UNIVERSITAT DE BARCELONA

Modelos basados en distancias con aplicación a la gestión del riesgo en el ámbito actuarial

Mª Teresa Costa Cor

PROGRAMA DE DOCTORADO EN EMPRESA

TESIS DOCTORAL:

MODELOS BASADOS EN DISTANCIAS CON APLICACIÓN A LA GESTIÓN DEL RIESGO EN EL ÁMBITO ACTUARIAL

Mª Teresa Costa Cor

Universitat de Barcelona

Directores:

Dra. Eva Boj del Val

Dr. Josep Fortiana Gregori

Índice

Capítulo 1

Introducción

El presente trabajo, **Modelos basados en distancias con aplicación a la gestión del riesgo en el ámbito actuarial**, está dedicado al estudio de metodologías estadísticas, complementarias o alternativas a las existentes en la bibliografía estadística y actuarial, para la solución de problemas reales de las carteras de seguros no vida. Estos problemas son, por ejemplo, el de la tarificación *a priori*, el de *credit scoring* y el de cálculo de provisiones técnicas.

En el proceso de tarificación *a priori* el objetivo es asignar una prima, precio del seguro, a una nueva póliza que se incorpora en la cartera de seguros de un asegurador. El primer paso del proceso de construcción del modelo es la selección de las variables de tarifa, que consiste en la elección de los factores de riesgo o características que se utilizan para distinguir a los asegurados/pólizas con diferentes riesgos asociados. En los modelos de predicción utilizados para la tarificación es fundamental poder evaluar la importancia relativa de los factores de riesgo o predictores y calibrar su influencia en la respuesta, es decir, en la siniestralidad esperada, para realizar su predicción.

En *credit scoring* se busca una regla que permita clasificar nuevos individuos entre aquellos que podrán hacer frente a sus obligaciones crediticias con alta probabilidad y entre aquellos que, por el contrario, resultarán fallidos. A partir de un conjunto de observaciones cuya pertenencia a una determinada clase es conocida *a priori*, con el análisis de las características del solicitante y de las características de la operación, se podrán inducir las reglas que posteriormente se aplicarán a nuevas solicitudes, determinando así su clasificación. Para una entidad es fundamental elegir un buen modelo de puntuaciones (*scorings*) para el cálculo de las primas por riesgo de crédito a partir de las probabilidades de insolvencia de los riesgos.

Tradicionalmente, el cálculo de provisiones de los seguros de no vida se hacía de forma determinista. Sin embargo, Solvencia II (Directiva 2009/138/CE del Parlamento Europeo y del Consejo de 25 de noviembre de 2009) establece que las entidades aseguradoras deben ser capaces de calcular sus provisiones técnicas con la mejor estimación de los pagos futuros y con un margen de riesgo, lo que da pie a la introducción de métodos estocásticos en dichos cálculos.

Los métodos deterministas únicamente hacen suposiciones sobre el valor esperado de los pagos futuros. Los modelos estocásticos también permiten modelizar la variación de los pagos futuros, es decir, no sólo proporcionan estimaciones del valor esperado de los pagos futuros sino también de la variación respecto al valor esperado.

Los objetivos del presente trabajo son:

- Realizar una revisión de algunas metodologías estadísticas de la literatura actuarial aplicadas a los problemas de tarificación, *credit scoring* y cálculo de provisiones técnicas de los seguros no vida.

- Realizar el estudio teórico de los métodos basados en distancias para constituir una herramienta alternativa o complementaria en la tarificación, el *credit scoring* y el cálculo de provisiones técnicas de los seguros no vida, fundamentalmente el modelo lineal generalizado basado en distancias.

- Ilustrar la aplicabilidad a tarificación, *credit scoring* y cálculo de provisiones técnicas de los seguros no vida de las metodologías ya existentes y de la propuesta en el trabajo con datos de carteras de seguros no vida.

El inicio de esta investigación se remonta al año 2011, cuando la Dra. Eva Boj me ofreció comenzar a trabajar con los métodos basados en distancias aplicados en el ámbito actuarial siguiendo la trayectoria de investigación iniciada por ella desde hacía años en Boj *et al.* (2004), con la propuesta de dirigir mi Tesis Doctoral junto con el Dr. Josep Fortiana.

En mayo de 2012 me incorporé al Proyecto de Investigación "Métodos semiparamétricos y basados en distancias con aplicaciones en bioinformática, finanzas y gestión del riesgo" con número de referencia MTM2010-17323, cuya investigadora principal era la Dra. Aurea Grané y del cual formaban parte como investigadores la Dra. Eva Boj y el Dr. Josep Fortiana.

Posteriormente, desde enero de 2015, junto con la Dra. Eva Boj y el Dr. Josep Fortiana, formo parte del grupo de investigadores del Proyecto de Investigación "Selección de factores en la gestión del riesgo. Contribuciones al sector asegurador con especial atención a la longevidad y dependencia", con número de referencia MTM2014-56535-R, cuya investigadora principal es la Dra. Aurea Grané.

Sin duda, la oportunidad de poder participar en dichos proyectos de investigación financiados ha supuesto una gran ayuda a nivel económico para poder, sobretodo, asistir a congresos tanto nacionales como internacionales. Se han presentado distintos trabajos a lo largo de estos años en congresos de carácter estadístico y actuarial y la asistencia a los mismos ha permitido conocer la investigación que se está realizando en nuestro ámbito de estudio.

El trabajo se estructura de la siguiente manera:

En el Capítulo 2 se revisan los modelos de regresión. En primer lugar, se presenta teóricamente el modelo lineal generalizado (MLG) y después, en el ámbito de las distancias, se describen el modelo de regresión lineal basado en distancias (MLBD) y el modelo lineal generalizado basado en distancias (MLGBD) a nivel teórico (ver Boj *et al.*, 2015b). Por último se presentan las metodologías de *bootstrap* que se van a aplicar en los modelos de regresión en capítulos siguientes.

El Capítulo 3 se dedica al análisis de coeficientes de influencia locales para el MLGBD, que se definen para medir la importancia relativa de cada variable observada y son válidos entorno a un valor dado en el espacio predictor (ver Boj *et al.*, 2015a). Además, se construyen intervalos de confianza para los coeficientes de influencia anteriores basados en el percentil de la distribución *bootstrap* a partir de una adaptación del test de Wald (ver Boj y Costa, 2015a).

El Capítulo 4 está dedicado al riesgo de crédito y el cálculo de puntuaciones, aplicando el modelo de regresión logística basado en distancias (BD) para estimar las probabilidades de insolvencia de los nuevos asegurados que se incorporan en la cartera (ver Costa *et al.*, 2012). Uno de los objetivos es minimizar la probabilidad de mala clasificación de los nuevos individuos para evitar conceder un crédito a un mal riesgo de crédito o denegarlo a un buen riesgo de crédito y analizar los costes de dicha clasificación incorrecta.

El Capítulo 5 incluye el desarrollo de la metodología aplicada al cálculo de provisiones técnicas en los seguros no vida. Se realiza un primer análisis de los métodos deterministas que se han aplicado durante décadas en el cálculo de provisiones y, a continuación, se presentan los principales métodos estocásticos para la estimación de los pagos futuros que deberá realizar la entidad aseguradora. Entre los modelos estocásticos clásicos cabe destacar el MLG, para el cuál se construye la formulación relativa al error de predicción cometido en los pagos futuros por años de calendario (ver Boj *et al.*, 2014b y Boj y Costa, 2015c). Como alternativa al MLG clásico para el problema del cálculo de provisiones se propone su versión basada en distancias, es decir, el MLGBD. Por último, se definen diferentes formas de añadir márgenes de riesgo al mejor estimador teniendo en cuenta el contexto de Solvencia II.

A lo largo del trabajo se desarrollan aplicaciones prácticas para ilustrar la aplicabilidad de las dos metodologías estudiadas con detalle en el trabajo: el MLG y el MLGBD, que se corresponde con el tercer objetivo.

La primera aplicación utiliza datos del seguro de automóviles a terceros de Suecia que se refieren a la frecuencia de siniestralidad, considerando tres factores de riesgo. Estos datos se describen en Hallin e Ingenbleek (1983) y se utilizan para ilustrar el cálculo de los coeficientes de influencia definidos en el trabajo y la construcción de intervalos de confianza apropiados en el MLGBD.

La segunda aplicación utiliza dos conjuntos de datos reales sobre riesgo de crédito, disponibles en http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) y en http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval), de una entidad

financiera alemana y de una entidad financiera australiana, respectivamente. Se usan para ilustrar la aplicación del modelo de regresión logística BD en el problema del *credit scoring.*

En la tercera aplicación se usan los datos incluidos en el trabajo de Taylor y Ashe (1983) referentes a importes de siniestros pagados durante diez años, recogidos en un triángulo *run-off,* que han sido utilizados por diferentes autores en sus ilustraciones numéricas dentro de la literatura actuarial. Con estos datos se realiza la estimación de los pagos futuros y el cálculo de las provisiones siguiendo las recomendaciones de Solvencia II.

En las aplicaciones se utiliza el software *R* (R Development Core Team, 2015), ejecutando las funciones incluidas en algunas librerías o programando los cálculos que se requieren. Cabe destacar el uso de la librería *dbstats* de Boj *et al.* (2014a) en la que se han implementado los modelos de regresión basados en distancias. Este software se va describiendo durante el trabajo.

El trabajo finaliza con un capítulo de conclusiones, en el que se destacan las aportaciones que se han realizado.

Capítulo 2

Modelos de regresión

En este capítulo se describen modelos de regresión. En primer lugar el MLG como una extensión del modelo de regresión lineal clásico.

En segundo lugar se describen los modelos de regresión basados en distancias, tanto el MLBD como el MLGBD.

Finalmente, se comentan dos metodologías de remuestreo (*bootstrap*) que son adecuadas para su aplicación en los modelos de regresión.

Este capítulo se estructura en tres apartados:

En el primer apartado, 2.1., se describe el MLG desde el punto de vista teórico. Se detallan sus características, las expresiones que permiten calcular la esperanza y varianza para una determinada familia paramétrica de distribuciones, y la estimación de sus parámetros.

En el segundo apartado, 2.2, en primer lugar se muestra el concepto de distancia entre individuos de una población para, a continuación, presentar los modelos de regresión en el ámbito de las distancias. Se describen, desde un punto de vista teórico, los principales elementos y características del MLBD y del MLGBD. Y, por último, se describe el software informático que se utiliza en las aplicaciones prácticas de los capítulos posteriores de este trabajo. En particular la librería *dbstats* (Boj *et al.*, 2014a) de *R* donde se han implementado los modelos basados en distancias.

