD freMTPL2sev

Variable	Descripción
IDpol	Identificador de la póliza
ClaimAmount	Coste de la reclamación

3.1.2. Depuración y transformación

El objetivo de este paso es realizar una primera revisión y tratamiento de los datos con el objetivo de disponer posteriormente una versión de la BBDD fiable, libre de errores y de observaciones atípicas, en la medida de lo posible. En primer lugar, se transformarán las BBDD en tablas con nombres más simples para un manejo más ágil (Anexo 2 – C1) y (Anexo 2 – C2). A continuación, se procederá a revisar los datos de *freMTPL2freq*, comenzando con la variable "IDpol". Se trata del número de póliza en un año, por lo tanto, no debería haber números repetidos y la frecuencia máxima de los valores de la columna "IDpol" debería de ser 1.

IDpol

```
max(table(A[,IDpol]))
[1] 1
```

El resultado coincide con el valor esperado, por lo tanto, en este aspecto los datos tienen sentido. En la siguiente variable, número de reclamaciones "ClaimNB", se buscarán valores atípicos que pudiera contener. Para ello, se crea la una tabla de frecuencias del número de siniestros (Anexo 2 – C3) como muestra el Cuadro 4:

Cuadro 4

Frecuencia del número de reclamaciones (ClaimNb)

Reclamaciones	Número de pólizas (frecuencias)
0	643.953
1	32.178
2	1.784
3	82
4	7
5	2
6	1
8	1
9	1
11	3
16	1

Efectivamente, se observan valores atípicos extremos. Debido a su poco peso sobre el volumen total de pólizas. Con la idea de mantener una forma de distribución decreciente por la derecha, se eliminarán de ambas BBDD las pólizas con más de 6 reclamaciones (Anexo 2 – C4).

A continuación, se muestra el dominio de la variable "Exposure":

Exposure

```
min(A$Exposure);max(A$Exposure)
[1] 0.00273224
[1] 2.01
```

El valor mínimo coincide con la fracción de un día sobre un año por lo que parece correcto. Sin embargo, el valor máximo es de más de dos años. De acuerdo a la descripción de la BBDD se

trata de las observaciones de un año, por lo que únicamente podrían tratarse de seguros bianuales, problemas de cobro o simplemente errores. Por consiguiente, los registros con exposición mayor a 1 año serán eliminados del estudio (Anexo 2 – C5).

Se analiza a continuación como ha repercutido la depuración en el número de registros de la BBDD.

Resultado depuración BBDD freMTPL2freq

```
1-(nrow(A)/nrow(freMTPL2freq))
[1] 0.002249219
```

Porcentualmente la pérdida de registros es pequeña, de solo el 0,22 %, por lo que se puede determinar que la pequeña variación no tendrá gran repercusión en las conclusiones que se extraigan.

Se realiza a continuación el análisis de la BBDD *freMTPL2sev*. Será de especial interés que el dominio del valor de los siniestros no muestre cifras anormales.

ClaimAmount

```
min(B$ClaimAmount);max(B$ClaimAmount)
[1] 1
[1] 4075401
```

Los valores observados aparentan ser normales, el máximo es realmente grande, pero puede tratarse de un siniestro con varios vehículos involucrados, grandes indemnizaciones por fallecimiento,...

Para obtener unos registros de valor de siniestros coherente con las pólizas vivas se eliminarán los registros de la BBDD *freMTPL2sev* cuyo "IDpol" no esté presente en la BBDD *freMTPL2freq* y además se le acoplarán el resto de variables disponibles (Anexo 2 – C6).

Resultado depuración BBDD freMTPL2sev

```
1-(nrow(B)/nrow(freMTPL2sev))
[1] 0.007320095
```

Porcentualmente la pérdida de registros en esta BBDD también es pequeña, de solo el 0,73 %, por lo que se puede determinar igualmente que la pequeña variación no tendrá gran repercusión en las conclusiones que se extraigan.

De la depuración realizada y comentada en esta sección, se ha reducido el número de observaciones de la muestra, pasando a ser de 676.783.

3.1.3. Análisis descriptivo univariante y adaptación

Se profundizará ahora en el estudio de los datos mediante un análisis estadístico básico. El análisis descriptivo univariante tiene como objetivo describir, caracterizar y extraer conclusiones individuales sobre cada una de las variables de una muestra de datos, en este caso los asegurados de una compañía de seguros. Esto se realizará mediante el cálculo de los estadísticos más útiles dependiendo del tipo de variable de que se trate, así como de las representaciones gráficas tratando de realizar una simplificación de ellas.

Número de reclamaciones

La primera variable a analizar será el número de reclamaciones "ClaimNB". Se trata de una variable cuantitativa compuesta por números enteros no–negativos cuyos estadísticos principales son (Anexo 2–C7):

Cuadro 5

Estadísticos descriptivos del número de reclamaciones

Polizas	Media	Varianza	Coef. Asimetría	Curtosis
676.783	0,05316	0,05656	4,75661	28,52091

Para reflejar comparativamente el volumen de los asegurados agrupados por el número de reclamaciones que hayan comunicado se utilizará el siguiente gráfico de barras en la Figura 1 (Anexo 2–C8):

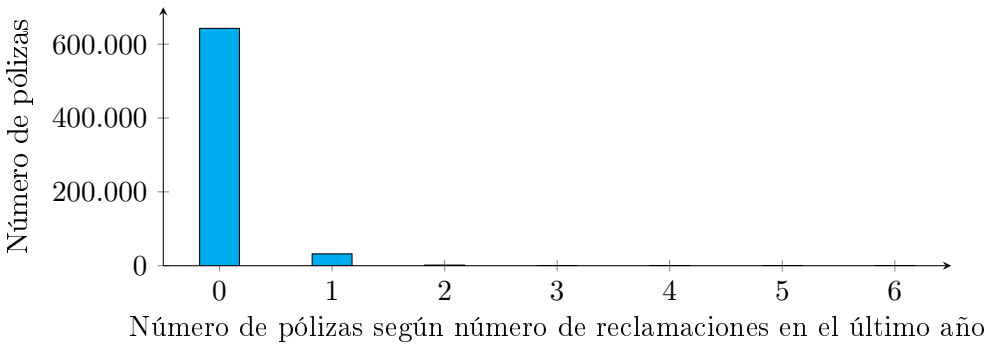

Figura 1: Frecuencia del número de reclamaciones. Elaboración propia.

El valor ($\approx 0,05$) del número medio de reclamaciones obtenido en el Cuadro 5 indica el número de siniestros medio que un asegurado puede tener durante el tiempo de exposición medio (en torno al medio año como se verá en el siguiente apartado), por lo tanto, en torno a 1 de cada 20 asegurados realizará una reclamación durante el tiempo de exposición medio. Es de especial interés la relación entre la media y la varianza, esta última al ser mayor que la media, indica un claro indicio de sobre–dispersión en la distribución de la variable aleatoria. El coeficiente de asimetría con un valor positivo indica que existe asimetría positiva con respecto a la media y qué por lo tanto, la cola larga de la distribución estará sobre valores superiores a la media. La curtosis o apuntamiento indica la cantidad de datos que se sitúan en torno a la media, por lo tanto. Un valor superior a 0 indica que se trata de una distribución leptocúrtica, con la mayoría de los datos en cercanos a la media y con un apuntamiento superior al de una distribución normal.

Exposición

La variable tiempo de exposición, "Exposure", es una variable cuantitativa discreta conformada por valores de fraccionados (nº de días de duración entre 365) calculados para cada una de las pólizas. Por lo que se asemeja a una variable continua. La media aritmética de esta variable se calcula,

$$\overline{Exposure} = \frac{\sum_{i=1}^{676783} Exposure_i = 357134{,}1}{N = 676783} = 0{,}5276937. \tag{3.1}$$

Este tiempo medio de exposición ($\approx 0,53$) que resulta ser ligeramente superior al medio año, indica que la contratación de las pólizas se ha realizado de un modo uniforme a lo largo del año

y que los datos suministrados corresponden al mismo instante de tiempo, donde el tiempo entre la realización del cobro y el momento de grabación de los datos diferirá para cada póliza. Para el actual proceso de tarificación no será necesario ahondar en el análisis de este concepto ya que la fijación de la tarifa será anual independientemente del momento de pago de la prima.

Área de residencia

La variable del área de residencia del asegurado "Area", es una variable cualitativa conformada por letras de la A a la F. Tras agrupar las pólizas por cada una de las mismas en función del número de reclamaciones y del tiempo de exposición, se obtiene la frecuencia anual del número de reclamaciones hechas, condicionada al área (Anexo 2–C9):

$$\mu_{Area=i} = E(ClaimNb|Area = Area_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}},\ i = A, B, ..., F\ ,\ j = 1, 2, ..., 676783\ , \tag{3.2}$$

donde $\mu_{Area=i}$ es el número medio anual de reclamaciones en el área $i = (A, B, ..., F)$. En la Figura 2 puede observarse la frecuencia con la que un asegurado realiza una reclamación dependiendo del área donde resida:

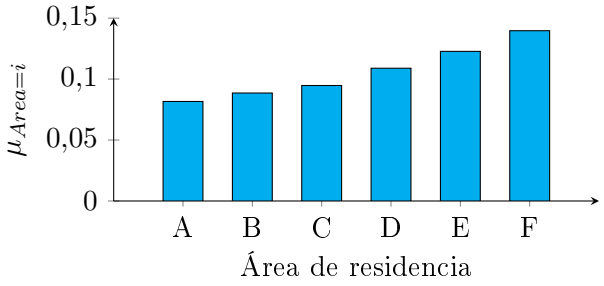

Figura 2: Distribución del número medio de reclamaciones por área de residencia. Elaboración propia.

Como puede observarse en la Figura 2, la segmentación que realizó la compañía de la que proceden los datos sigue un orden incremental en el riesgo de realizar una reclamación en cada área. Por lo tanto, con el único objetivo de simplificar los valores de la variable, y para conservar la tendencia creciente del riesgo, se hace una reclasificación agrupando en torno a los grupos 0, 1 y 2, las áreas incluidas en el primer tercio, el segundo tercio y el tercer tercio del rango de frecuencias condicionadas obtenido en (3.2) respectivamente (Anexo 2–C10). Sea el rango de las frecuencias condicionadas de la variable,

$$Rango = max\{\mu_i\} - min\{\mu_i\}, \tag{3.3}$$

se realiza la reclasificación, tal que,

$$Grupo\,0 : Area_i \in \left[min\{\mu_i\}; min\{\mu_i\} + \frac{Rango}{3}\right], \tag{3.4}$$

$$Grupo\,1 : Area_i \in \left[min\{\mu_i\} + \frac{Rango}{3}; min\{\mu_i\} + \frac{2 \cdot Rango}{3}\right], \tag{3.5}$$

$$Grupo\,2 : Area_i \in \left[min\{\mu_i\} + \frac{2 \cdot Rango}{3}; min\{\mu_i\} + Rango\right]. \tag{3.6}$$

Los resultados de este proceso, así como la agrupación de las distintas áreas que da lugar a la nueva clasificación reflejada en la variable "Area_c" se muestra en el Cuadro 6.