En el tercer apartado, 2.3., se comentan las principales características de dos metodologías de remuestreo que se aplican en los modelos de regresión: *bootstrapping residuals* y *bootstrapping pairs*, y que se utilizan en algunas aplicaciones prácticas de este trabajo.

2.1 Modelo lineal generalizado

2.1.1. Descripción del modelo

En este capítulo se describe a grandes rasgos el MLG con el objetivo de seguir el resto de capítulos del trabajo. Para una descripción detallada del MLG y sus características se puede consultar, por ejemplo, McCullagh y Nelder (1989) y Wood (2006), y para el detalle de su aplicación al problema de tarificación *a priori* de los seguros no vida Boj *et al.* (2004).

Se supone la variable aleatoria *Y*, con $\boldsymbol{y}$ un vector de dimensión $n\times 1$, y_i para $i=1,2,\ldots,n$ observaciones independientes, las cuales recogen la siniestralidad a explicar y juegan el papel de variable respuesta en el modelo. Se suponen *p* predictores o factores potenciales de la estructura de riesgo $\boldsymbol{F}_1,\boldsymbol{F}_2,\ldots,\boldsymbol{F}_p$, vectores de dimensión $n\times 1$, de manera que se tienen f_{ij} para $i=1,2,\ldots,n$ y $j=1,2,\ldots,p$.

En el modelo clásico de regresión lineal por mínimos cuadrados ordinarios, en el que se supone una distribución del error ε_i Normal centrada y con varianza constante, $\varepsilon_i \sim N\left(0,\sigma^2\right)$, la relación lineal de la respuesta con la estructura sistemática dada por los predictores es:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} + \varepsilon_i. \tag{2.1}$$

Se dispone de observaciones independientes $y_i \sim N\left(\mu_i,\sigma^2\right)$, con esperanza

$$E\left[y_i\right] = \mu_i = \boldsymbol{x_i}\cdot\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p}\beta_j f_{ij}, \tag{2.2}$$

y varianza constante

$$V\left[y_i\right] = \sigma^2. \tag{2.3}$$

En el MLG se siguen teniendo y_i para $i = 1, 2, \ldots, n$ observaciones independientes de la respuesta, unos errores centrados $E[\varepsilon_i] = 0$, y un predictor lineal determinista al que se simboliza por η_i:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}. \tag{2.4}$$

Las dos extensiones respecto al modelo de regresión lineal son:

1) La distribución de *Y* puede ser distinta de la distribución Normal. En primer lugar, se considera que puede ser cualquier distribución derivada de la familia exponencial de McCullagh y Nelder (ver McCullagh y Nelder, 1989), que se caracterizan por tener una función de densidad de probabilidad en un punto de la forma:

$$f(y_i; \theta_i, \phi_i) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\}, \tag{2.5}$$

para funciones especificadas $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$. En esta función, a θ_i se le denomina parámetro canónico y a ϕ_i se le denomina parámetro de dispersión.

El logaritmo de la función de verosimilitud es:

$$l(\theta_i; y_i) = \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i). \tag{2.6}$$

A partir de esta formulación se puede deducir la media y la varianza de *Y* a partir de las relaciones:

$$E\left[\frac{\partial l}{\partial \theta_i} \right] = 0$$

$$E\left[\frac{\partial^2 l}{\partial \theta_i^2} \right] + E\left[\frac{\partial l}{\partial \theta_i} \right]^2 = 0$$

De la expresión (2.5) se cumple que:

$$\frac{\partial l}{\partial \theta_i} = \frac{\left\{ y_i - b'(\theta_i) \right\}}{a_i(\phi_i)}$$
$$\frac{\partial l^2}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a_i(\phi_i)}$$

y se deduce que:

$$E[y_i] = \mu_i = b'(\theta_i), \tag{2.7}$$

$$V[y_i] = b''(\theta_i)\, a(\phi_i). \tag{2.8}$$

De esta manera, la varianza es el producto de dos componentes:

- la primera, $b''(\theta_i)$, depende únicamente del parámetro canónico θ_i (y por tanto de la esperanza μ_i por (2.7)). Esta componente se denomina función de varianza y se explicita su dependencia respecto de la esperanza: $b''(\theta_i) = V(\mu_i)$,

- la segunda, $a(\phi_i)$, depende sólo del parámetro de dispersión ϕ_i y usualmente adopta la forma $a(\phi_i) = \frac{\phi}{w_i}$, con parámetro de dispersión constante para todas las observaciones, ϕ, y unos pesos especificados *a priori*, w_i, que varían de observación a observación.

Teniendo en cuenta lo anterior se puede reescribir:

$$V[y_i] = \frac{\phi V(\mu_i)}{w_i}, \tag{2.9}$$

donde ϕ es el parámetro de dispersión, $V(\mu_i)$ es la función de varianza y w_i es el posible peso especificado *a priori* de la observación *i*. Si se supone $\phi = 1$, los recíprocos de los pesos pueden reinterpretarse como parámetros de escala no constantes: $1/w_i = \phi_i$.

Se puede utilizar para la función de varianza la familia paramétrica de distribuciones de

error (2.5), de manera que:

$$V(\mu_i) = \mu_i^{\xi}. \tag{2.10}$$

Se observa que con esta familia de distribuciones del error se obtienen algunos casos particulares de la familia exponencial, por ejemplo:

- Si $\xi = 0$, la distribución del error es Normal.
- Si $\xi = 1$, la distribución del error es Poisson.
- Si $\xi = 2$, la distribución del error es Gamma.
- Si $\xi = 3$, la distribución del error es Inversa Gaussiana.

2) La respuesta está ligada con el predictor lineal a través de una función $g(\cdot)$, denominada función de enlace, de manera que:

$$\eta_i = g(\mu_i) = g(E[y_i]) = \boldsymbol{x_i} \cdot \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}. \tag{2.11}$$

De aquí se deriva que:

$$E[y_i] = \mu_i = g^{-1}(\eta_i) = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \right). \tag{2.12}$$

A la función de enlace, $g(\cdot)$, se le exige que sea monótona y diferenciable. Existen, para algunas distribuciones de la familia exponencial, funciones de enlace denominadas canónicas, para las que se cumple que el predictor lineal coincide con el parámetro canónico: $\theta(\mu_i) = \eta_i$.

En la Tabla 2.1 se recogen las distribuciones más conocidas que forman parte de la familia exponencial definida en (2.5), junto con el enlace canónico asociado.

Tabla 2.1. Tabla de propiedades para casos particulares del MLG.

	Normal o Gaussiana $N\left(\mu,\sigma^2\right)$	Binomial $B(m,\pi)/m$	Poisson $P(\mu)$	Gamma $G(\mu,\nu)$	Inversa Gaussiana $GI\left(\mu,\sigma^2\right)$
Rango de Y	$(-\infty,+\infty)$	$0,1,\ldots,m/m$	$0,1,\ldots,\infty$	$(-\infty,+\infty)$	$(-\infty,+\infty)$
Peso w	1	1	1	1	1
Parámetro de dispersión: ϕ	σ^2	$1/m$	1	υ^{-1}	σ^2
$b(\theta)$	$\frac{\theta^2}{2}$	$\log\left(1+e^{\theta}\right)$	$\exp(\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y;\theta)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$log\left(\begin{matrix}m\\my\end{matrix}\right)$	$-\log(\boldsymbol{y}!)$	$\nu\log(\nu y)-\log(y)$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\left(\log\left(2\pi\phi y^3\right)+\frac{1}{\phi y}\right)$
$\mu(\theta)=E(Y;\theta)$	θ	$\frac{e^{\theta}}{\left(1+e^{\theta}\right)}$	$\exp(\theta)$	$\frac{-1}{\theta}$	$(-2\theta)^{-1/2}$
Función de enlace canónica: $\theta(\mu)$	Identidad μ	Logit: $\log\left(\frac{\mu}{1-\mu}\right)$	Logarítmico: $\log(\mu)$	Recíproco: $\frac{1}{\mu}$	$\frac{1}{\mu^2}$
Función de varianza: $V(\mu)$	1	$\mu(1-\mu)$	μ	μ^2	μ^3

En general se opta por cualquier enlace que no sea el canónico y es usual utilizar uno derivado de la familia de enlaces paramétricos:

$$\eta_i = g(\mu_i) = \begin{cases} \mu_i^{\lambda} & \text{para} \quad \lambda \neq 0 \\ \log(\mu_i) & \text{para} \quad \lambda = 0. \end{cases} \tag{2.13}$$

Al aplicar MLG a unos datos, se decide sobre la distribución del error y sobre la función de enlace a utilizar en el modelo. La elección del enlace canónico tiene la ventaja de simplificar la formulación, pero no tiene porqué implicar que sea el más adecuado para unos datos particulares. Si el objetivo es seleccionar un modelo, la simplicidad de la función de enlace no debe sustituir a la calidad del ajuste como criterio.

En el MLG se obtiene un modelo aditivo si se combina cualquier distribución del error con la función enlace identidad, y se obtiene un modelo multiplicativo si se utiliza la función de enlace logarítmica. Según (2.11), que relaciona la respuesta con el predictor lineal, se tiene que para el enlace identidad:

$$\mu_i = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}. \tag{2.14}$$

Mientras que para el enlace logarítmico:

$$\log(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij},$$

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}\right) = e^{\beta_0} e^{\beta_1 f_{i1}} e^{\beta_2 f_{i2}} \cdots e^{\beta_p f_{ip}}. \tag{2.15}$$

Se observa que, partiendo de la base β_0, en el modelo aditivo debe sumarse (o restarse) las *p* cantidades $\sum_{j=1}^{p} \beta_j f_{ij}$. En el modelo multiplicativo, partiendo de la base e^{β_0}, debe multiplicarse por las *p* cantidades de incremento (o decremento) $\prod_{j=1}^{p} e^{\beta_j f_{ij}}$.

En el caso particular de predictores binarios, partiendo del efecto global, se suma β_j o se

multiplica por e^{β_j}, sólo cuando para el individuo *i* se dé la característica *j*, para $j = 1, \ldots, p$, ya que en tal caso se cumple que $f_{ij} = 1$.