Cuadro 6

Reclasificación de variable "Area"

Area	Intervalo (Freq.Recl.)	Area_c	Pólizas	Proporción
A,B,C	[0,08162;0,10096]	0	370.734	54,8 %
D	[0,10096;0,12031]	1	151.304	22,3 %
E,F	[0,12031;0,13965]	2	154.745	22,9 %

Potencia del vehículo

La variable potencia del vehículo, "VehPower", es una variable cualitativa conformada por números del 4 al 15. Representando para cada uno de estos valores, el número de reclamaciones medio anual (Anexo 2–C11), tal que,

$$\mu_{VehPower=i} = E(ClaimNb|VehPower_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}}, \quad i = 4, 5, ..., 15, j = 1, 2, ..., 676783, \tag{3.7}$$

donde $\mu_{VehPower=i}$ es el número medio anual de reclamaciones por grupo de potencia $i = (4, 5, ..., 15)$. En la Figura 3, puede observarse la frecuencia del número de reclamaciones anual según la potencia del vehículo.

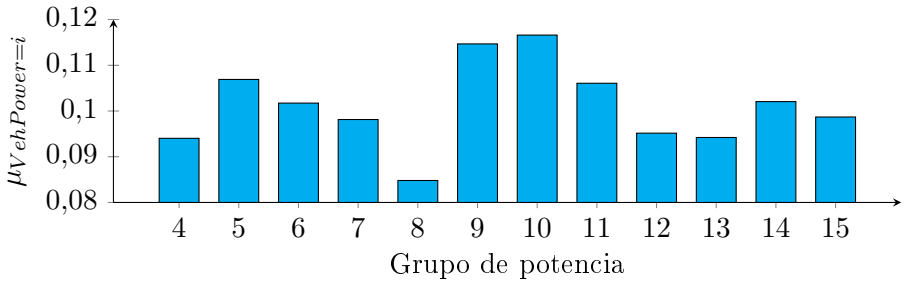

Figura 3: Distribución del número medio de reclamaciones por potencia del vehículo. Elaboración propia.

En vista de las diferencias en cuanto a la frecuencia de reclamaciones se simplifican los grupos a 0, 1 y 2. Para ello, se repite el proceso de reclasificación realizado con la variable "Area" (3.3-6), aunque en este caso en base a la frecuencia de reclamaciones por grupo de potencia (3.7). El resultado del proceso (Anexo 2–C12), reflejado en la variable "VehPower_c", puede verse en el Cuadro 7.

Cuadro 7

Reclasificación de variable "VehPower"

VehPower	Intervalo (Frec.Recl.)	VehPower_c	Pólizas	Proporción
4,8,12,13	[0,08480;0,09540]	0	173.463	25,6 %
6,7,14,15	[0,09540;0,10600]	1	299.118	44,2 %
5,9,10,11	[0,10600;0,11660]	2	204.202	30,2 %

Edad del vehículo

La variable de edad del vehículo "VehAge" es una variable cuantitativa compuesta por números enteros no–negativos. Representando para cada una de estas edades, el número de reclamaciones medio anual (Anexo 2–C13), tal que,

$$\mu_{VehAge=i} = E(ClaimNb|VehAge_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}}, \quad i = 0, 1, 2, ..., 29, \ , \ j = 1, 2, ..., 676783, \tag{3.8}$$

donde $\mu_{VehAge=i}$ es el número medio anual de reclamaciones por cada edad del vehículo $i = (0, 1, 2, ..., 30)$. En la Figura 4, puede observarse la frecuencia del número de reclamaciones anual según la edad del vehículo (se limita la representación a 30 años de antigüedad).

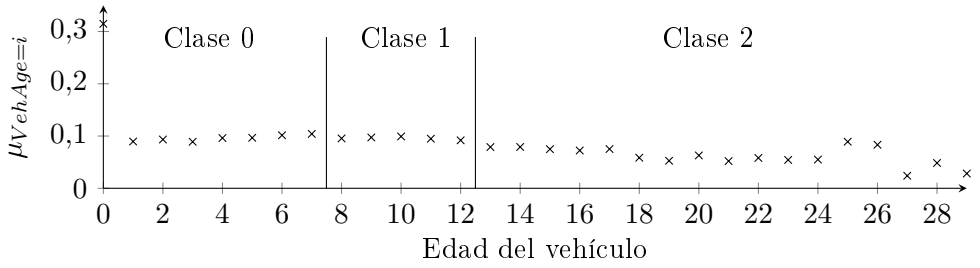

Figura 4: Distribución del número medio de reclamaciones por edad del vehículo. Elaboración propia.

Se observa una importante cantidad de reclamaciones el primer año y posteriormente una tendencia decreciente de la misma en el largo plazo. En torno a los 25 años comienza a observarse un incremento en la volatilidad, pudiendo deberse a la falta de datos. En vista de las diferencias en cuanto a frecuencia de reclamaciones se simplifican los grupos a 0, 1, 2 agrupando en base a la antigüedad. Suponiendo una tendencia lineal decreciente en el largo plazo del número de reclamaciones medio anual, se asigna el grupo 0 a los vehículos de menos de 8 años, el grupo 1 los de 8 a 12 años y en el grupo 2 los de más de 12 años, tal y como se muestra en la Figura 4. No se penalizará el importante exceso de siniestralidad del primer año por fines comerciales. El resultado del proceso (Anexo 2–C14), reflejado en la variable "VehAge_c" puede verse en el Cuadro 8.

Cuadro 8

Reclasificación de variable "VehAge"

VehAge	VehAge_c	Pólizas	Proporción
[0; 7]	0	388.663	57,4 %
[8; 12]	1	161.172	23,8 %
[13; ∞)	2	126.948	18,8 %

Edad del asegurado

La variable de edad del asegurado "DrivAge", es una variable cuantitativa compuesta por números enteros no–negativos. Representando para cada uno de estos valores, el número de

reclamaciones medio anual (Anexo 2–C15), tal que,

$$\mu_{DrivAge=i} = E(ClaimNb|DrivAge_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}}, i = 18, 19, ..., 100, , j = 1, 2, ..., 676783, \tag{3.9}$$

donde $\mu_{DrivAge=i}$ es el número medio anual de reclamaciones por cada edad del asegurado $i = (18, 19, ..., 100)$. En la Figura 5, puede observarse la frecuencia del número de reclamaciones anual según la edad del asegurado.

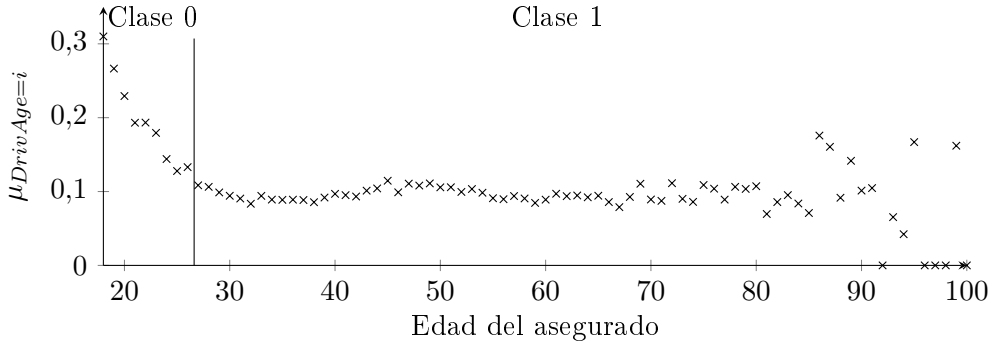

Figura 5: Distribución del número medio de reclamaciones por edad del asegurado. Elaboración propia.

Se observa una importante cantidad de reclamaciones en los conductores más noveles y un descenso importante hasta su estabilización en torno a los 27 años. Posteriormente se observa un leve repunte de la siniestralidad en torno a los 45 años. A partir de los 65 años se comienza a mostrar un aumento de la volatilidad debido posiblemente al descenso de la población conductora y por lo tanto del número de asegurados. En vista de las diferencias en cuanto a frecuencia de reclamaciones, se simplificarán los grupos a 0 y 1 agrupando en base a esta, concretamente en el grupo 0 los de menores de 27 años y en el grupo 1 el resto. El resultado del proceso (Anexo 2–C16), reflejado en la variable "DrivAge_c" puede verse en el Cuadro 9.

Cuadro 9

Reclasificación de variable "DrivAge"

DrivAge	DrivAge_c	Pólizas	Proporción
$[0; 26]$	0	49.169	7,3 %
$[27; \infty)$	1	627.614	92,7 %

Marca del vehículo

La variable de marca del vehículo "VehBrand", es una variable cualitativa conformada por grupos del B1 al B6 y del B10 al B14. Representando para cada uno de estos valores, el número de reclamaciones medio anual (Anexo 2–C17), tal que,

$$\mu_{VehBrand=i} = E(ClaimNb|VehBrand_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}}, i = B1, ..., B14, , j = 1, 2, ..., 676783, \tag{3.10}$$

donde $\mu_{VehBrand=i}$ es el número medio anual de reclamaciones del grupo de marcas $i = (B1, ..., B14)$. En la Figura 6, puede observarse la frecuencia del número de reclamaciones anual según la marca del vehículo.

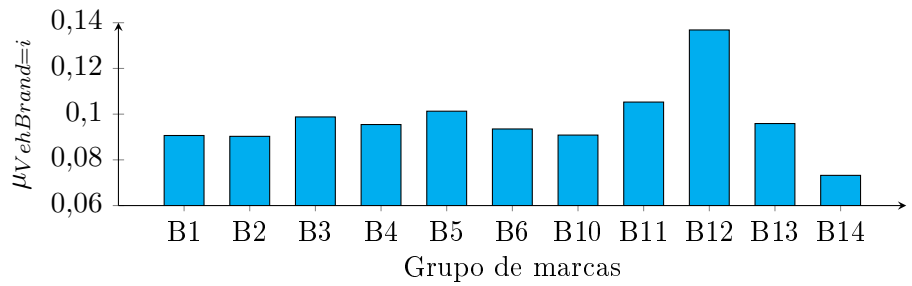

Figura 6: Distribución del número medio de reclamaciones por marca del vehículo. Elaboración propia.

En vista de las diferencias en cuanto a la frecuencia de reclamaciones se simplifican los grupos a 0, 1, 2. Para ello, se repite el proceso de reclasificación realizado con la variable "Area" (3.3-6), aunque en este caso en base a la frecuencia de reclamaciones por marca del vehículo (3.10). El resultado del proceso (Anexo 2–C18), reflejado en la variable "VehBrand_c", puede verse en el Cuadro 10.

Cuadro 10

Reclasificación de variable "VehBrand"

VehBrand	Intervalo (Freq.Recl.)	VehBrand_c	Pólizas	Proporción
B1, B2, B6, B10, B14	[0,07321;0,09442]	0	372.045	55,0 %
B3, B4, B5, B11, B13	[0,09442;0,11562]	1	138.779	20,5 %
B12	[0,11562;0,13682]	2	165.959	24,5 %

Combustible del vehículo

La variable de combustible del vehículo "VehGas", es una variable cualitativa conformada por Regular o Diésel que simplemente se transforman en 0 y 1 respectivamente (Anexo 2–C19). El resultado del proceso, reflejado en la variable "VehGas_c", puede observarse en el Cuadro 11.