2.1.2. Estimación de parámetros

La estimación de los parámetros β_j del predictor lineal se realiza mediante la maximización del logaritmo de la función de verosimilitud total: $l(\boldsymbol{\theta}; y) = \sum_{i=1}^{n} l(\theta_i; y_i)$, que ya se ha definido en (2.6).

En el caso en que se supone una distribución del error Normal y la función de enlace identidad se obtiene como caso particular la solución por mínimos cuadrados ordinarios del modelo de regresión lineal.

La versión más general del MLG no exige que la distribución del error de *Y* pertenezca a la familia de distribuciones exponenciales caracterizadas por (2.5). En dicha versión más general se sigue teniendo y_i para $i = 1, 2, \ldots, n$ observaciones independientes de la respuesta, y para éstas se conocen sólo los dos primeros momentos (2.7) y (2.9). La estimación de los parámetros β_j se realiza maximizando los logaritmos de las funciones de cuasi-verosimilitud:

$$q(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} w_i \int_{y_i}^{\mu_i} \frac{y_i - s}{\phi V(s)} ds. \tag{2.16}$$

Para maximizar (2.16) debe resolverse el siguiente sistema de ecuaciones lineales:

$$\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j \tag{2.17}$$

mediante la aplicación de algún método numérico. Para este caso general de la familia exponencial, las cuasi-verosimilitudes juegan el papel de verosimilitudes.

En McCullagh y Nelder (1989) y Wood (2006) se describe y se justifica el algoritmo de

estimación de parámetros, β_j, del predictor lineal por máxima verosimilitud, mediante un proceso iterativo de mínimos cuadrados ponderados iterativos. En dicho proceso la variable dependiente de las sucesivas regresiones no es la respuesta ***y*** original, sino ***Z***, una forma linealizada de la función de enlace que se aplica a ***y***, y los pesos de cada iteración, ***w***, dependen únicamente de los valores ajustados, $\hat{\boldsymbol{\mu}}$. El proceso es iterativo porque tanto la variable dependiente ajustada ***Z*** como los pesos ***w*** dependen de los sucesivos valores. Esta característica es importante, pues en el MLGBD se utilizará también el algoritmo iterativo de mínimos cuadrados ponderados para la estimación del modelo y, al depender únicamente de los valores iniciales ajustados de la respuesta, es posible aplicar el método cuando la información inicial dada del espacio predictor es únicamente una matriz de distancias al cuadrado.

Con algo de detalle, el proceso iterativo consta de los siguientes pasos. Sea $\hat{\boldsymbol{\eta}}_0$ el estimador actual del predictor lineal, con los valores ajustados correspondientes $\hat{\boldsymbol{\mu}}_0$ derivados de la función de enlace, $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. Se construye la variable dependiente ajustada con valor típico:

$$\boldsymbol{z}_0 = \hat{\boldsymbol{\eta}}_0 + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0) \cdot \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}} \right)_0, \tag{2.18}$$

donde se calcula la derivada del enlace en $\hat{\boldsymbol{\mu}}_0$. Se define el peso cuadrático como:

$$\boldsymbol{w}_{\boldsymbol{o}}^{-1} = \left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}} \right)_0^2 \cdot \frac{\boldsymbol{V}_0}{\boldsymbol{w}}, \tag{2.19}$$

donde $\boldsymbol{V}_{\boldsymbol{o}}$ es la función de varianza calculada en $\hat{\boldsymbol{\mu}}_0$. Entonces se hace la regresión de $\boldsymbol{z}_0$ sobre las covarianzas con pesos $\boldsymbol{w}_0$ para tener nuevas estimaciones de los parámetros $\hat{\boldsymbol{\beta}}_1$; a partir de éstos se construye una nueva estimación $\hat{\boldsymbol{\eta}}_1$ del predictor lineal. Se repite el proceso hasta que los cambios son suficientemente pequeños.

2.1.3. Desvianza y residuos

En el MLG la variabilidad no explicada por un modelo (fijada una función de enlace, una distribución del error y unos predictores) se plasma en la desvianza escalada.

Si $l(\boldsymbol{\theta};\boldsymbol{y})$ denota el logaritmo de la función de verosimilitud total del modelo en estudio y $l(\boldsymbol{\theta}^*;\boldsymbol{y})$ el logaritmo de la función de verosimilitud total del modelo saturado (aquel modelo que cumple que tiene tantos parámetros como individuos, y por lo tanto se cumple que $\hat{\mu}_i = y_i$ para $i = 1,2,\ldots,n$), entonces:

$$Dev(\boldsymbol{\theta};\boldsymbol{y}) = 2\left[l(\boldsymbol{\theta}^*;\boldsymbol{y}) - l(\boldsymbol{\theta};\boldsymbol{y})\right]\phi = \sum_{i=1}^{n} 2w_i\left[y_i(\theta_i^* - \theta_i) - b(\theta_i^*) + b(\theta_i)\right], \quad (2.20)$$

donde $\boldsymbol{\theta}^*$ es el estimador de máxima verosimilitud de $\boldsymbol{\theta}$ en el modelo saturado.

La desvianza escalada disminuye a mayor número de número de predictores incluidos en el modelo, hasta llegar a explicar la variabilidad total de los datos.

También se pueden calcular las desvianzas a partir de las cuasi-verosimilitudes definidas en (2.16) como:

$$Dev(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} 2w_i \frac{y_i - s}{V(s)} d(s) = -2\phi\, q(\boldsymbol{\theta};\boldsymbol{y}). \quad (2.21)$$

En este caso, para el cálculo de las desvianzas asociadas a un modelo tan sólo se necesita conocer los dos primeros momentos.

En la Tabla 2.2 se muestra la expresión que toma en algunos de los casos particulares de la familia exponencial, en términos de desvianzas no escaladas:

$$Dev(\boldsymbol{\theta};\boldsymbol{y}) = \phi Dev^*(\boldsymbol{\theta};\boldsymbol{y}). \quad (2.22)$$

Tabla 2.2. Desvianzas para casos particulares del MLG.

	Desvianzas
Normal o Gaussiana $N(\mu,\sigma^2)$	$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$
Binomial $B(m,\pi)/m$	$2\sum_{i=1}^{n}\{y_i\log(y_i/\hat{\mu}_i)+(m-y_i)\log[(m-y_i)/(m-\hat{\mu}_i)]\}$
Poisson $P(\mu)$	$2\sum_{i=1}^{n}\{y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)\}$
Gamma $G(\mu,\nu)$	$2\sum_{i=1}^{n}\{-\log(y_i/\hat{\mu}_i)+(y_i-\hat{\mu}_i)/\hat{\mu}_i\}$
Inversa Gaussiana $GI(\mu,\sigma^2)$	$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

Se definen, entre otros, dos tipos de residuos:

- Residuos de Pearson:

$$r_i^P=\frac{y_i-\hat{\mu}_i}{\left(V(\hat{\mu}_i)/w_i\right)^{1/2}}, \tag{2.23}$$

- Residuos de desvianza:

$$r_i^D=signo(y_i-\hat{\mu}_i)\sqrt{d_i}, \tag{2.24}$$

donde d_i es la i-ésima componente de (2.21).

Para la estimación del parámetro de dispersión ϕ de un modelo con $p+1$ coeficientes se puede utilizar la fórmula:

$$\hat{\phi}^D = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(r_i^D\right)^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} d_i, \tag{2.25}$$

o bien el estimador de momentos basado en los residuos generalizados de Pearson:

$$\hat{\phi}^P = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(r_i^P\right)^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} w_i \frac{\left(y_i-\hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)}. \tag{2.26}$$

Normalmente se indica que el MLG más apropiado para unos datos es aquél que ofrece una menor desvianza. Hay diferentes maneras de reducirla (Millenhall, 1999): si se varía la función de enlace; si se varía la distribución del error; y/o si se varían los factores de riesgo incluidos en el predictor lineal.

2.2 Modelos de regresión basados en distancias

2.2.1 Introducción

Las técnicas estadísticas que se basan en distancias, o similaridades, entre los individuos de una muestra tienen una amplia tradición.

La utilidad de los métodos estadísticos basados en distancias radica en el hecho de que, a menudo, se presentan situaciones en las que la única posibilidad de conocer las relaciones entre unidades estadísticas es calcular una matriz de distancias entre ellas. En Boj *et al.* (2015b) se citan algunos ejemplos de estas situaciones:

- Los individuos que participan en estudios de marketing o de psicología pueden señalar muchas veces cómo de similares o de diferentes son pares de objetos o estímulos pero encuentran dificultades a la hora de describirlos con un número finito de características medibles. Por lo tanto, un resultado común de estos estudios es una matriz de distancias (o similaridades) entre objetos (o estímulos).

- En una red social se puede observar una respuesta continua (por ejemplo, el consumo telefónico del último año) para varios individuos (por ejemplo, aquellos que tienen contrato con una determinada compañía telefónica) que se relacionan entre sí. En este caso, la compañía puede estar interesada en predecir el consumo potencial de otros individuos en la red social que no tienen contrato con ella y enfocar sus esfuerzos de marketing para atraer a aquellos clientes potenciales con un consumo esperado más elevado.

- Los conjuntos de variables de tipo mixto (mezcla de variables cuantitativas, binarias y cualitativas) y los datos funcionales constituyen ejemplos adicionales de estructuras de datos que pueden beneficiarse de las técnicas basadas en distancias.

2.2.2 Distancias y similaridades

Una distancia δ sobre un conjunto (finito o no) Ω es una aplicación que, a cada par de individuos $(i, j) \in \Omega \times \Omega$, le hace corresponder un número real $\delta(i, j) = \delta_{ij}$, que como mínimo cumple las siguientes propiedades básicas: $\delta_{ij} \geq 0, \delta_{ii} = 0$ y $\delta_{ij} = \delta_{ji}$. Se puede hablar en tal caso de disimilaridad. Cuando, además, se cumple la desigualdad triangular, $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$, y $\delta_{ij} = 0$ si y sólo si $i = j$, entonces la distancia es métrica.

Si para una distancia se cumplen las propiedades básicas, la desigualdad triangular y además se pueden encontrar puntos $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots x_{ir})$, $\boldsymbol{x}_j = (x_{j1}, x_{j2}, \ldots x_{jr})$ de $\mathbb{R}^r$ tales que permiten reproducir las distancias originales:

$$\delta_{ij}^2 = (\boldsymbol{x}_i - \boldsymbol{x}_j) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j)^T, \tag{2.27}$$

entonces la distancia es Euclídea y (Ω, δ) puede representarse mediante el espacio euclídeo $(\mathbb{R}^r, \delta)$.

Si Ω es un conjunto finito, que se define como $\Omega = (1, 2, \ldots, n)$, las distancias δ_{ij} se expresan mediante la matriz simétrica $\boldsymbol{\Delta}$, que se denomina matriz de distancias sobre Ω:

$$\boldsymbol{\Delta} = \begin{pmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn} \end{pmatrix}, \text{ con } \delta_{ii} = 0,\ \delta_{ij} = \delta_{ji}\,.$$

A continuación se define la distancia (o similaridad) en función de si se calculan para datos cuantitativos, datos cualitativos o datos mixtos.

Se supone que cada individuo i de Ω viene representado por un punto $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots x_{ip}) \in \mathbb{R}^p$. En el caso de datos cuantitativos, la distancia más familiar entre dos individuos i, j es la distancia ℓ^2:

$$d_2(i,j) = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{2.28}$$

que, además, es una distancia Euclídea.

Si los datos son cualitativos es conveniente trabajar con la disimilaridad, que es un concepto dual del de distancia. Una similaridad s sobre un conjunto Ω es una aplicación que a cada par de individuos $(i, j) \in \Omega \times \Omega$, le hace corresponder un número real $s_{ij} = s(i, j)$, que cumple las siguientes propiedades: $0 \le s_{ij} \le s_{ii} = 1,\ s_{ii} = 1$ y $s_{ij} = s_{ji}$.

Si Ω es un conjunto finito, que definimos como $\Omega = (1, 2, \ldots, n)$, se tiene la matriz simétrica ***S***, que se denomina matriz de similaridades:

$$\boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}, \text{ con } s_{ii} = 1,\ s_{ij} = s_{ji}\,.$$

El valor $s_{ij} = s(i, j)$ mide el grado de semejanza entre dos elementos i, j de manera que se aproxima a 1 cuando ambos son muy parecidos.

El concepto de similaridad se utiliza especialmente cuando se han introducido p características cualitativas sobre Ω, que se asocian a variables binarias, de manera que toman valor 1 cuando la característica está presente y 0 cuando está ausente.

Para pasar de similaridad a distancia puede hacerse de distintas maneras, entre las que se destacan:

$$\delta_{ij} = 1 - s_{ij}, \tag{2.29}$$

o bien

$$\delta_{ij} = \sqrt{1 - s_{ij}}. \tag{2.30}$$

En el segundo caso se obtiene siempre una distancia métrica e incluso Euclídea para muchas similaridades.

Cuando se dispone de un conjunto de variables de tipo mixto, es decir, mezcla de variables cuantitativas, binarias y cualitativas, un coeficiente apropiado para tratar la similaridad de estos datos es el coeficiente de similaridad de Gower, definido en Gower (1971) y que toma la siguiente expresión:

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - \frac{\left| x_{ih} - x_{jh} \right|}{G_h} \right) + a + \alpha_{ij}}{p_1 + (p_2 - d) + p_3}, \tag{2.31}$$

donde p_1 es el número de variables cuantitativas, a y d son el número de coincidencias positivas y negativas, respectivamente, para las p_2 variables dicotómicas, α_{ij} es el número de coincidencias para las p_3 variables cualitativas y G_h es el rango de la h-ésima variable cuantitativa.

Este coeficiente cumple $0 \leq s_{ij} \leq 1$ y para pasar de similaridad a distancia se puede aplicar tanto (2.29) como (2.30).

El coeficiente de similaridad de Gower es la suma diferentes coeficientes apropiados para cada tipo de variables (ver Boj *et al.*, 2004 para una explicación detallada), de manera que:

- Si sólo se tienen variables de tipo cuantitativo, utilizando (2.31) el coeficiente se reduce a la distancia:

$$d_{ij} = \left(\frac{1}{p} \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{G_k} \right)^{\frac{1}{2}}, \tag{2.32}$$

  que es métrica y Euclídea.

- Si sólo se tienen variables binarias, el coeficiente se reduce al coeficiente de Jaccard:

$$s_{ij} = \frac{a}{a+b+c}, \tag{2.33}$$

  donde a, b y c son las frecuencias de $(1,1)$, $(1,0)$ y $(0,1)$, respectivamente, de manera que $a+b+c+d=p$. Esta similaridad tiene un rango entre 0 y 1 y es métrica y Euclídea.

- Si sólo se tienen variables cualitativas, el coeficiente se reduce al coeficiente de coincidencias:

$$s_{ij} = \frac{\alpha_{ij}}{p}. \tag{2.34}$$

Esta similaridad tiene un rango entre 0 y 1 y posee tanto la propiedad métrica como Euclídea.

2.2.3 Modelo de regresión lineal basado en distancias

El MLBD fue inicialmente construido y estudiado en Cuadras (1989), Cuadras y Arenas (1990) y Cuadras *et al.* (1996), y posteriormente estudiado en Boj *et al.* (2004, 2007, 2010),

Esteve (2003) y Esteve *et al.* (2009). En estos trabajos se asume que para un conjunto de individuos puede disponerse de una matriz de distancias entre individuos así como del valor de una variable respuesta continua para cada individuo. La idea principal del MLBD es usar las coordenadas principales como variables explicativas en un modelo de regresión lineal.

En este apartado se describen los principales aspectos teóricos del MLBD.

Sea $\Omega = \{\Omega_1, \ldots, \Omega_n\}$ un conjunto de n individuos de una población dada. Para cada uno de ellos se observa la variable respuesta continua $Y : Y_1, \ldots Y_n$, por tanto, sea $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ un vector de dimensión $n \times 1$ con los valores observados. Sean $w_i \in (0,1)$ los pesos constantes positivos de Ω_i. El vector de pesos $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ de dimensión $n \times 1$ se estandariza para que la suma sea igual a 1, es decir, $\mathbf{1}^T \cdot \boldsymbol{w} = 1$, donde $\mathbf{1}$ es el vector de unos de dimensión $n \times 1$. Se asume que el vector de respuestas $\boldsymbol{y}$ es $\boldsymbol{w}$-centrado, es decir, $\boldsymbol{w}^T \cdot \boldsymbol{y} = 0$.

Los individuos de Ω se describen a partir de un conjunto de variables, es decir, predictores observados, que pueden ser datos de tipo mixto (cuantitativos, cualitativos y binarios) o incluso otro tipo datos, como cadenas de caracteres o funciones. Sea $\boldsymbol{F} = (F_1, \ldots, F_p)$ la matriz de dimensión $n \times p$ que contienen los valores de los p predictores mixtos.

Se define una distancia δ en Ω, que es una función de los predictores $\boldsymbol{F}$, que satisfaga la propiedad Euclídea y se denota por $\boldsymbol{\Delta}$ la matriz de dimensión $n \times n$ que contiene las distancias al cuadrado $\delta^2(\Omega_i, \Omega_j)$.

Se define la matriz de dimensión $n \times n$ con los productos internos como:

$$\boldsymbol{G}_w = -\frac{1}{2} \boldsymbol{J}_w \cdot \boldsymbol{\Delta} \cdot \boldsymbol{J}_w^T, \tag{2.35}$$

donde $\boldsymbol{J}_w$ es la matriz $\boldsymbol{w}$-centrada definida como $\boldsymbol{J}_w = \boldsymbol{I} - \mathbf{1} \cdot \boldsymbol{w}^T$.

Se calcula $\boldsymbol{X}_w$, la matriz $\boldsymbol{w}$-centrada de dimensión $n \times k$ que cumple $\boldsymbol{G}_w = \boldsymbol{X}_w \cdot \boldsymbol{X}_w^T$, tal que

$\boldsymbol{w}^T \cdot \boldsymbol{X}_w = 0$ y $k \geq r \equiv \text{rango}\left(\boldsymbol{G}_w\right)$. A esta matriz $\boldsymbol{X}_w$ se le denomina configuración Euclídea de $\boldsymbol{\Delta}$.

La descomposición $\boldsymbol{G}_w = \boldsymbol{X}_w \cdot \boldsymbol{X}_w^T$ existe si y sólo si $\boldsymbol{G}_w$ es una matriz semidefinida positiva y en tal caso $\boldsymbol{\Delta}$ se denomina Euclídea. Se denota como $\boldsymbol{g}_w$ al vector de dimensión $1 \times n$ que contiene los elementos de la diagonal de $\boldsymbol{G}_w$, que son necesariamente no negativos.

La traza de $\boldsymbol{G}_w$ dividido entre n extiende el concepto de variación total (es decir, la traza de la matriz de covarianzas) en el campo de las distancias. Se define la variabilidad geométrica de $\boldsymbol{\Delta}$ (Boj *et al.*, 2004) como:

$$V\left(\boldsymbol{\Delta}\right) = \frac{1}{2n^2}\mathbf{1}_n^T \cdot \boldsymbol{\Delta}^{(2)} \cdot \mathbf{1}_n = \frac{1}{n} tr\, \boldsymbol{G}_w. \tag{2.36}$$

Se dice que la respuesta *Y*, los pesos $\boldsymbol{w}$ y la matriz de distancias al cuadrado $\boldsymbol{\Delta}$ siguen un MLBD cuando la esperanza $\boldsymbol{w}$-centrada $\boldsymbol{\mu} = E\left[Y\right]$ (que es igual a $\boldsymbol{J}_w \cdot \boldsymbol{\mu}$) pertenece al espacio de columnas φ de $\boldsymbol{G}_w$. Se puede observar que φ es también el espacio de columnas de cualquier configuración Euclídea $\boldsymbol{X}_w$ de $\boldsymbol{\Delta}$, porque $\boldsymbol{G}_w = \boldsymbol{X}_w \cdot \boldsymbol{X}_w^T$.

Sean $\boldsymbol{y}$ los valores observados de la variable respuesta *Y*. La estimación de la regresión lineal basada en distancias con respuesta $\boldsymbol{y}$, pesos $\boldsymbol{w}$ y matriz de predicción $\boldsymbol{\Delta}$ se obtiene haciendo una regresión por mínimos cuadrados ponderados de $\boldsymbol{y}$ sobre una configuración Euclídea $\boldsymbol{w}$-centrada de $\boldsymbol{\Delta}$, $\boldsymbol{X}_w$, una configuración Euclídea latente.