Cuadro 11

Reclasificación de variable "VehGas"

VehGas	VehGas_c	Pólizas	Proporción
Regular	0	342.252	51,0 %
Diesel	1	331.531	49,0 %

Densidad de población de la localidad de residencia

La variable de densidad de población de la localidad de residencia "Density", es una variable cuantitativa compuesta por números enteros no–negativos. Representando para cada uno de estos valores, el número de reclamaciones medio anual (Anexo 2–C20), tal que,

$$\mu_{Density=i} = E(ClaimNb|Density_i) = \frac{\sum_j ClaimNb_{ij}}{\sum_j Exposure_{ij}}, \quad i = 1, ..., 27000, \ , \ j = 1, 2, ..., 676783, \tag{3.11}$$

donde $\mu_{Density=i}$ es el número medio anual de reclamaciones en la localidad donde la densidad de población es $i = (1, ..., 27000)$. En la Figura 7, puede observarse la frecuencia del número de reclamaciones anual según la densidad de población del lugar de residencia del asegurado.

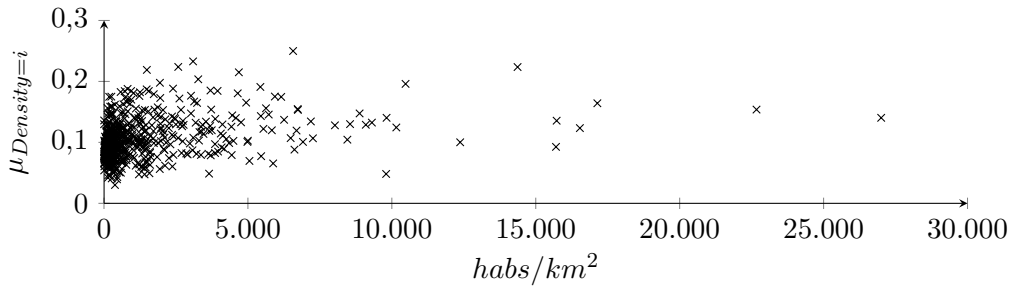

Figura 7: Distribución del número medio de reclamaciones por densidad de habitantes del lugar de residencia. Elaboración propia.

Sin embargo, no se observa ninguna tendencia clara de la que se pudiese extraer una conclusión clara a la hora de agrupar, por lo que se optará por reclasificar las pólizas en las clases 0, 1 y 2 en base a los terciles del valor de la densidad. Este reparto agrupará el tercio de localizaciones con más baja densidad de población en la clase 0, el segundo tercio en la clase 1 y el tercer tercio en la clase 2. El resultado del proceso (Anexo 2–C21), reflejado en la variable "Density_c", puede verse en el Cuadro 12.

Cuadro 12

Reclasificación de variable "Density"

Density	Density_c	Pólizas	Proporción
$[0; 539,98]$	0	377.122	55.7 %
$(539,98; 1439,72]$	1	116.580	17,2 %
$(1439,72; 27000]$	2	183.081	27,1 %

3.1.4. Análisis multivariante

El análisis multivariante conlleva el análisis de más de una variable al mismo tiempo para extraer conclusiones de la relación entre ellas. El estadístico principal en el estudio de la relación entre las variables dos a dos es el coeficiente de correlación, mediante el cálculo de este, de todas las combinaciones de variables (Anexo 2–C22) se obtendrá la matriz de correlaciones del Cuadro 13. Esta es un matriz cuadrada, simétrica y con su diagonal principal formada por unos, ya qué aunque carece de sentido, indica la relación de la variable consigo misma.

En el Cuadro 13 se observa una fuerte correlación entre la variable "Area_c" y la variable "Density_c", lo cual era de esperar al tratarse ambas de variables que hacen referencia a la localización de la residencia del asegurado o a sus características. Por otro lado, también puede verse una moderada correlación negativa entre la variable "VehAge_c" y la variable "VehBrand_c", posiblemente debido a la evolución temporal de la popularidad de distintas marcas o a la variación del poder adquisitivo. El resto de las variables no muestran ninguna relación destacable en el comportamiento.

La frecuencia media de ocurrencia de la reclamación de un siniestro para un asegurado de la

Cuadro 13

Matriz de correlaciones

	Area_c	VehPower_c	VehAge_c	DrivAge_c	VehBrand_c	VehGas_c	Density_c
Area_c	1,00	-0,01	-0,07	-0,01	0,12	-0,12	0,97
VehPower_c	-0,01	1,00	-0,02	0,01	-0,04	0,05	-0,01
VehAge_c	-0,07	-0,02	1,00	-0,05	-0,37	-0,13	-0,07
DrivAge_c	-0,01	0,01	-0,05	1,00	0,04	0,03	-0,01
VehBrand_c	0,12	-0,04	-0,37	0,04	1,00	-0,03	0,12
VehGas_c	-0,12	0,05	-0,13	0,03	-0,03	1,00	-0,11
Density_c	0,97	-0,01	-0,07	-0,02	0,12	-0,11	1,00

cartera en un periodo de exposición completo de 1 año será la siguiente (Anexo 2–C23):

$$\overline{frecuencia_{sin}} = \frac{\sum_{i=1}^{676783} ClaimNb_i}{\sum_{i=1}^{676783} Exposure_i} = 0{,}1007521 \tag{3.12}$$

Se observa una siniestralidad anual de entorno al 10 %, indicando que cada póliza realiza por término medio la reclamación de un siniestro cada 10 años.

Cada uno de esos siniestros tendrá una severidad, reflejado por su coste económico (Anexo 2–C24).

$$\overline{coste_{sin}} = \frac{\sum_{i=1}^{676783} ClaimAmount_i}{\sum_{i=1}^{676783} ClaimNb_i} = 1659,144 \tag{3.13}$$

El coste medio de un siniestro durante el tiempo de exposición grabado en los datos ha sido de 1659,144 €.

La media de la prima pura indica el montante económico de la aportación que los asegurados deberán hacer en media para mantener el equilibrio del sistema (Anexo 2–C25).

$$\overline{Prima\,pura} = \frac{\sum_{i=1}^{676783} ClaimAmount_i}{\sum_{i=1}^{676783} Exposure_i} = 167,1621 \tag{3.14}$$

Esta es una prima que quebranta uno de los objetivos principales de la tarificación en la ciencia actuarial, que es ajustar la prima al riesgo que comporta el asegurado, lo cual, se tratará de realizar en los siguientes apartados.

3.2. Tarificación *a priori*

3.2.1. Selección de modelo y variables

El método de selección de variables para los Modelos Lineales Generalizados será Stepwise mediante el Criterio Informativo de Akaike (AIC) como se vio en el Capítulo 1, utilizando el logaritmo de la variable "Exposure" como variable de desplazamiento, ya que la variable dependiente no se encuentra observada en periodos completos. Este procedimiento proporcionará también un valor con el que seleccionar el modelo que mejor se ajuste a los datos. En primer lugar, se realizará el análisis de un modelo con distribución de Poisson (Anexo 2–C26).

Stepwise AIC Poisson

```
Step:  AIC=289544,7
incurridos~VehPower_c+VehAge_c+DrivAge_c+VehBrand_c+VehGas_c+Density_c+offset(log(exposicion))
             Df Deviance    AIC
<none>            220329 289545
+ Area_c      2   220326 289546
- VehGas_c    1   220357 289571
- VehPower_c  2   220523 289734
- Density_c   2   220757 289969
- VehBrand_c  2   220805 290017
- VehAge_c    2   220821 290032
- DrivAge_c   1   221180 290394
```

Se obtiene un resultado que recomienda la omisión de la variable "Area_c" y que obtiene un AIC de 289544,7. Será interesante analizar la sobre–dispersión de los datos sobre el modelo, para ello se realizará el siguiente contraste de hipótesis propuesto en el apartado 1.4.1 (Anexo 2–C27):

$$\begin{aligned} H_0: &\quad E(Y) = \mu \quad \text{y} \quad Var(Y) = \mu, \\ H_a: &\quad E(Y) = \mu \quad \text{y} \quad Var(Y) = \mu + \alpha \cdot f(\mu) \quad \forall \quad \alpha > 0 \end{aligned} \tag{3.15}$$

Test de sobre–dispersión

```
data:  fit_po
z = 9,2695, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0,181708
```

Por lo tanto, se rechaza la hipótesis nula y se puede confirmar que existe sobre–dispersión, por consiguiente, se buscará otro modelo que se ajuste mejor a los datos. En segundo lugar, se comprobará la efectividad de un MLG con distribución Binomial Negativa y se seleccionarán las variables mediante la técnica escalonada (Anexo 2–C28).

Stepwise AIC Binomial Negativa

```
Step:  AIC=288577,1
ClaimNb~VehPower_c+VehAge_c+DrivAge_c+VehBrand_c+VehGas_c+ Density_c+offset(log(Exposure))
             Df    AIC
<none>          288577
+ Area_c      2 288579
- VehGas_c    1 288607
- VehPower_c  2 288762
- Density_c   2 288972
- VehAge_c    2 289030
- VehBrand_c  2 289059
- DrivAge_c   1 289391
```

En este caso se recomienda la omisión de las variables "Area_c" y "VehGas_c" obteniendo un AIC de 288577,1.

Las diferencias en el AIC indican que el modelo Binomial Negativo sin las variables "Area_c" y "VehGas_c" se ajusta mejor a los datos. Llegado a este punto será una buena opción analizar los dos modelos propuestos (Anexo 2–C29).

Cuadro 14

Parámetros y errores de los modelos analizados

		Poisson		Bin. Neg.	
Variable	Coeficiente	Valor estimado	Error est.	Valor estimado	Error est.
$(Intercept)$	β_0	-1,9534	0,0225	-1,9297	0,0235
$VehPower\_c1$	β_1	0,1483	0,0139	0,1530	0,0144
$VehPower\_c2$	β_2	0,1970	0,0146	0,2032	0,0152
$VehAge\_c1$	β_3	-0,0936	0,0132	-0,0945	0,0138
$VehAge\_c2$	β_4	-0,3493	0,0162	-0,3515	0,0167
$DrivAge\_c1$	β_5	-0,5664	0,0180	-0,5817	0,0189
$VehBrand\_c1$	β_6	0,0327	0,0137	0,0330	0,0142
$VehBrand\_c2$	β_7	0,3054	0,0140	0,3229	0,0146
$VehGas\_c1$	β_8	-0,0575	0,0109	-0,0644	0,0113
$Density\_c1$	β_9	0,1536	0,0145	0,1559	0,0150
$Density\_c2$	β_{10}	0,2517	0,0124	0,2541	0,0129

El Cuadro 14 muestra el valor de los parámetros estimados y errores estándar de los mismos, de los modelos de Regresión de Poisson y Binomial Negativo. Como puede observarse, al tratarse de variables cualitativas a las que no se las ha designado como incrementales, se han separado en nuevas variables imaginarias.

Cuadro 15

Parámetros

Parámetro	Poisson	Bin. Neg.
$\hat{\mu}$	0,0531	0,0540
θ	–	0,8408
τ	–	15,5685

El Cuadro 15 muestra la media de los valores medios del número de reclamaciones estimados por los modelos y los parámetros estimados del modelo binomial negativo para el tiempo de exposición dado.

Cuadro 16

Frecuencia del número de siniestros (Predicción)

Siniestros en el último año	Observados	Poisson	BN	Error Poisson	Error BN
0	643953	642209	643201	1744	752
1	32178	33211	30903	-1033	1275
2	1784	1320	2422	464	-638
3	82	42	229	40	-147
4	7	1	24	6	-17
5	2	0	3	2	-1
6	1	0	0	1	1

Por último, se muestra en el Cuadro 16 las pólizas de la cartera agrupadas por número de reclamaciones efectuadas, tanto observadas, como estimadas por los modelos.

Para seleccionar el modelo más adecuado se realizan varias medidas de la bondad del ajuste vistas en el Capítulo 1 (Anexo 2–C31).

Cuadro 17

Medidas de la bondad del ajuste

Medida	Poisson	Bin. Neg.
ECM	617937,4	374296,1
logV	-144761,5	-144277,5
D	220328,9	188367,7
AIC	289544,7	288577,1

El Cuadro 17 muestra en primer lugar, el valor del ECM, que como se vio anteriormente, valores menores indicarán una menor desviación del valor estimado de la variable sobre el valor observado. En el valor de la log-verosimilitud, aquellos valores mayores indicarán un mejor ajuste sobre los datos, ya que la estimación del modelo se realiza mediante su maximización. En tercer lugar, se muestra el valor de la desviación, en este caso un menor valor indica una menor desviación. En último lugar se muestra el Criterio Informativo de Akaike, que tal y como se vio, un valor menor llevará a seleccionar un modelo que se ajuste mejor a los datos, así como más simple.

Las cuatro medidas de la bondad del ajuste indican que el modelo BN es el que mejor se ajusta a los datos y por lo tanto será el seleccionado para la tarificación *a priori*.

3.2.2. Estimación y validación

De acuerdo al apartado anterior se estimará el modelo siguiente:

$$\begin{aligned}\ln(\hat{\mu}_i) =& \hat{\beta}_0 + VehPower\_c1_i \cdot \hat{\beta}_1 + VehPower\_c2_i \cdot \hat{\beta}_2 + VehAge\_c1_i \cdot \hat{\beta}_3 \\ &+ VehAge\_c2_i \cdot \hat{\beta}_4 + DrivAge\_c1_i \cdot \hat{\beta}_5 + VehBrand\_c1_i \cdot \hat{\beta}_6 \\ &+ VehBrand\_c2_i \cdot \hat{\beta}_7 + VehGas\_c1_i \cdot \hat{\beta}_8 + Density\_c1_i \cdot \hat{\beta}_9 \\ &+ Density\_c2_i \cdot \hat{\beta}_{10} + \ln(Exposure_i)\end{aligned} \tag{3.16}$$

El cómputo global del número de siniestros relación del cómputo global del número de siniestros observados y esperados indicará la precisión global del modelo (Anexo 2–C32)

$$Precision = \frac{\sum_{i=1}^{n} \hat{\mu}_i}{\sum_{i=1}^{n} ClaimNb_i} - 1 = \frac{36549,76}{35982} - 1 = 0,01577 \tag{3.17}$$

La ecuación (3.17) indica que el error global del modelo a la hora de estimar el número de siniestros es únicamente del 1,57 %.

A continuación, con el objetivo de asegurar la independencia del modelo seleccionado respecto de los datos en contra de la problemática del sobreajuste se realizará una validación cruzada de 100 iteraciones (Anexo 2–C33,C34).

En la Figura 8 se observa una distribución del Error Cuadrático Medio aparentemente aleatorio en cada una de las iteraciones ya que no forman ninguna tendencia definida, corroborando la independencia del modelo sobre los datos, aunque no denotando gran estabilidad.

No se contemplan otras técnicas de validación a partir de los residuos ya que carecen de sentido debido a la naturaleza de los datos.

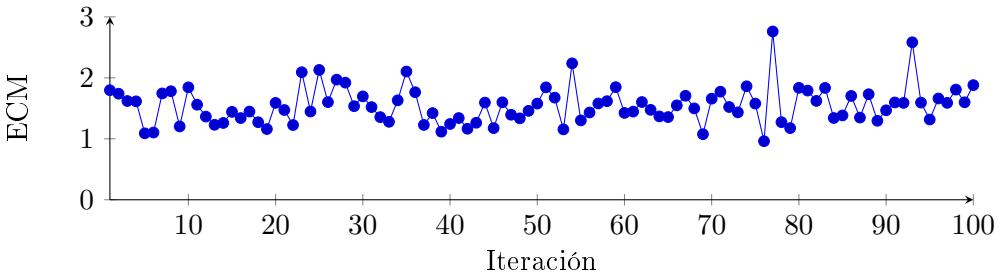

Figura 8: Validación cruzada del modelo Binomial Negativo con $k = 100$ iteraciones. Las iteraciones son representadas por el eje de abscisas, mientras que el eje de ordenadas muestra el Error Cuadrático Medio obtenido en cada una de éstas. Elaboración propia.

3.2.3. Tarificación

Se selecciona la póliza 115255 a modo de ejemplo en el proceso de fijación de la tarifa. Esta es una póliza perteneciente a un vehículo de clase de potencia 4 (transformado a la clase 0), 5 años de antigüedad (transformado a la clase 0), grupo de marcas B2 (transformado a la clase 0), combustible Gasolina (transformado a la clase 0), cuyo tomador tiene 20 años (transformado a la clase 0) y reside en una localidad con una densidad de 109 $habs/km^2$ (transformado a la clase 0).

Para fijar las tarifas para un periodo de un año, se fija en primer lugar el valor de la variable de exposición ($Exposure$) igual a 1, bajo la hipótesis de que los siniestros se distribuyen uniformemente a lo largo del año. A continuación, se obtiene la estimación del número medio de reclamaciones anual a partir de la ecuación (3.16) (Anexo 2–C35).

$$\begin{aligned}\ln(\hat{\mu}) - \ln(Exposure) &= \hat{\beta}_0 + VehPower\_c1 \cdot \hat{\beta}_1 + VehPower\_c2 \cdot \hat{\beta}_2 \\ &+ VehAge\_c1 \cdot \hat{\beta}_3 + VehAge\_c2 \cdot \hat{\beta}_4 + DrivAge\_c1 \cdot \hat{\beta}_5 + VehBrand\_c1 \cdot \hat{\beta}_6 \\ &+ VehBrand\_c2 \cdot \hat{\beta}_7 + VehGas\_c1 \cdot \hat{\beta}_8 + Density\_c1 \cdot \hat{\beta}_9 + Density\_c2 \cdot \hat{\beta}_{10}\end{aligned} \tag{3.18}$$

Aplicando las propiedades de los logaritmos y la función exponencial a ambos lados de la igualdad:

$$\begin{aligned}exp\Big\{\ln\left(\frac{\hat{\mu}}{Exposure}\right)\Big\} = exp\Big\{ &\hat{\beta}_0 + VehPower\_c1 \cdot \hat{\beta}_1 + VehPower\_c2 \cdot \hat{\beta}_2 \\ &+ VehAge\_c1 \cdot \hat{\beta}_3 + VehAge\_c2 \cdot \hat{\beta}_4 + DrivAge\_c1 \cdot \hat{\beta}_5 + VehBrand\_c1 \cdot \hat{\beta}_6 \\ &+ VehBrand\_c2 \cdot \hat{\beta}_7 + VehGas\_c1 \cdot \hat{\beta}_8 + Density\_c1 \cdot \hat{\beta}_9 + Density\_c2 \cdot \hat{\beta}_{10}\Big\}\end{aligned} \tag{3.19}$$

Sustituyendo las variables independientes por sus valores correspondientes:

$$\begin{aligned}exp\Big\{\ln\left(\frac{\hat{\mu}}{1}\right)\Big\} = exp\Big\{ &\hat{\beta}_0 + 0 \cdot \hat{\beta}_1 + 0 \cdot \hat{\beta}_2 + 0 \cdot \hat{\beta}_3 + 0 \cdot \hat{\beta}_4 \\ &+ 0 \cdot \hat{\beta}_5 + 0 \cdot \hat{\beta}_6 + 0 \cdot \hat{\beta}_7 + 0 \cdot \hat{\beta}_8 + 0 \cdot \hat{\beta}_9 + 0 \cdot \hat{\beta}_{10}\Big\}\end{aligned} \tag{3.20}$$

Por último, sustituyendo por los parámetros estimados se obtendrá el número medio de reclamaciones anuales estimado de dicha póliza.

$$\hat{\mu} = e^{\hat{\beta}_0} = e^{(-1,9297)} = 0,14519 \tag{3.21}$$

La fijación de la tarifa *a priori* se hará en función del riesgo relativo que represente la póliza sobre el conjunto de la cartera. Mediante la relación (1.51) se obtiene el valor relativo para esta póliza (Anexo 2–C36):

$$r = \frac{\hat{\mu}}{\sum_{i=1}^{n} \hat{\mu}_i} = \frac{0,14519}{72600,06} = 0,000001999885 \tag{3.22}$$

A continuación, será necesario anualizar el montante económico del conjunto de las reclamaciones, para ello se establece la hipótesis de que es un cantidad proporcional al número de reclamaciones (Anexo 2–C37).

$$Q = \sum_{i=1}^{n} ClaimAmount_i \cdot \left(\frac{\sum_{i=1}^{n} Expuestos_i}{\sum_{i=1}^{n} Exposure_i} \right) = 59699307 \cdot \frac{676783}{357134,1} = 113132496 \tag{3.23}$$

Por último, se obtiene la prima *a priori* de dicha póliza de acuerdo a la relación (1.52) (Anexo 2–C38):

$$P_0 = r \cdot Q = 0,0000019986 \cdot 113132496 = 226,26 \tag{3.24}$$

Para una simple y rápida visualización de las tarifas se conforma la tabla de tarifas *a priori* (Anexo 2–C39) que muestra el Anexo 3. Esta contiene una única tarifa para cada grupo homogéneo de riesgo, es decir, para cada póliza que tenga las mismas características. Existe un pequeño número de grupos homogéneos de riesgo que no contienen ninguna póliza y por lo tanto no se les asigna ninguna tarifa.

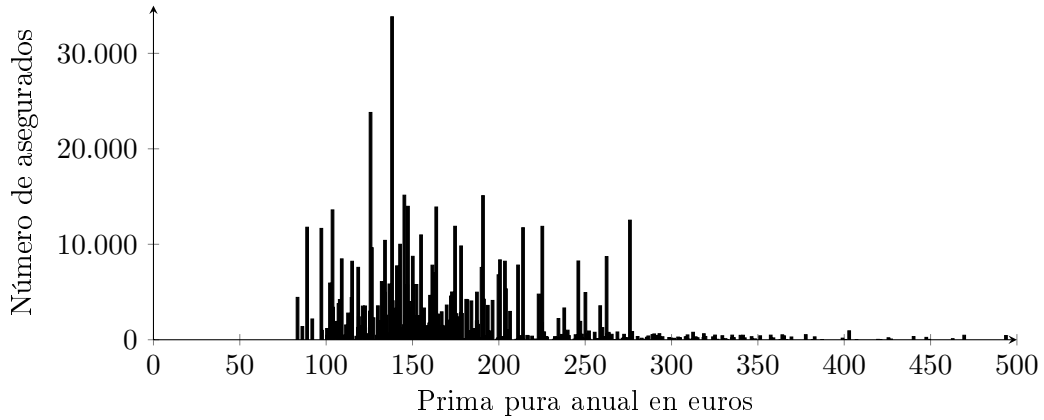

Figura 9: Frecuencia de las tarifas a priori. El eje de abscisas muestra el valor de la prima pura a pagar y el eje de ordenadas el número de pólizas con la prima correspondiente. Elaboración propia.