Se asume que se dispone de un nuevo individuo Ω_{n+1} y que se tiene el vector $\boldsymbol{\delta}_{n+1}$ de dimensión $1 \times n$ con las distancias al cuadrado en el espacio predictor entre el nuevo individuo Ω_{n+1} y los n individuos previamente conocidos. Entonces, Ω_{n+1} se puede representar como un vector de dimensión *k*, $\boldsymbol{x}_{n+1}$, en las filas de $\boldsymbol{X}_w$. De esta manera, el valor de $\boldsymbol{y}$ predicho para Ω_{n+1} es $\boldsymbol{x}_{n+1} \cdot \hat{\boldsymbol{\beta}}$, donde $\hat{\boldsymbol{\beta}}$ es el vector de coeficientes de regresión estimados.

En la regresión basada en distancias, por lo general, no necesita explicitarse una configuración Euclídea, ni tampoco $\hat{\boldsymbol{\beta}}$ o $\boldsymbol{x}_{n+1}$, ya que los valores finales se obtienen directamente a partir de las distancias.

La matriz de predicción es:

$$\boldsymbol{H}_w = \boldsymbol{G}_w \cdot \left(\boldsymbol{D}_w^{1/2} \cdot \boldsymbol{F}_w^{+} \cdot \boldsymbol{D}_w^{1/2} \right), \tag{2.37}$$

donde $\boldsymbol{D}_w = diag(\boldsymbol{w})$ es la matriz diagonal cuyos elementos son los pesos $\boldsymbol{w}$, $\boldsymbol{F}_w = \boldsymbol{D}_w^{1/2} \cdot \boldsymbol{G}_w \cdot \boldsymbol{D}_w^{1/2}$ y $\boldsymbol{F}_w^{+}$ es la pseudo-inversa de Moore-Penrose de $\boldsymbol{F}_w$. Entonces, la respuesta predicha es:

$$\hat{\boldsymbol{y}} = \bar{y}_w 1 + \boldsymbol{H}_w \cdot \left(\boldsymbol{y} - \bar{y}_w 1 \right), \tag{2.38}$$

donde $\bar{y}_w = \boldsymbol{w}^T \cdot \boldsymbol{y}$ es la media ponderada de $\boldsymbol{y}$.

Finalmente, para la predicción de un nuevo individuo Ω_{n+1}, dadas $\boldsymbol{\delta}_{n+1}$, las distancias al cuadrado con los n individuos previamente conocidos, se tiene que:

$$\hat{y}_{n+1} = \bar{y}_w + \frac{1}{2} \left(\boldsymbol{g}_w - \boldsymbol{\delta}_{n+1} \right) \cdot \left(\boldsymbol{D}_w^{1/2} \cdot \boldsymbol{F}_w^{+} \cdot \boldsymbol{D}_w^{1/2} \right) \cdot \left(\boldsymbol{y} - \bar{y}_w 1 \right). \tag{2.39}$$

Por tanto, las estimaciones obtenidas con (2.38) y (2.39) son cantidades intrínsecas, que significa que se pueden expresar directamente como una función de las distancias o, de manera equivalente, de los productos internos (2.35).

En la regresión lineal basada en distancias, el rango r de la matriz de predicción (2.37), como sucede en la regresión lineal ordinaria, es igual al número de predictores linealmente independientes. Dado que para n casos, según la métrica elegida, r puede ser tan alta como $n-1$, podría dar lugar a un modelo sobreparametrizado. Un procedimiento adecuado para esta situación es reemplazar la pseudo-inversa $\boldsymbol{F}_w^{+}$ por una aproximación con menor rango. En Boj *et al.* (2012) se detalla cómo obtener la mejor aproximación de cualquier rango dado

$k, 1 \leq k \leq r,$ a partir de los valores singulares de $\boldsymbol{F}_w^+$. Al rango k usado para definir la pseudo-inversa $\boldsymbol{F}_w^+$ se le denomina rango efectivo.

Para seleccionar un valor adecuado de rango efectivo k se pueden tener en cuenta diferentes criterios. En la función *dblm* del paquete *dbstats* se han implementado los criterios que a continuación se listan y que se encuentran explicados con detalle en Boj *et al.* (2015b):

- OCV (*Ordinary Cross-Validation*): con este criterio se selecciona el rango efectivo que minimiza el estadístico de validación cruzada ordinaria que se corresponde con *leave-one-out*, es decir la eliminación de un caso para cada regresión.

- GCV (*Generalized Cross-Validation*): con este criterio se selecciona el rango efectivo que minimiza el estadístico de validación cruzada generalizada.

- AIC: con este criterio se selecciona el rango efectivo que minimiza el estadístico dado por la expresión del Criterio de Akaike.

- BIC: con este criterio se selecciona el rango efectivo que minimiza el estadístico dado por la expresión del Criterio de Información Bayesiana.

- *rel.gvar* o variabilidad geométrica relativa: este criterio es una alternativa que se fundamenta en el hecho de que la suma de todos los valores singulares de $\boldsymbol{F}_w$ es igual a la variabilidad geométrica de $\boldsymbol{\Delta}$. De esta manera, se permite fijar una proporción de variabilidad geométrica a explicar por el modelo resultante que debe alcanzarse con la suma de los valores singulares más altos de $\boldsymbol{F}_w$. Los valores que se pueden asignar deben estar entre 0 y 1, y el valor por defecto es 0.95

- *eff.rank* o rango efectivo: con este criterio el usuario puede elegir exactamente las dimensiones que desea incluir en el modelo. Los valores enteros que se pueden asignar deben comprenderse entre 1 y el número total de observaciones menos uno como máximo.

La regresión basada en distancias reproduce los resultados del modelo clásico de regresión lineal de mínimos cuadrados ponderados: si se parte de una matriz $\boldsymbol{X}_w$ $\boldsymbol{w}$-centrada de dimensión $n \times r$ de r predictores continuos correspondientes a n individuos y se define $\boldsymbol{\Delta}$ como la matriz de distancias al cuadrado Euclídeas entre las filas de $\boldsymbol{X}_w$, entonces $\boldsymbol{X}_w$ es una configuración Euclídea de $\boldsymbol{\Delta}$ y, en este caso, la matriz de predicción, la respuesta y las predicciones coinciden con los respectivos valores que se obtienen con mínimos cuadrados ponderados en el modelo lineal ordinario.

2.2.4 Modelo lineal generalizado basado en distancias

El MLGBD se ha definido y estudiado en Boj *et al.* (2012, 2014a y 2015b).

En este modelo se tienen los mismos elementos que en el MLBD. Las diferencias entre ambos casos son las mismas que hay entre el modelo de regresión lineal clásico y el MLG clásico:

1) La distribución de Y puede ser cualquier distribución derivada de la familia exponencial de McCullagh y Nelder (ver McCullagh y Nelder, 1989), como en cualquier modelo lineal generalizado

2) La relación entre el predictor lineal $\boldsymbol{\eta} = \boldsymbol{X}_w \cdot \boldsymbol{\beta}$, obtenido a partir de la configuración Euclídea latente $\boldsymbol{X}_w$, y la respuesta esperada, $\boldsymbol{\mu}$, viene dada por una función de enlace $g(\cdot): \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$

Se dice que la respuesta $\boldsymbol{y}$, los pesos $\boldsymbol{w}$ y la matriz de distancias al cuadrado $\boldsymbol{\Delta}$ siguen un MLGBD cuando la esperanza $\boldsymbol{w}$-centrada y transformada por la función de enlace, $\boldsymbol{\mu} = E[Y]$, es un vector del espacio de columnas φ de $\boldsymbol{G}_w$, que coincide con el espacio de columnas de cualquier configuración Euclídea de $\boldsymbol{X}_w$.

Para estimar un MLGBD se aplica un algoritmo estándar de mínimos cuadrados ponderados iterativo, como el que se ha descrito para el MLG clásico, en el que la regresión lineal basada

en distancias sustituye a la regresión lineal clásica en las fórmulas (2.18) y (2.19) para hacer la regresión de $\boldsymbol{z}_0$ sobre las covarianzas con pesos $\boldsymbol{w}_o$ y así obtener una nueva estimación $\hat{\boldsymbol{\eta}}_1$. Este proceso de estimación de mínimos cuadrados ponderados iterativo no depende de una $\boldsymbol{X}_w$ específica, de manera que los valores finales se obtienen directamente a partir de las distancias originales.

En el primer paso de este proceso iterativo es necesario un valor inicial $\hat{\boldsymbol{\mu}}_o$. Entonces se calcula $\hat{\boldsymbol{\eta}}_0$ y la derivada $\left(\frac{d\boldsymbol{\eta}}{d\boldsymbol{\mu}}\right)_0$. Estos dos elementos sólo dependen de la función de enlace. Finalmente, se calcula $\boldsymbol{V}_o$, es decir el valor de la varianza (2.9) para $\hat{\boldsymbol{\mu}}_o$, que únicamente depende de los valores ajustados $\hat{\boldsymbol{\mu}}$ en cada paso.

Las predicciones para nuevas observaciones también son independientes de la matriz $\boldsymbol{X}_w$ que se haya elegido. Dado un nuevo individuo Ω_{n+1}, dadas $\boldsymbol{\delta}_{n+1}$, las distancias al cuadrado con los *n* individuos previamente conocidos, se tiene que la predicción $\hat{\eta}_{n+1}$ se calcula con la expresión (2.39), con los valores del último paso del proceso de mínimos cuadrados ponderados iterativos. Y finalmente, se calcular $\hat{\mu}_{n+1} = g^{-1}(\hat{\eta}_{n+1})$.

El MLGBD incluye al MLG como caso particular: si se parte de una matriz $\boldsymbol{X}_w$ $\boldsymbol{w}$-centrada de dimensión $n \times r$ de r predictores continuos correspondientes a n individuos y se define $\boldsymbol{\Delta}$ como la matriz de distancias al cuadrado Euclídeas entre las filas de $\boldsymbol{X}_w$, entonces $\boldsymbol{X}_w$ es una configuración Euclídea de $\boldsymbol{\Delta}$ y, en este caso, la matriz de predicción, la respuesta y las predicciones coinciden con los respectivos valores que se obtienen con mínimos cuadrados ponderados iterativos en el MLG.

2.2.5 Software estadístico. La librería *dbstats* de *R*.

Para realizar los cálculos computacionales en los modelos de regresión basados en distancias se encuentra disponible la librería de *R* creada con el nombre de *dbstats* (ver Boj *et al.*,

2014a), y que puede descargarse en http://CRAN.R-project,org/package=dbstats desde el *Comprehensive R Archive Network.*

En la librería *dbstats* se incluye la función *dblm*, que ajusta MLBD, y la función *dbglm*, que ajusta MLGBD.