La Figura 9 muestra la distribución del número de asegurados que tienen asignado cada nivel de tarifa, como se puede observar, la mayoría de las pólizas tienen tarifas en el rango de 80 y 280 euros.

3.3. Tarificación *a posteriori*

Como se comentó en el capítulo anterior, el objetivo de la tarificación *a posteriori* es ajustar la tarifa al comportamiento individual del asegurado, ajustando la prima a las variables que no son observables fácilmente. Para ello se realizará un modelo de Bonus–Malus aplicable mediante

una tabla de ajuste a una serie de años historial de reclamaciones será simulado en base al modelo de reclamaciones *a priori*.

El sistema de Bonus–Malus funcionará de acuerdo a las siguientes reglas:

- La memoria del sistema será de 5 años, una vez pasados estos, la reclamación no se tendrá en cuenta.
- El número máximo de reclamaciones a tener en cuenta será de 6.

A partir de estas premisas se configurarán unos escenarios de simulación.

3.3.1. Escenarios de simulación

En primer lugar, se generan las probabilidades de realizarse exactamente 0 reclamaciones, exactamente 1 reclamación, y así hasta 6 reclamaciones, mediante la función de probabilidad BN vista en (1.39) (Anexo 2–C40). Siguiendo con el ejemplo de la póliza 115255, se obtiene la probabilidad de que no se realice ninguna reclamación:

$$P(Y=0) = \frac{\Gamma(0+0,8408)}{\Gamma(0+1)\Gamma(0,8408)} \left(\frac{0,8408}{0,14519+0,8408}\right)^{0,8408} \left(\frac{0,14519}{0,14519+0,8408}\right)^{0} = 0,8746443 \tag{3.25}$$

Continuando del mismo modo hasta alcanzar $y = 6$ se obtienen las probabilidades necesarias.

A continuación, se realiza un muestreo aleatorio ponderado de la serie de valores $\{0, 1, 2, 3, 4, 5, 6\}$ para cada póliza, durante 15 periodos. La ponderación de cada valor de la serie serán las probabilidades obtenidas anteriormente mediante la distribución BN (Anexo 2–C41). Estas serán las reclamaciones simuladas en cada periodo para cada póliza.

Posteriormente se acumulan según avanza el tiempo y, de acuerdo las reglas del sistema, una vez pasados cinco años se eliminan (Anexo 2–C42).

3.3.2. Tabla Bonus–Malus

Dado que el modelo elegido ha sido el BN, se aplica la ecuación 2.6 para obtener la prima BM. Para ello en primer lugar, se actualiza el valor de τ, ya que el valor del número medio de reclamaciones ha variado a causa de la anualización.

$$\hat{\mu} = \frac{\theta}{\tau} \tag{3.26}$$

$$\tau = \frac{\theta}{\hat{\mu}} = \frac{\theta}{\frac{\sum_i^n \hat{\mu}_i}{n}} = \frac{0,8407}{0,1072} = 7,8378 \tag{3.27}$$

A continuación, se aplica la ecuación (2.6) en base 100 (Anexo 2–C43):

$$P_{BM}(y_1, y_2, ..., y_t) = 100 \cdot \frac{\tau(\theta + t\bar{y})}{\theta(\tau + t)} = 100 \cdot \frac{7,8378 \cdot (0,8407 + t\bar{y})}{0,8407 \cdot (7,8378 + t)} \tag{3.28}$$

En el Cuadro 18 se muestran los valores relativos obtenidos a aplicar sobre las primas *a priori* de cada una de las pólizas.

Cuadro 18

Tabla Bonus–Malus

t	$t\bar{y}$						
	0	1	2	3	4	5	6
0	100,00	218,94	337,87	456,81	575,75	694,68	813,62
1	88,69	194,16	299,64	405,12	510,60	616,08	721,56
2	79,67	174,43	269,19	363,94	458,70	553,46	648,21
3	72,32	158,33	244,35	330,36	416,38	502,39	588,40
4	66,21	144,96	223,71	302,45	381,20	459,95	538,70
5	61,05	133,67	206,28	278,89	351,51	424,12	496,74

3.3.3. Análisis de la simulación de la tarificación *a posteriori*

Para obtener la prima final *a posteriori* se aplicará la ecuación (2.7) a cada póliza en cada periodo t (Anexo 2–C44). Siguiendo con la póliza de ejemplo, en $t = 0$, su prima final será:

$$P_0 = P_0 \cdot P_{BM} = P_0 \cdot \frac{\tau(\theta + t \cdot \bar{y}_i)}{\theta(\tau + \cdot t)} = P_0 \cdot \frac{\tau\left(\theta + t \cdot \frac{\sum_{j=t-5}^{t} y_j}{5}\right)}{\theta(\tau + \cdot t)} = P_0 \cdot \frac{\tau(\theta + 0 \cdot \bar{y}_i)}{\theta(\tau + \cdot 0)} = P_0 = 226,26 \tag{3.29}$$

En el periodo $t = 5$ ha pasado 5 años sin realizar reclamaciones, por lo tanto su prima final será:

$$P_5 = P_0 \cdot P_{BM} = 226,26 \cdot 61,05 = 138,13 \tag{3.30}$$

En el periodo $t = 7$ sufre un siniestro, por lo tanto su prima final pasa a ser:

$$P_7 = P_0 \cdot P_{BM} = 226,26 \cdot 133,67 = 302,43 \tag{3.31}$$

En el periodo $t = 12$ han pasado 5 años sin realizar reclamaciones, por lo tanto su prima final será:

$$P_{12} = P_0 \cdot P_{BM} = 226,26 \cdot 61,05 = 138,13 \tag{3.32}$$

Para verificar la validez del modelo y su independencia respecto de los datos utilizados, se realizará una simulación de coste de las reclamaciones en 15 años consecutivos (Anexo 2–C45), tal que,

$$Coste_t = \sum_{i=1}^{n} ClaimAmount \cdot \left(\frac{\sum_{i=1}^{n} ClaimNb_t}{\sum_{i=1}^{n} ClaimNb}\right), \quad t = 1, ..., 15. \tag{3.33}$$

En la Figura 10 pueden compararse la suma de las primas simuladas en cada periodo y la suma del coste total de las reclamaciones simuladas.

Los resultados arrojados indican tal y como puede verse en la Figura 11 a partir del 5º año que el modelo Bonus–Malus hace un buen ajuste, mostrando estabilidad en las primas totales pagadas, aunque con un pequeño exceso de prima. Se analiza también su relación con el coste de los siniestros simulados (Anexo 2–C46).

$$Desviación = \frac{\sum_{t=5}^{15} Prima\ total_t - \sum_{t=5}^{15} Coste_t}{\sum_{t=5}^{15} Coste_t} \cdot 100 = 8,07\,\%, \quad t = 5, ..., 15 \tag{3.34}$$

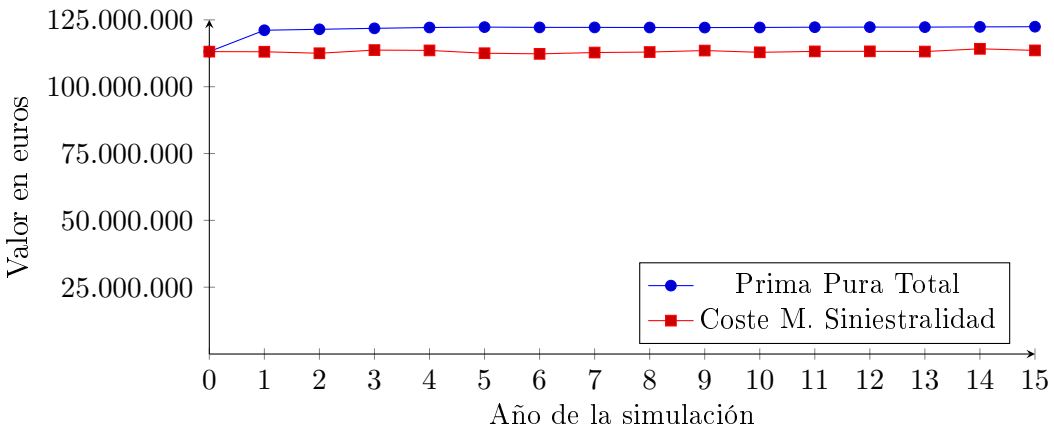

Figura 10: Simulación del comportamiento de la cartera de asegurados. Elaboración propia.

Una desviación de las primas pagadas sobre el coste de los siniestros de únicamente el $8,07\,\%$ indica que el modelo de tarificación está bien ajustado a la siniestralidad.

Por otro lado, el ajuste individual de las primas lleva a obtener una nueva distribución de las mismas, disolviendo la anterior segmentación en base a los grupos de riesgo (Anexo 2–C47).

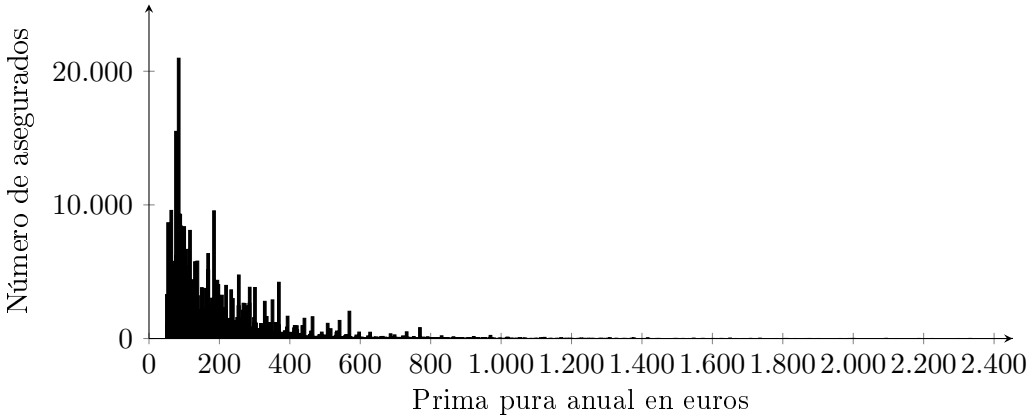

Figura 11: Frecuencia de las tarifas en el décimo año. Elaboración propia.

En la Figura 11 puede observarse que el número de asegurados que pagan la prima más habitual cuyo valor representa la moda, es menor al valor previo a la tarificación *a posteriori*. Esto ocurre debido al ajuste individual de la prima y por los tanto a la mayor variabilidad de esta. En está ocasión también se observa la aparición de valores más extremos, lo cual es la consecuencia principal de la aplicación de este sistema.

A continuación, se analiza el comportamiento de varias pólizas seleccionadas por sus particulares características (Anexo 2–C48).

La Figura 12 muestra el comportamiento de la tarifa de 3 pólizas. Debido a que inician su tarificación en un momento que no tienen un historial de siniestralidad, pueden considerarse como pólizas de tomadores que nunca antes han contratado un seguro o que en cualquier caso no existe registro disponible.