La función *dbglm* se explica en este apartado con más detalle que la función *dblm*, ya que ha sido la que se ha utilizado en las distintas aplicaciones prácticas del trabajo.

A nivel de notación, en la librería se considera que *y* es la variable respuesta del modelo (variable dependiente), que debe ser un objeto de clase vector numérico, factor, matriz o *data.frame*, siempre teniendo en cuenta que en los modelos de regresión implementados la variable respuesta es univariante. Las variables explicativas del modelo se indican como *Z* o $\boldsymbol{z}$ mientras que *X* o *x* denotan la configuración Euclídea $\boldsymbol{X}_w$.

Cabe indicar que en la ejecución de la función *dbglm* se incluyen distintas opciones para la introducción de las distancias entre individuos y, por tanto, la sintaxis es distinta para cada caso.

En primer lugar, se puede utilizar como argumento principal de la función un objeto de clase *formula* de la forma $y \sim Z$, que indica la relación entre la variable respuesta y las variables explicativas del modelo. Este argumento también se utiliza en la función *glm* de la librería *stats* de *R*, donde se ajusta el MLG ordinario.

Si se indica el argumento *formula* se introduce el argumento *data* que incluye las variables en el modelo, ya sean la variable respuesta y las variables explicativas (o bien las variables observadas, *Z*, o bien una configuración Euclídea *X*). En caso contrario se requiere el argumento *y*, que es la variable respuesta (dependiente) y debe ser un vector numérico, un factor, una matriz o un *data.frame.*

Si no se utiliza el argumento principal *formula* se consideran tres opciones distintas:

- El argumento *distance* es un objeto de clase *dist* o *dissimilarity*.

  - En el caso de clase *dist* una posibilidad es hacer uso de la función con este mismo nombre de la librería *stats*, que permite calcular matrices de distancias usando una métrica de distancia específica para obtener las distancias entre las filas de una matriz de datos.
  - En el caso de *dissimilarity* una posibilidad es hacer uso de la función *daisy* de la librería *cluster*, que calcula disimilaridades entre las observaciones del conjunto de datos.

- El argumento *D2* es un objeto de clase *D2*, que contiene la matriz de distancias al cuadrado.
- El argumento *G* es un objeto de clase *Gram*. En este caso se utiliza la matriz de productos internos centrados $G = X \cdot X^T$.

Cuando las distancias se calculan a partir de las variables explicativas observadas se pueden aplicar distintas métricas. En concreto se destaca la métrica Euclídea y el índice de similaridad de Gower que están descritas en apartados anteriores de este capítulo y que son las dos métricas que se utilizan en las aplicaciones de este trabajo. Cabe notar que, para indicar en la función *dbglm* que se aplique alguna de estas dos métricas directamente en el caso de *formula*, es posible utilizar el parámetro *metric*, que será igual a "*euclidean*" (que es la opción por defecto) o "*gower*", respectivamente, para estos dos casos.

En la aplicación del MLGBD es necesario indicar la distribución del error y la función de enlace asumidas en el modelo. Esta información se recoge en el argumento *family*, que es una cadena de caracteres donde se indica la distribución elegida y, opcionalmente, la función de enlace (por defecto se utiliza el enlace canónico) y que es el mismo argumento que se utiliza en la función *glm* de *stats*. Algunos de los valores que puede tomar este argumento son:

- "gaussian" para distribución Normal (por defecto)
- "binomial" para distribución Binomial
- "Gamma" para distribución Gamma
- "inverse.gaussian" para distribución Inversa Gaussiana

- “poisson” para distribución Poisson
- “quasipoisson” para distribución Poisson sobredispersa

Las funciones de enlace viene explicitada en el argumento *link* de la función *family*, y algunos de los valores que puede tomar son:

- “identity” para enlace identidad
- “inverse” para enlace inverso
- “log” para enlace logarítmico
- “logit” para enlace logit

Se pueden asignar también los pesos de cada uno de los individuos que se aplicarán en el proceso de ajuste del modelo con el argumento *weights*, que por defecto serán igual a 1.

El argumento *method* permite elegir cómo se decide el rango efectivo del modelo resultante. La función *dbglm* permite aplicar cinco criterios diferentes que se definen en el argumento *method*: “GCV”, “AIC”, “BIC”, “*rel.gvar*” y “*eff.rank*” al igual que se explica para la función *dblm*. La diferencia con la función *dblm* es que para la función *dbglm* no está implementado el criterio OCV debido al costoso tiempo computacional.

Cuando se elige en el argumento *method* algunas de las opciones siguientes: “AIC”, “BIC” o “GVC”, entonces es necesario especificar qué procedimiento de optimización se desea utilizar para minimizar el criterio elegido a través del argumento *full.search* (argumento lógico). Se puede asignar el valor TRUE y en este caso buscará el mejor valor global después de evaluar el criterio para todos los posibles rangos, o bien el valor FALSE, en cuyo caso pasará por aplicar la función *optimize.*

Para obtener las predicciones de nuevos individuos se ejecuta el comando *predict* de forma similar que para la función *glm*, el cual tiene los siguientes argumentos:

- El argumento *object* es un objeto de clase *dbglm*, es decir, el resultado de aplicar la función *dbglm*.

- El argumento *newdata* es un *data.frame* o matriz relacionado con el argumento *type.var*, de manera que:
    - Si *type.var* es igual a "*Z*" indica que *newdata* contiene los valores de las variables explicativas.
    - Si *type.var* es igual a "D2" indica que *newdata* contiene las distancias al cuadrado entre los k nuevos individuos y los n individuos originales.
    - Si *type.var* es igual a "G" indica que *newdata* contiene los productos internos.
- El argumento *type.pred* es el tipo de predicción, y puede ser igual a "*link*" cuando se trata de los predictores lineales (que es el valor que toma por defecto), o igual a "*response*" cuando se trata de la variable respuesta.

2.3 Metodología *bootstrap* aplicada a los modelos de regresión

Los métodos *bootstrap* (término implementado por Efron, 1979) implican la estimación de un modelo muchas veces usando datos simulados, de manera que los valores que se obtienen a partir de los datos simulados permiten hacer inferencias sobre los datos reales.

Los métodos *bootstrap* admiten aplicaciones en distintos contextos, como contrastes de hipótesis, modelos de regresión y estimación de errores estándar e intervalos de confianza. En Efron y Tibshirani (1998) y por ejemplo MacKinnon (2006) puede consultarse un tratamiento detallado de esta metodología y de su aplicabilidad.

Este apartado se centra en algunas metodologías *bootstrap* que resultan adecuadas en los modelos de regresión, en concreto:

- *Bootstrapping residuals*: dónde cada muestra *bootstrap* de vectores respuesta de n elementos se obtiene a partir de remuestrear n residuos.
- *Bootstrapping pairs*: dónde cada muestra *bootstrap* consiste en n pares de respuesta-predictor a partir de los datos originales.

La diferencia entre ambos métodos es que en el primer caso las variables latentes (o los predictores) se consideran fijos. Se asume que el modelo de regresión básico es correcto y que los residuos se pueden considerar iguales. Esta metodología puede conducir a resultados erróneos cuando, por ejemplo, los residuos tengan diferentes varianzas. El *bootstrapping pairs* es menos sensible a hipótesis erróneas sobre el modelo y es más robusto que el *bootstrapping residuals*, aunque en algunos contextos sólo puede implementarse este último método, por ejemplo, cuando existe dependencia entre algunas observaciones y los estimadores de los parámetros.

Bootstrapping residuals

Para definir los residuos más adecuados para este *bootstrap* es importante tener en cuenta que:

- El remuestreo se basa en la hipótesis de que los errores son independientes e idénticamente distribuidos. Además, no es necesario asumir que los errores se distribuyen según una distribución Normal o cualquier otra distribución conocida.
- Es indiferente el remuestreo de los residuos o de los residuos multiplicados por una constante, siempre y cuando se tenga este hecho en cuenta en el proceso de generación de datos.

Para ilustrar el procedimiento a seguir en el *bootstrapping pairs* se supone un modelo de regresión lineal. El primer paso es calcular los estimadores de los parámetros $\hat{\boldsymbol{\beta}}$ del modelo y de los residuos, $\hat{r}_i$, $i = 1,...,n$, que deben ser ajustados por los grados de libertad:

$$r_i^{'} = \sqrt{\frac{n}{n-p}}\hat{r}_i, \tag{2.40}$$

donde n el número de observaciones y p el número de parámetros estimados en el modelo.

El proceso de generación de datos del *bootstrappimg residuals*, utilizando los residuos ajustados, genera una observación de la muestra *bootstrap* mediante la ecuación:

$$y_i^* = \boldsymbol{x}_i \cdot \hat{\boldsymbol{\beta}}_i + r_i^*, \tag{2.41}$$

donde r_i^* siguen la distribución empírica de $r_i^{'}$. Se dice entonces que los errores *bootstrap* r_i^* se remuestrean a partir de $r_i^{'}$. Es decir, se han extraído de la función de distribución empírica de $r_i^{'}$, que asigna probabilidad $1/n$ a cada $r_i^{'}$. Así, cada uno de los términos del error *bootstrap* puede tomar n valores posibles, es decir, los valores de $r_i^{'}$, cuya probabilidad para cada uno de ellos es $1/n$.

Este mismo procedimiento puede aplicarse para un MLG, como se hace en una aplicación práctica incluida en el Capítulo 5 de este trabajo, donde se detalla todo el proceso de cálculo.

Bootstrapping pairs

Este método fue propuesto por Freedman (1981) y se basa en remuestrear los datos en lugar de los residuos.

Si se supone un modelo de regresión lineal, se remuestrea a partir de la matriz $[\boldsymbol{y}, \boldsymbol{X}]$ con una fila típica $[y_i, \boldsymbol{x}_i]$. Cada observación de la muestra *bootstrap* es $[y_i^*, \boldsymbol{x}_i^*]$, una fila elegida aleatoriamente de $[\boldsymbol{y}, \boldsymbol{X}]$. Este método se denomina *bootstrapping pairs* porque la variable respuesta y_i^* y las variables explicativas $\boldsymbol{x}_i^*$ siempre se seleccionan por pares.

Este método no impone ninguna condición sobre $\boldsymbol{X}$; de hecho, cada muestra *bootstrap* tiene una matriz $\boldsymbol{X}^*$ distinta. Se asume implícitamente que cada observación $[y_i, \boldsymbol{x}_i]$ es un valor aleatorio independiente de una distribución multivariante.

El *bootstrapping pair*s es muy fácil de implementar y se puede aplicar a una amplia gama de modelos, aunque tiene dos inconvenientes:

- El primero es que en el proceso de generación de datos no se impone ninguna restricción sobre los parámetros $\boldsymbol{\beta}$.

- El segundo es que, en comparación con el *bootstrapping pairs*, generalmente no proporciona resultados muy exactos, básicamente porque no impone ninguna condición sobre $\boldsymbol{X}$.

En este trabajo se utiliza el *bootsrapping pairs* aplicado a los modelos de regresión para el MLGBD en una aplicación práctica incluida en el Capítulo 5.

Capítulo 3

Coeficientes de influencia para el modelo lineal generalizado basado en distancias

En este capítulo se definen coeficientes de influencia en el MLGBD con el objetivo de evaluar la importancia relativa de los predictores en el modelo y calibrar su influencia en la respuesta. En segundo lugar, se construyen intervalos de confianza adecuados con la metodología *bootstrapping pairs* para contrastar su significación.

La definición de los coeficientes de influencia junto con el cálculo de errores estándar y *p*-valores con *bootstrap* se ha presentado en distintos congresos nacionales e internacionales:

- En la *5th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on COMPUTING & STATISTICS*, celebrada en Oviedo del 1 al 3 de Diciembre de 2012, se ha presentado "Relative predictor importance in distance-based generalized linear models", de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa.

- En la *15th Applied Stochastic Models and Data Analysis International Conference*, celebrada en Mataró (Barcelona) del 25 al 28 de junio de 2013, se ha presentado "Assessing the importance of risk factors in distance-based generalized linear models", de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa.

- En el *XXIV Congreso Nacional de Estadística e Investigación Operativa*, celebrado en Castellón del 10 al 13 de setiembre de 2013, se ha presentado "Evaluación de la importancia relativa de los predictores observados en el modelo lineal generalizado basado en distancias", de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa.

Posteriormente, se han presentado dos trabajos donde se incluyen los intervalos de confianza *bootstrap* y el test de Wald:

- En la *3rd Stochastic Modeling Techniques and Data Analysis International Conference*, celebrada en Lisboa del 11 al 14 de junio de 2014, se ha presentado "Wald test and distance-based generalized linear models. Actuarial application", de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa.

- En el *International Workshop on Proximity Data, Multivariate Analysis, and Classification*, celebrado en Granada del 9 al 10 de octubre de 2014, se ha presentado "Bootstrap confidence intervals and the Wald test", de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa.

Por último, cabe resaltar dos artículos publicados con contenidos que se incluyen en este capítulo:

- "Assessing the importance of risk factors in distance-based generalized linear models" de Boj, E.; Esteve, A.; Fortiana, J. y T. Costa, en la revista *Methodology and Computing in Applied Probability*, en 2015, que se puede consultar en http://link.springer.com/article/10.1007/s11009-014-9415-6.

- "Wald test and distance-based generalized linear models. Actuarial application" de Boj, E. y T. Costa, en la revista *Global Journal of Pure and Applied Mathematics*, Volumen 11, Número 1, en 2015.

Este capítulo se estructura en tres apartados:

El primer apartado, 3.1., está dedicado a la definición de los coeficientes de influencia. Se detalla el procedimiento de cálculo para el caso de predictores categóricos o binarios y para el caso de predictores cuantitativos.

En el segundo apartado, 3.2., en primer lugar se propone la metodología *bootstraping pairs* para generar *B* muestras de los coeficientes de influencia y calcular, a partir de ellas, el error estándar. Después se construyen intervalos de confianza para los coeficientes de influencia, que pueden obtenerse de dos maneras distintas: basándose en la distribución Normal estandarizada o basándose en la distribución *bootstrap* del estadístico del test de Wald.

Finalmente, en el tercer apartado, 3.3, se utilizan unos datos de frecuencia de siniestralidad en el seguro de automóviles, a los que se aplica el MLGBD. Se calculan los once coeficientes de influencia para los tres predictores del modelo, de los cuales dos son cuantitativos y uno es categórico nominal. Se aplica a los datos la metodología *bootstrapping pairs* generando 1000 muestras a partir de las cuáles se construyen los intervalos de confianza correspondientes para contrastar su significación.

3.1 Definición

Se asume un conjunto de n individuos con un vector de pesos asociado $\boldsymbol{w}$ para los que se tiene el vector de respuesta $\boldsymbol{y}$ y un conjunto $\boldsymbol{F}$ de p factores de riesgo o predictores.

En un modelo lineal ordinario se cumple la siguiente relación entre la respuesta y los predictores:

$$\hat{\boldsymbol{y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{F}_1 + \hat{\beta}_2 \boldsymbol{F}_2 + \cdots + \hat{\beta}_p \boldsymbol{F}_p .$$

En un MLG ordinario se tiene la siguiente relación entre la respuesta y el predictor lineal:

$$\hat{\boldsymbol{y}} = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{F}_1 + \hat{\beta}_2 \boldsymbol{F}_2 + \cdots + \hat{\beta}_p \boldsymbol{F}_p\right).$$

Para medir la influencia de cada predictor $\boldsymbol{F}_j$ en un modelo lineal se pueden usar los coeficientes estimados $\hat{\beta}_j$ para $j = 1, \ldots, p$ y en un MLG se pueden utilizar los coeficientes estimados $\hat{\beta}_j$ para $j = 1, \ldots, p$ del predictor lineal.

Pero en la predicción basada en distancias hay que tener en cuenta la métrica, que actúa como intermediario entre cada predictor observable $\boldsymbol{F}_j$ con las respuestas. La relación indirecta entre el predictor y la respuesta impide interpretar los coeficientes del predictor lineal como se hace en los modelos lineales ordinarios o en los MLG.

En este sentido, en Boj *et al.* (2007), ya se ha propuesto una aproximación parcial a este problema para el MLBD, con la definición de una versión del test estadístico *F* para seleccionar variables explicativas, con *p*-valores haciendo uso de *bootstrap*.

En este apartado se definen y estudian coeficientes de influencia locales de los p predictores $\boldsymbol{F}$ para el MLGBD. Estos coeficientes también son válidos para el MLBD. La idea que yace bajo el concepto de influencia es imitar los coeficientes $\hat{\beta}_j$ del modelo lineal y del MLG clásicos.

Sean $\boldsymbol{f}^0 = \left(f_1^0, f_2^0, \ldots, f_p^0 \right)$ los p valores de los predictores de un individuo virtual o de referencia, que se toma como origen o referencia. Las influencias $\boldsymbol{F}_j$ que se quieren cuantificar dependerán de este individuo de referencia. Así, por ejemplo, $\boldsymbol{f}^0$ puede estar formado por la media o la mediana en coordenadas numéricas y por la moda en coordenadas binarias o cualitativas. Los valores de los coeficientes de influencia tienen una validez local, es decir, son válidos entorno a un valor dado en el espacio predictor, ya que dependen del individuo virtual que se elige.

A continuación se definen coeficientes de influencia diferenciando el caso en que se trate de predictores categóricos o binarios y el caso en que se trate de predictores cuantitativos.

3.1.1 Coeficientes de influencia para predictores categóricos (o binarios)

Para predictores categóricos (o binarios) se definen los coeficientes de influencia $\hat{\beta}_j$ para $j = 1,\ldots, p$ como el incremento en el predictor lineal estimado $\hat{\boldsymbol{\eta}}$ cuando el valor del predictor j-ésimo de $\boldsymbol{f}^0$ cambia a otro nivel.

Esto se expresa a partir de la notación:

$$\beta_j = \Delta j\, \hat{\boldsymbol{\eta}}\big|_{\boldsymbol{f}^0}, \quad j = 1,\ldots, p. \tag{3.1}$$

Los pasos a seguir para calcular los coeficientes de influencia para predictores categóricos (o binarios) son:

- Se define $\boldsymbol{f}^0 = (0,0,....,0)$, los valores de los predictores del individuo de referencia, por ejemplo para predictores binarios, y se calcula su predicción, $\hat{y}_{n+1}^{(0)}$.

- Se cambia la coordenada j-ésima del predictor a 1, $\boldsymbol{f}^0 = \left(0,\ldots,0,\underset{j-\acute{e}simo}{1},0,\ldots,0\right)$ y se calcula su predicción, $\hat{y}_{n+1}^{(1)}$.

- Con las nuevas predicciones $\hat{y}_{n+1}$ para los dos individuos virtuales (antes y después del cambio), $\hat{y}_{n+1}^{(0)}$ y $\hat{y}_{n+1}^{(1)}$ se calcula el incremento, $\hat{\beta}_j = \hat{y}_{n+1}^{(1)} - \hat{y}_{n+1}^{(0)}$.

Si en lugar de una predictor binario es un predictor categórico, cuando el número de niveles es superior a 2 se puede calcular el incremento desde el nivel básico hasta los otros haciendo la descomposición en variables *dummy*.

3.1.2 Coeficientes de influencia para predictores cuantitativos

Para predictores cuantitativos se definen los coeficientes de influencia $\hat{\beta}_j$ para $j = 1,\ldots, p$ como:

$$\beta_j = \left.\frac{\partial \hat{\eta}}{\partial F_j}\right|_{\boldsymbol{f}^0}, \quad j = 1, \ldots, p, \tag{3.2}$$

la velocidad a la que el predictor lineal estimado $\hat{\boldsymbol{\eta}}$ se mueve cuando $\boldsymbol{f}^0$ se mueve a lo largo de la curva:

$$\boldsymbol{f}^0 + t \cdot s_j \left(0, \ldots, 0, \underset{j-th}{1}, 0, \ldots, 0\right), \quad t \in (-\varepsilon, +\varepsilon), \tag{3.3}$$

donde s_j es la desviación estándar del predictor cuantitativo *j*-ésimo.