La primera de ellas, la póliza 26421, se trata de una póliza de un automóvil de potencia categoría 5, 12 años de antigüedad, del conjunto de marcas B3, combustible gasolina y cuyo tomador tiene 66 años de edad y es residente en una localidad con $45 habs/km^2$ de densidad y que no ha sufrido ningún siniestro a lo largo de los periodos simulados, por lo que su tarifa se

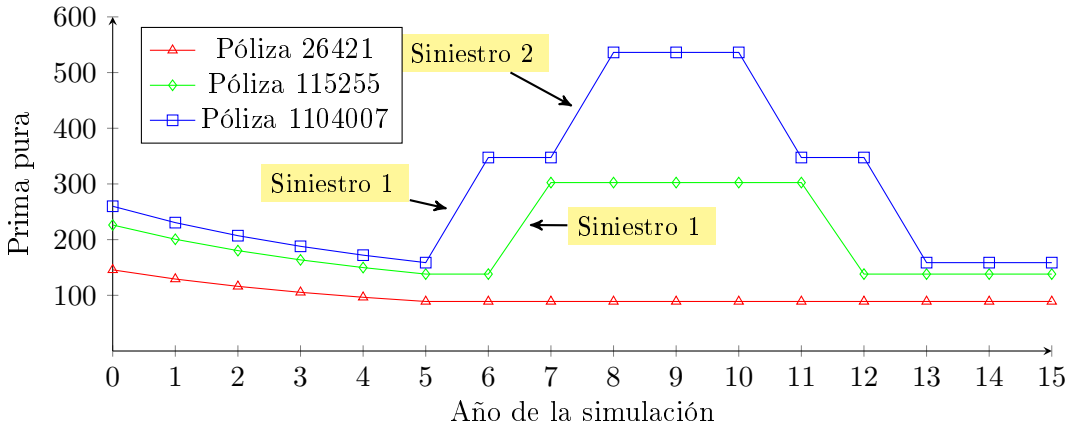

Figura 12: Simulación del comportamiento individual de las pólizas. Elaboración propia.

mantiene inalterada a partir de la estabilización del sistema en el 5º año.

En segundo lugar, la póliza 115255, se trata de la póliza de ejemplo utilizada anteriormente. Como se vio, ha sufrido un siniestro a lo largo de los periodos simulados, por lo que su tarifa sufre un incremento a partir del siguiente año y durante 5 años.

En tercer lugar, la póliza 1104007, se trata de una póliza de un automóvil de potencia categoría 5, 4 años de antigüedad, del conjunto de marcas B1, combustible diesel y cuyo tomador tiene 25 años de edad y es residente en una localidad con $81 habs/km^2$ de densidad y que ha sufrido dos siniestros a lo largo de los periodos simulados, por lo que su tarifa sufre un incremento a partir del siguiente año de cada uno de ellos y durante 5 años a partir de cada uno de ellos.

La ejecución de todo el programa que lleva a los resultados obtenidos en este capítulo se puede realizar ejecutando el siguiente comando en R:

Programa completo

```
source("http://dl.dropbox.com/s/15lxmd0uz7f5vwl/script_TFG_Daniel_Del_Castillo.r")
```

Capítulo 4

Conclusiones

El presente trabajo ha sido presentado con el objetivo fundamental de mostrar varias técnicas estadísticas de tarificación utilizadas comúnmente en el mercado asegurador, tanto desde una perspectiva teórica como práctica. Como enfoque central de todas ellas se encuentra la estimación del número de siniestros que los tomadores de las pólizas pueden realizar.

Partiendo de una base de datos pública que contiene una cartera asegurados perteneciente a un asegurador francés, tras una revisión teórica de la metodología más común en el sector, se ha llevado a cabo un proceso de tarificación que ha constado de todas las fases habituales dentro del alcance estadístico. Apoyándose en un análisis descriptivo de las variables características disponibles y tras eliminar aquellas que se identifiquen como redundantes o distorsionantes se realiza una segmentación de las pólizas de acuerdo a sus factores de riesgo, siempre tratando de otorgarle un sentido a la misma. Se confirma analíticamente que la variable Área y la Densidad resultan redundantes, tras mostrar una fuerte correlación y posteriormente la variable Área muestra su escasa aportación al modelo al realizarse la selección de variables, por lo que se elimina esta última.

Desde el punto de vista de los Modelos Lineales Generalizados se muestra la necesidad de recurrir a un modelo más flexible que el de Poisson al observarse sobre–dispersión, se opta en segundo lugar por el modelo Binomial Negativo. Se consigue finalmente el mejor ajuste mediante el modelo Binomial Negativo, obteniendo una escala de tarifas bien ajustada a los riesgos de cada grupo homogéneo de riesgo, mostrando además una buena independencia sobre los datos y por lo tanto un buen poder de predicción sobre cualquier muestra.

Por último, se pone en práctica el sistema Bonus–Malus óptimo durante una simulación temporal, observandose tanto una buena estabilidad financiera a lo largo del tiempo como el efecto deseado sobre el reajuste de la tarifa a la experiencia individual. Se puede concluir, por lo tanto, que el proceso de análisis teórico y posterior puesta en práctica ha producido el resultado buscado.

A su vez este estudio podría profundizarse utilizando otros modelos de regresión más avanzados como los inflados de ceros, añadir un modelo del coste de los siniestros o realizar una tarificación individualizada sin recurrir al grado de segmentación que proporciona el transformar algunas variables en cualitativas.

Anexo 1

Instrucciones de instalación del software estadístico libre R, así como de los paquetes necesarios para la ejecución del código a lo largo del caso práctico.

1. Instalar R Base
   1.1. Acceder al sitio oficial de R: `http://cran.at.r-project.org/`
   1.2. Seleccionar la versión que corresponda a nuestro sistema operativo.
   1.3. Descargar e instalar
2. Instalar R Studio
   2.1. Acceder al sitio oficial de R: `https://www.rstudio.com/products/rstudio/download/`
   2.2. Seleccionar la versión que corresponda a nuestro sistema operativo.
   2.3. Descargar e instalar
3. Instalar paquetes necesarios
   3.1. Ejecutar R/Rstudio
   3.2. Seleccionar todas las líneas de código dentro del siguiente recuadro

Preámbulo 1

```
list.of.packages <- c("zoo", "xts", "sp", "moments", "MASS","AER","data.table")
new.packages <- list.of.packages[!(list.of.packages %in%
    installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
library("zoo"); library("xts"); library("sp"); library("moments");
    library("MASS"); library("AER"); library("data.table")

freMTPL2sev <-
    read.csv(url("https://www.dropbox.com/s/f3skpp0rd0vdmni/freMTPL2sev.csv?raw=1"))
freMTPL2freq <-
    read.csv(url("https://www.dropbox.com/s/l1c0zlu8xkj7t7r/freMTPL2freq.csv?raw=1"))
```

   3.3. Copiarlas (Ctrl+c), pegarlas en la consola de R (Ctrl+v) y ejecutar (Intro).

Anexo 2

En este apartado se muestran las líneas de código en el lenguaje R a las que se hace referencia en el caso práctico.

Depuración y transformación

C1

```
A <- data.table(freMTPL2freq)
```

C2

```
B <- data.table(freMTPL2sev)
```

C3

```
ClaimNB_freq <- data.table(table(A[,ClaimNb]))
```

C4

```
A <- A[!A$ClaimNb>6,]
```

C5

```
A <- A[!A$Exposure>1,]
```

C6

```
B <- merge(A, B, by="IDpol")
```

Análisis descriptivo univariante y adaptación

C7

```
n_polizas <- nrow(A); stats_A <- data.table(n_polizas)
stats_A$media <- mean(A$ClaimNb)
stats_A$varianza <- var(A$ClaimNb)
stats_A$asimetria <- skewness(A$ClaimNb)
stats_A$curtosis <- kurtosis(A$ClaimNb)
```

C8

```
cont_reclam <- data.table(table(A$ClaimNb))
```

C9

```
area <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)), by=Area]
area$freq <- area$total_claim/area$total_exp
```

C10

```
#ATENCION: revisar que no haya espacios dentro del operador "%in%" a la hora de ejecutar el
    código
A$Area_c <- 1
A$Area_c[A$Area %in%
    area[freq<(min(area$freq)+((1/3)*(max(area$freq)-min(area$freq)))),][[1]]] <- 0
A$Area_c[A$Area %in%
    area[freq>(min(area$freq)+((2/3)*(max(area$freq)-min(area$freq)))),][[1]]] <- 2
Area_c <- data.table(table(A$Area_c))
```

C11

```
vehpower <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)),
    by=VehPower]
vehpower$freq <- vehpower$total_claim/vehpower$total_exp
```

C12

```
A$VehPower_c <- 1
A$VehPower_c[A$VehPower %in%
    vehpower[freq<(min(vehpower$freq)+((1/3)*(max(vehpower$freq)-min(vehpower$freq)))),][[1]]]
    <- 0
A$VehPower_c[A$VehPower %in%
    vehpower[freq>(min(vehpower$freq)+((2/3)*(max(vehpower$freq)-min(vehpower$freq)))),][[1]]]
    <- 2
VehPower_c <- data.table(table(A$VehPower_c))
```

C13

```
vehage <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)),
    by=VehAge]
vehage$freq <- vehage$total_claim/vehage$total_exp
```

C14

```
A$VehAge_c <- 0
A$VehAge_c[A$VehAge %in% c(8:12)] <- 1
A$VehAge_c[A$VehAge %in% c(13:150)] <- 2
VehAge_c <- data.table(table(A$VehAge_c))
```

C15

```
drivage <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)),
    by=DrivAge]
drivage$freq <- drivage$total_claim/drivage$total_exp
```

C16

```
A$DrivAge_c <- 0
A$DrivAge_c[A$DrivAge %in% c(27:150)] <- 1
DrivAge_c <- data.table(table(A$DrivAge_c))
```

C17

```
veh_brand <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)),
    by=VehBrand]
veh_brand$freq <- veh_brand$total_claim/veh_brand$total_exp
```

C18

```
A$VehBrand_c <- 1
A$VehBrand_c[A$VehBrand %in% veh_brand[freq<(min(veh_brand$freq) +
    ((1/3)*(max(veh_brand$freq) - min(veh_brand$freq)))),][[1]]] <- 0
A$VehBrand_c[A$VehBrand %in% veh_brand[freq>(min(veh_brand$freq) +
    ((2/3)*(max(veh_brand$freq) - min(veh_brand$freq)))),][[1]]] <- 2
VehBrand_c <- data.table(table(A$VehBrand_c))
```

C19

```
A$VehGas_c <- 0
A$VehGas_c[A$VehGas %in% "Diesel"] <- 1
A$VehGas_c <- as.factor(A$VehGas_c)
```

C20

```
density <- A[,.(total_claim=sum(ClaimNb), total_exp=sum(Exposure), n=length(ClaimNb)),
    by=Density]
density$freq <- density$total_claim/density$total_exp
```

C21

```
A$Density_c <- 1
A$Density_c[A$Density %in% c(1:quantile(density$Density, c(.33)))] <- 0
A$Density_c[A$Density %in% c(round(quantile(density$Density,c(.66))) +
    1:max(density$Density))] <- 2
Density_c <- data.table(table(A$Density_c))
```

Análisis multivariante

C22

```
M_corr<-cor(A[,13:19, with=FALSE])
```

C23

```
sum(A$ClaimNb)/sum(A$Exposure)
```

C24

```
sum(B$ClaimAmount)/sum(A$ClaimNb)
```

C25

```
sum(B$ClaimAmount)/sum(A$Exposure)
```

Selección de modelo y variables

C26

```
A$Area_c <- as.factor(A$Area_c)
A$VehPower_c <- as.factor(A$VehPower_c)
A$VehAge_c <- as.factor(A$VehAge_c)
A$DrivAge_c <- as.factor(A$DrivAge_c)
A$VehBrand_c <- as.factor(A$VehBrand_c)
A$VehGas_c <- as.factor(A$VehGas_c)
A$Density_c <- as.factor(A$Density_c)

fit_po <- glm(ClaimNb~Area_c + VehPower_c + VehAge_c + DrivAge_c + VehBrand_c + VehGas_c +
    Density_c + offset(log(Exposure)), fam=poisson(link = log), data=A)
stepAIC(fit_po, direction = "both")
fit_po <- glm(ClaimNb~VehPower_c + VehAge_c + DrivAge_c + VehBrand_c + VehGas_c + Density_c +
    offset(log(Exposure)), fam=poisson(link = log), data=A)
```

C27

```
dispersiontest(fit_po,trafo=1)
```

C28

```
fit_nb <- glm.nb(ClaimNb~Area_c + VehPower_c + VehAge_c + DrivAge_c + VehBrand_c + VehGas_c +
    Density_c + offset(log(Exposure)), link = log, data=A)
step_nb <- stepAIC(fit_nb, direction = "both")
fit_nb <- glm.nb(ClaimNb~VehPower_c + VehAge_c + DrivAge_c + VehBrand_c + VehGas_c +
    Density_c + offset(log(Exposure)), link = log, data=A)
```

C29

```
fits <- list("Poisson" = fit_po, "NB" = fit_nb)
coefs <- data.table(sapply(fits, function(x) coef(x)[1:11]))
std_e <- cbind("Poisson" = sqrt(diag(vcov(fit_po))), sapply(fits[-1], function(x)
    sqrt(diag(vcov(x)))[1:11]))
rownames(coefs) <- rownames(std_e)
mu_po <- mean(fit_po$fitted.values)
mu_bn <- mean(fit_nb$fitted.values)
theta <- fit_nb$theta
tau <- theta/mu_bn
```

C30

```
step_nb <- stepAIC(fit_nb, direction = "both")
fit_nb <- glm.nb(incurridos~VehPower_c + VehAge_c + DrivAge_c + VehBrand_c + Density_c +
    offset(log(exposicion)), link = log, data=tarifas_sub)
```

C31

```
ecm_po <- sum(comparacion$error_Poisson^2)/nrow(comparacion)
ecm_nb <- sum(comparacion$error_nb^2)/nrow(comparacion)
D_po <- fit_po$deviance
D_nb <- fit_nb$deviance
log_L <- sapply(fits, function(x) logLik(x))
aic_glm <- AIC(fit_po,fit_nb)
b_ajuste <- rbind(ECM = c(ecm_po,ecm_nb), logL = log_L, D = c(D_po, D_nb), AIC = t(aic_glm))
b_ajuste <- b_ajuste[c(1,2,5,4,3),]
```

Estimación y validación

C32

```
precision <- (sum(A$n_med_sin_nb)/sum(A$ClaimNb))-1
```

C33

```
n_VehPower_c <- length(unique(A$VehPower_c))
n_VehAge_c <- length(unique(A$VehAge_c))
n_DrivAge_c <- length(unique(A$DrivAge_c))
n_VehBrand_c <- length(unique(A$VehBrand_c))
n_VehGas_c <- length(unique(A$VehGas_c))
n_Density_c <- length(unique(A$Density_c))
clases_riesgo <- n_VehPower_c * n_VehAge_c * n_DrivAge_c * n_VehBrand_c * n_VehGas_c *
    n_Density_c
tarifas <- data.table(matrix(1:clases_riesgo))
tarifas$VehPower_c <- as.factor(rep(0:(n_VehPower_c-1), each=clases_riesgo/(n_VehPower_c),
    len=clases_riesgo))
tarifas$VehAge_c <- as.factor(rep(0:(n_VehAge_c-1), each=clases_riesgo/(n_VehPower_c *
    n_VehAge_c), len=clases_riesgo))
tarifas$DrivAge_c <- as.factor(rep(0:(n_DrivAge_c-1), each=clases_riesgo/(n_VehPower_c *
    n_VehAge_c * n_DrivAge_c), len=clases_riesgo))
tarifas$VehBrand_c <- as.factor(rep(0:(n_VehBrand_c-1), each=clases_riesgo/(n_VehPower_c *
    n_VehAge_c * n_DrivAge_c * n_VehBrand_c), len=clases_riesgo))
tarifas$VehGas_c <- as.factor(rep(0:(n_VehGas_c-1), each=clases_riesgo/(n_VehPower_c *
    n_VehAge_c * n_DrivAge_c * n_VehBrand_c * n_VehGas_c), len=clases_riesgo))
tarifas$Density_c <- as.factor(rep(0:(n_Density_c-1), each=clases_riesgo/(n_VehPower_c *
    n_VehAge_c * n_DrivAge_c * n_VehBrand_c * n_VehGas_c * n_Density_c), len=clases_riesgo))
tarifas_v<-tarifas
```

C34

```
set.seed(1234)
k_folds <- 100
A_val <- A[sample(nrow(A)), ]
folds <- cut(seq(1,nrow(A_val)), breaks = k_folds, labels = FALSE)
ecm <- c()
for(i in 1:k_folds){
        testIndexes <- which(folds==i, arr.ind=TRUE)
        A_val_test <- A_val[testIndexes, ]
        A_val_ent <- A_val[-testIndexes, ]
        tarifas_val_ent <- tarifas_v
        tarifas_val_test <- tarifas_v
        groups_val_ent <- A_val_ent[,.(Exposure=sum(Exposure), incurridos=sum(ClaimNb)),
    by=c("VehPower_c", "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c", "Density_c")]
        groups_val_ent <- transform(groups_val_ent, clase=paste0(VehPower_c, VehAge_c,
    DrivAge_c, VehBrand_c, VehGas_c, Density_c))
        tarifas_val_ent <- transform(tarifas_val_ent, clase=paste0(VehPower_c, VehAge_c,
    DrivAge_c, VehBrand_c, VehGas_c, Density_c))
        groups_val_ent[,c("VehPower_c", "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c",
    "Density_c")] <- NULL
        tarifas_val_ent <- merge(tarifas_val_ent,groups_val_ent, by="clase", all.x=TRUE)
        tarifas_val_ent[is.na(tarifas_val_ent)] <- 0
        fit_nb_vc <- glm.nb(incurridos~VehPower_c + VehAge_c + DrivAge_c + VehBrand_c +
    VehGas_c + Density_c + offset(log(Exposure)), link = log, data =
    tarifas_val_ent[Exposure>0])
        groups_val_test <- A_val_test[,.(Exposure=sum(Exposure), incurridos=sum(ClaimNb)),
    by=c("VehPower_c", "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c", "Density_c")]
        groups_val_test <- transform(groups_val_test, clase=paste0(VehPower_c, VehAge_c,
    DrivAge_c, VehBrand_c, VehGas_c, Density_c))
        tarifas_val_test <- transform(tarifas_val_test, clase=paste0(VehPower_c, VehAge_c,
    DrivAge_c, VehBrand_c, VehGas_c, Density_c))
        groups_val_test[,c("VehPower_c", "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c",
    "Density_c")] <- NULL
        tarifas_val_test <- merge(tarifas_val_test,groups_val_test, by="clase", all.x=TRUE)
        tarifas_val_test[is.na(tarifas_val_test)] <- 0
        tarifas_val_test <- tarifas_val_test[Exposure>0]
        tarifas_val_test$esperados <- predict(fit_nb_vc, tarifas_val_test, type = "response")
        s <- c()
        for(j in 1:nrow(tarifas_val_test)){s[j] = ((tarifas_val_test$incurridos[j] -
    tarifas_val_test$esperados[j]))^2}
        tarifas_val_test <- cbind(tarifas_val_test, error2 = s)
        ecm[i] = mean(tarifas_val_test$error2)
}
vc = data.table(cbind(1:k_folds,ecm))
```

Tarificación

C35

```
A$Exposure_parcial <- A$Exposure
A$Exposure <- 1
A$prob_m_rec_anual <- predict(fit_nb, A, type = "response")
```

C36

```
A$peso_rec <- A$prob_m_rec_anual/sum(A$prob_m_rec_anual)
```

C37

```
v_rec_anual <- sum(B$ClaimAmount) * (sum(A$Exposure)/sum(A$Exposure_parcial))
```

C38

```
A$pp_apriori <- A$peso_rec * v_rec_anual
```

C39

```
groups <- A[,.(pp_apriori=unique(pp_apriori), Exposure=sum(Exposure)), by=c("VehPower_c",
    "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c", "Density_c")]
groups <- transform(groups, clase=paste0(VehPower_c, VehAge_c, DrivAge_c, VehBrand_c,
    VehGas_c, Density_c))
tarifas <- transform(tarifas, clase=paste0(VehPower_c, VehAge_c, DrivAge_c, VehBrand_c,
    VehGas_c, Density_c))
groups[,c("VehPower_c", "VehAge_c", "DrivAge_c", "VehBrand_c", "VehGas_c", "Density_c")] <-
    NULL
tarifas <- merge(tarifas,groups, by="clase", all.x=TRUE)
tarifas[is.na(tarifas)] <- 0
tarifas_print <- tarifas[,c(3:9), with=FALSE]
tarifas_print$pp_apriori <- round(tarifas_print$pp_apriori, digits = 2)
tarifas_print_wide <- dcast(tarifas_print,VehPower_c+VehAge_c+ VehBrand_c~
    DrivAge_c+VehGas_c+Density_c,value.var="pp_apriori")
```

Tarificación *a posteriori*

C40

```
t_max <- 5; t_mem <- t_max
k_max <- 6
t_max_sim <- 15
A$p_0rec_anual <- dnbinom(0, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_1rec_anual <- dnbinom(1, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_2rec_anual <- dnbinom(2, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_3rec_anual <- dnbinom(3, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_4rec_anual <- dnbinom(4, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_5rec_anual <- dnbinom(5, mu = A$prob_m_rec_anual, size = fit_nb$theta)
A$p_6rec_anual <- dnbinom(6, mu = A$prob_m_rec_anual, size = fit_nb$theta)
```

Escenarios de simulación

C41

```
for(i in 1:t_max_sim) {set.seed(123+i);for(j in 1:nrow(A)){s[j] = sample.int(7, 1, prob =
    A[j, c("p_0rec_anual", "p_1rec_anual", "p_2rec_anual", "p_3rec_anual", "p_4rec_anual",
    "p_5rec_anual", "p_6rec_anual"), with=FALSE])-1};A <- cbind(A, s); colnames(A)[ncol(A)]
    <- paste("ClaimNb_t", i, sep="")}
```

C42

```
A$ClaimNb_acum_t1 <- A$ClaimNb_t1+A$ClaimNb
for(i in 2:t_mem) {ClaimNb_acum <- eval(parse(text = paste("A$ClaimNb_t", i, sep = ""))) +
    eval(parse(text = paste("A$ClaimNb_acum_t", i-1, sep=""))); A <- cbind(A,
    ac=ClaimNb_acum); colnames(A)[ncol(A)] <- paste("ClaimNb_acum_t", i, sep="")}
for(i in (t_mem+1):t_max_sim) {ClaimNb_acum <- eval(parse(text = paste("A$ClaimNb_t", i, sep
    = ""))) + eval(parse(text = paste("A$ClaimNb_acum_t", i-1, sep=""))) - eval(parse(text =
    paste("A$ClaimNb_t", i-t_mem, sep = ""))); A <- cbind(A, ac=ClaimNb_acum);
    colnames(A)[ncol(A)] <- paste("ClaimNb_acum_t", i, sep="")}
for(i in 1:t_max_sim) {A[[paste("ClaimNb_acum_t", i, sep="")]] <- sapply(eval(parse(text =
    paste("A$ClaimNb_acum_t", i, sep=""))), function(x) {result <- min(x,
    6);return(result)})}
```

Bonus–Malus

C43

```
set.seed(1234)
bm_base <- 100
bm <- data.table(matrix(1:((1+t_max)*(1+k_max))))
bm$t <- rep(0:t_max, each = (k_max+1))
bm$k <- rep(0:k_max, times = (t_max+1))
theta <- fit_nb$theta
tau <- theta/mean(A$prob_m_rec_anual)
bm$prima_aj <- round(bm_base*(tau/theta)*((theta+bm$k)/(tau+bm$t)), 2)
```

Análisis de la simulación de la tarificación *a posteriori*

C44

```
for(i in 1:t_max_sim) {A <- merge(A,bm[t==(min(i,t_mem)),],
    by.x=paste("ClaimNb_acum_t",i,sep=""), by.y="k"); colnames(A)[ncol(A)] <- paste("bm_t",
    i, sep=""); A$V1 <- NULL; A$t <- NULL}
for(i in 1:t_max_sim) {A[[paste("pp_t", i, sep="")]] <- A$pp_apriori*eval(parse(text =
    paste("A$bm_t", i, sep="")))/bm_base}
bm_print_wide <- dcast(bm[,c(2:4), with=FALSE],t ~ k,value.var="prima_aj")
```

C45

```
s <- c()
s[1] = sum(A$pp_apriori)
for(i in 2:(t_max_sim+1)) {s[i]=sum(eval(parse(text =paste("A$pp_t", (i-1), sep=""))))}
sim <- t(data.table(c(0,sum(A$prob_m_rec_anual))))
for(i in 1:t_max_sim) {sim <- rbind(sim,c(i,sum(eval(parse(text =paste("A$ClaimNb_t", i,
    sep=""))))))}
sim <- cbind(sim,s)
sim <- data.table(sim)
for(i in 1:(t_max_sim+1)) {s[i]=v_rec_anual*(sim$V2[[i]]/sim$V2[[1]])}
sim <- cbind(sim,s); colnames(sim) <- c("n_sim","n_rec","prima_pura_total","v_rec_simulada")
```

C46

```
desv <- ((sum(sim$prima_pura_total[6:16]) - sum(sim$v_rec_simulada[6:16])) /
    sum(sim$v_rec_simulada[6:16])) * 100
```

C47

```
p_sim <- data.table(table(eval(parse(text =paste("A$pp_t", (t_max_sim), sep="")))))
p_sim$V1 <- as.numeric(as.character(p_sim$V1))
```

C48

```
AA <- A[,c(1:16,40,78:92),with=FALSE]
pol_ej <- cbind(c(0:15), t(AA[AA$IDpol==1104007|AA$IDpol==115255|AA$IDpol==26421, c(17:32),
    with=FALSE]))
colnames(pol_ej) <- c("t","p1","p2","p3")
```

Anexo 3

Cuadro 19

Guía tabla de tarifas

Dimensión	Variable	*Codigo*		
		0	1	2
VPo	Potencia del vehículo	4, 8, 12, 13	6, 7, 14, 15	5, 9, 10, 11
VAg	Antigüedad del vehículo (años)	<8	8-12	+12
VBr	Marca del vehículo	B1, 2, 6, 10, 14	B3, 4, 5, 11, 13	B12
DrivAge	Edad del asegurado (años)	<27	+27	-
VehGas	Combustible	Gasolina	Diesel	-
Density	Densidad lugar residencia ($habs/km^2$)	<535,98	535,98-1439,72	+1439,72

Cuadro 20

Tabla de tarifas

VPo	*VAg*	*VBr*	*DrivAge*											
			0	0	0	0	0	0	1	1	1	1	1	1
			VehGas											
			0	0	0	1	1	1	0	0	0	1	1	1
			Density											
			0	1	2	0	1	2	0	1	2	0	1	2
0	0	0	226,26	264,44	291,70	212,14	247,93	273,50	126,46	147,80	163,04	118,57	138,58	152,86
0	0	1	233,85	273,31	301,49	219,26	256,25	282,68	130,70	152,76	168,51	122,55	143,23	157,99
0	0	2	312,48	365,21	402,87	292,98	342,42	377,73	174,65	204,13	225,17	163,75	191,39	211,12
0	1	0	205,85	240,58	265,39	193,00	225,57	248,82	115,05	134,47	148,33	107,87	126,07	139,07
0	1	1	212,75	248,66	274,29	199,48	233,14	257,17	118,91	138,98	153,31	111,49	130,31	143,74
0	1	2	284,29	332,27	366,53	266,55	311,53	0,00	158,90	185,71	204,86	148,98	174,12	192,07
0	2	0	159,20	186,06	205,24	149,26	174,45	192,43	88,98	103,99	114,72	83,42	97,50	107,56
0	2	1	164,54	192,30	212,13	154,27	180,30	198,89	91,96	107,48	118,56	86,22	100,77	111,16
0	2	2	219,86	256,97	283,46	0,00	0,00	0,00	122,89	143,62	158,43	115,22	134,66	148,54
1	0	0	263,67	308,16	339,93	247,21	288,93	318,72	147,37	172,24	190,00	138,17	161,49	178,14
1	0	1	272,51	318,50	351,34	255,51	298,62	329,41	152,31	178,02	196,37	142,81	166,91	184,12
1	0	2	364,15	425,60	469,48	341,42	399,03	440,18	203,53	237,88	262,40	190,83	223,03	246,02
1	1	0	239,88	280,36	309,27	224,91	262,86	289,96	134,07	156,70	172,86	125,71	146,92	162,07
1	1	1	247,93	289,77	319,64	232,46	271,68	299,69	138,57	161,96	178,66	129,92	151,85	167,51
1	1	2	331,30	387,20	427,12	310,62	363,04	400,47	185,17	216,42	238,73	173,61	202,91	223,83
1	2	0	185,52	216,82	239,18	173,94	203,29	224,25	103,69	121,19	133,68	97,22	113,62	125,34
1	2	1	191,74	224,10	247,20	179,77	210,11	231,78	107,17	125,25	138,17	100,48	117,44	129,54
1	2	2	256,22	299,45	330,33	240,22	280,76	309,71	143,20	167,37	184,63	134,27	156,92	173,10
2	0	0	277,24	324,03	357,44	259,94	303,80	335,13	154,96	181,11	199,78	145,29	169,80	187,31
2	0	1	286,55	334,90	369,43	268,66	314,00	346,37	160,16	187,18	206,48	150,16	175,50	193,60
2	0	2	382,90	447,51	493,65	359,00	419,58	462,84	214,01	250,12	275,91	200,65	234,51	258,69
2	1	0	252,23	294,79	325,19	236,49	276,40	304,89	140,98	164,77	181,76	132,18	154,48	170,41
2	1	1	260,70	304,69	336,10	244,43	285,67	315,13	145,71	170,30	187,86	136,61	159,67	176,13
2	1	2	348,35	407,14	449,12	326,61	381,73	421,09	194,70	227,56	251,02	182,55	213,36	235,35
2	2	0	195,07	227,99	251,49	182,89	213,76	235,80	109,03	127,43	140,56	102,22	119,47	131,79
2	2	1	201,61	235,64	259,93	189,03	220,93	243,71	112,69	131,70	145,28	105,65	123,48	136,21
2	2	2	269,41	314,87	347,33	252,59	0,00	325,66	150,58	175,99	194,13	141,18	165,00	182,02

Bibliografía

Adamidis, K. (1999). Theory and methods: An em algorithm for estimating negative binomial parameters. *Australian and New Zealand Journal of Statistics*, 41(2):213–221.

Alcañiz Zanón, M. y Ayuso Gutiérrez, M. y. P. M. A. M. (2014). *El seguro basado en el uso*. Fundación MAPFRE.

Baione, F., Levantesi, S., y Menzietti, M. (2002). The development of an optimal bonus-malus system in a competitive market. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 32(01):159–170.

Bermudez, L., Denuit, M., y Dhaene, J. (2001). Exponential bonus-malus systems integrating a prior risk classification. *Journal of Actuarial Practice*, 9:84–112.

Bonche, S., Brau, L., y Olympio, N. (2005). Decreasing the deductive in an automobile insurance policy. *ISFA*, (02).

Cameron, Adrian Colin. y Trivedi, P. K. (1985). Regression based tests for overdispersion.

Carrillo, M., Bermúdez, L., y Guillén, M. (2004). Solidaridad entre asegurados: ¿existen alternativas a los sistemas bonus-malus. *Anales del Instituto de Actuarios Españoles*, Tercera Época(10):55–90.

Charpentier, A. (2015). *Computational actuarial science with R*. Boca Raton : CRC Press.

DeGroot, M. H. (1986). *Probability and statistics*. Addison-Wesley Pub. Co.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman Hall.

Frangos, N. E. y Vrontos, S. D. (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 31(01):1–22.

Frees, E. W. (2010). *Regression modeling with actuarial and financial applications*. Cambridge University Press.

Greene, W. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Greenwood, M. y Yule, G. (1920). An inquiry in to the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society*, 83:255–279.

Hickman, J. C. y Heacox, L. (1999). Credibility theory. *North American Actuarial Journal*, 3(2):1–8.

Holtan, J. (1994). Bonus made easy. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 24(01):61–74.

Instituto de Actuarios Españoles, I. (2014). Riesgos del automóvil. *Actuarios*, (34).

Kaas, R. (2009). *Modern actuarial risk theory: using R*. Springer.

Kupper, J. (1963). Some aspects of cumulative risk. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 3(01):85–103.

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance*. Kluwer Scademic Publishers.

Lemaire, J., Park, S. C., y Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 46(01):39–69.

Lemaire, J. y Zi, H. (1994). A comparative analysis of 30 bonus-malus systems. *ASTIN Bulletin: The Journal of the International Actuarial Association*, 24(02):287–309.

Lewis, F., Butler, A., y Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2):155–162.

Álvarez Jareño, J. A. y Rodríguez, P. M. (2010). Reparametrización de las principales distribuciones de probabilidad en el estudio del número de siniestros debido a las anomalías muestrales en las carteras del seguro de responsabilidad civil de automóviles. determinación del índice de dispersión. *Anales del Instituto de Actuarios Españoles*, Tercera Época(16):1–24.

McCullagh, P. y Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Melgar Hiraldo, M. d. C., Ordaz Sanz, J. A., y Guerrero Casas, F. M. (204). Una estimación del número de siniestros en el seguro del automóvil: comparación entre distintos modelos. *XII Jornadas de ASEPUMA*.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370.

Pérez Sánchez, J. M. (2011). Un modelo bonus-malus con asignación de tarifas más competitivas en el mercado de seguro de automóviles.