Los pasos a seguir para calcular los coeficientes de influencia para cada predictor continuo *j*, con $j = 1, \ldots, p$, son:

- Se definen $\boldsymbol{f}^0 = \left(\overline{f}_1, \overline{f}_2, \ldots, \overline{f}_p\right)$ los valores de los predictores del individuo de referencia.
- Se define $t \in (-\varepsilon, +\varepsilon)$ (T valores discretos).
- Se calcula la desviación estándar del predictor j, s_j.
- Se crean T individuos virtuales con $\boldsymbol{f}^0 + t \cdot s_j \left(0, \ldots, 0, \underset{j-ésimo}{1}, 0, \ldots, 0\right)$, $t \in (-\varepsilon, +\varepsilon)$.
- Se calcula el vector δ_{n+1} para los T nuevos individuos virtuales.
- Se calculan las predicciones $\hat{y}_{n+1}$ para los T nuevos individuos virtuales.
- Se calcula la velocidad del cambio de las predicciones $\hat{y}_{n+1}$. Para este cálculo se usa la función *smooth.spline* de *R* para ajustar un *spline* cúbico suavizado. Entonces, para estimar $\hat{\beta}_j$ se utiliza el comando *predict* para obtener la primera derivada del *spline* cuando t vale cero.

Este cálculo está inspirado en los *biplots* no lineales de Gower y Harding (1988).

3.2 Intervalos de confianza y *bootstrapping pairs*

Se propone usar la metodología *bootstrapping pairs* (ver, por ejemplo, Efron y Tibshirari, 1998 y Davidson y Hinkley, 1997). Cada muestra consiste en *n* pares de respuesta-predictor a partir de los datos originales. De esta manera, se pueden generar *B* muestras *bootstrap* a partir de las cuales se puede estimar la matriz *Beta* de coeficientes:

$$Beta = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_1^1 & \cdots & \hat{\beta}_1^B \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\beta}_p & \hat{\beta}_p^1 & \cdots & \hat{\beta}_p^B \end{bmatrix}_{p\times(B+1)},$$

donde en la primera columna se tienen los coeficientes estimados de la muestra original, $\hat{\beta}_j$ para $j = 1,\ldots, p$ y en las siguientes *B* columnas se tienen los coeficientes estimados para las *B* muestras *bootstrap* generadas, $\hat{\beta}_j^b$ para $j = 1,\ldots, p$ y para $b = 1,\ldots, B$.

Para estimar los errores estándar de los coeficientes de influencia se puede usar la fórmula habitual:

$$std^*\left(\hat{\beta}_j\right) = \sqrt{\text{var}^*\left(\hat{\beta}_j\right)} = \sqrt{\sum_{b=1}^{B} \frac{\left(\hat{\beta}_j^b - \overline{\hat{\beta}}_j^*\right)^2}{B-1}}, \tag{3.4}$$

donde

$$\overline{\hat{\beta}}_j^* = \sum_{b=1}^{B} \frac{\hat{\beta}_j^b}{B}. \tag{3.5}$$

Para contrastar la hipótesis nula $H_0 : \beta_j = \beta_0$ para un valor numérico fijado β_0 se pueden construir intervalos de confianza de distintas maneras. MacKinnon (2006) propone que la manera más simple es calcular el error estándar y usarlo para construir intervalos de confianza

basados en la distribución Normal estandarizada. Así, un intervalo de confianza simple de *bootstrap* a un nivel de confianza $1-\alpha$ se puede construir aplicando la fórmula:

$$\left[\hat{\beta}_j - std^*\left(\hat{\beta}_j\right) z_{1-\frac{\alpha}{2}}, \hat{\beta}_j + std^*\left(\hat{\beta}_j\right) z_{1-\frac{\alpha}{2}}\right], \tag{3.6}$$

donde $z_{1-\frac{\alpha}{2}}$ denota el cuantil $1-\frac{\alpha}{2}$ de la distribución Normal estandarizada. Por ejemplo, si $\alpha = 0.05$, entonces $\frac{\alpha}{2} = 0.025$ y el cuantil 0.975 de la distribución normal es $z_{1-\frac{0.05}{2}} = 1.96$.

El intervalo de confianza simple de *bootstrap* (3.6) se puede modificar para que esté centrado en un estimador corregido por sesgo $\hat{\beta}_j$. Simplemente hay que reemplazar $\hat{\beta}_j$ en (3.6) por la expresión:

$$\breve{\beta}_j = \hat{\beta}_j - \left(\bar{\hat{\beta}}_j^* - \hat{\beta}_j\right) = 2\hat{\beta}_j - \bar{\hat{\beta}}_j^*. \tag{3.7}$$

No hay ninguna razón teórica para creer que el intervalo simple (3.6) funcionará mejor, o peor, que un intervalo similar basado puramente en la teoría asintótica. Sin embargo, puede usarse cuando no hay manera de calcular un error estándar analíticamente o cuando los errores asintóticos son poco fiables. Otra ventaja es que el número de muestras *bootstrap, B*, no tiene que ser muy elevado.

El método *bootstrapping pairs* no impone ninguna restricción sobre β_j en el proceso de generación de datos, como se explica en MacKinnon (2002, 2006, 2007). Entonces, si se están contrastando restricciones sobre β_j, a diferencia de la estimación de los errores estándar, o se están definiendo intervalos de confianza es necesario modificar el estadístico del test de modo que se esté contrastando algo que será cierto en el proceso de generación de datos. O, alternativamente, se puede modificar el esquema de remuestreo para que la hipótesis nula sea respetada en el proceso de generación de datos del *bootstrap*, como se explica en Boj *et al.* (2007) y Flachaire (1999, 2005).

En este sentido, se describe a continuación cómo adaptar el test de Wald al MLGBD. El test de Wald contrasta la hipótesis nula $H_0 : \beta_j = \beta_0$. El estadístico:

$$\tau_j = \frac{\hat{\beta}_j - \beta_0}{std\left(\hat{\beta}_j\right)} \tag{3.8}$$

sigue una distribución t asintóticamente, en el MLG ordinario.

En (3.8) $\hat{\beta}_j$ es el estimador no restringido del parámetro β_j que está siendo contrastado y $std\left(\hat{\beta}_j\right)$ es su error estándar.

Se propone estimar una distribución t^* del *bootstrsap* para el MLGBD que siga la hipótesis nula del test, $H_0 : \beta_j = \beta_0$. El procedimiento se basa en utilizar el estadístico *bootstrap* modificado:

$$\hat{\tau}_j^b = \frac{\hat{\beta}_j^b - \hat{\beta}_j}{std^*\left(\hat{\beta}_j\right)} \quad \text{para} \quad b = 1, \ldots, B, \tag{3.9}$$

donde $\hat{\beta}_j^b$ es el estimador de β_j a partir de la *b*-ésima muestra para $b = 1, \ldots, B$, donde *B* es el tamaño de la muestra, y el denominador $std^*\left(\hat{\beta}_j\right)$ es el error estándar de la distribución de β_j.

Como el estimador de β_j a partir de las muestra *bootstrap* debería ser, en términos medios, igual a $\hat{\beta}_j$, la hipótesis nula que se contrasta con $\hat{\tau}_j^b$ es cierta en el proceso de generación de datos del *bootstrapping pairs*. De esta forma, se puede comparar el estadístico de la muestra original:

$$\tau_j = \frac{\hat{\beta}_j - \beta_0}{std\left(\hat{\beta}_j\right)}$$

con la distribución t^* dada por (3.9) y calcular un *p*-valor a partir de:

$$\hat{p}^*\left(\hat{\tau}_j\right)=\frac{1}{B}\sum_{b=1}^{B} I\left(\left|\hat{\tau}_j^b\right|>\left|\hat{\tau}_j\right|\right), \tag{3.10}$$

o, alternativamente, a partir de:

$$\hat{p}^*\left(\hat{\tau}_j\right)=2\min\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\tau}_j^b\le\hat{\tau}_j\right),\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{\tau}_j^b>\hat{\tau}_j\right)\right). \tag{3.11}$$

Un intervalo de confianza que tiene mejores propiedades que el intervalo de confianza simple de *bootstrap* es el intervalo de confianza del percentil de *t.* Un intervalo de confianza del percentil de t a un nivel de confianza $1-\alpha$ se define como:

$$\left[\hat{\beta}_j-std^*\left(\hat{\beta}_j\right)t^*_{1-\frac{\alpha}{2}},\hat{\beta}_j-std^*\left(\hat{\beta}_j\right)t^*_{\frac{\alpha}{2}}\right], \tag{3.12}$$

donde t^*_δ es el percentil δ de la distribución t^* del estadístico definido en (3.9).

Por ejemplo, si $\alpha=0.05$ entonces $\frac{\alpha}{2}=0.025$, $1-\frac{\alpha}{2}=0.975$ y $t^*_{1-\frac{\alpha}{2}}$ es el cuantil 0.975 de la distribución t^* del *bootstrap* y $t^*_{\frac{\alpha}{2}}$ es su cuantil 0.025.

La distribución t^* dada por (3.9) sigue la hipótesis nula, de esta manera el intervalo de confianza del percentil de t (3.12) resulta útil para contrastar la hipótesis con un valor real β_0 fijado. Con este tipo de intervalos de confianza no es necesario repetir los cálculos si se quiere cambiar el valor de β_0, a diferencia de lo que sucede con los *p*-valores.

3.3 Aplicación práctica

En esta aplicación práctica se utilizan unos datos del seguro de automóviles a terceros de Suecia del año 1977 que se describen en Hallin e Ingenbleek (1983) y que se han usado en aplicaciones anteriores, como Boj *et al.* (2012).

Los datos para el factor *Zone* = 1 pueden encontrarse en Andrews y Herzberg (1985). Estos datos corresponden a las ciudades de Estocolmo, Göteborg y Malmo y se obtuvieron a partir de un comité de estudio de primas de riesgo en seguros de automóviles.

Los datos están incluidos en la librería *faraway* de *R* bajo el nombre de *motorins* y pueden descargarse electrónicamente desde http://www.statsci.org/data/general/motorins.html.

El número total de observaciones para la *Zone* = 1 es $n = 295$, que corresponden a diferentes grupos de riesgo no vacíos. Se analiza la frecuencia de siniestralidad, a partir del número de siniestros sufridos por los automóviles asegurados y del número de asegurados en años de póliza.

Siguiendo, por ejemplo, a Haberman y Renshaw (1996) y Brockman y Wright (1992) se asume una distribución del error de Poisson y la función de enlace logarítmica.

Los factores de riesgo son tres:

- *Distance* (distancia): kilómetros recorridos por año. Se codifica el factor *Distance* como numérico usando las marcas de clase, como en Boj *et al.* (2012):
    - "<1000 Km por año" : 750 kilómetros recorridos por años
    - "1000-15000 Km por año": 8000 kilómetros recorridos por año
    - "15000-20000 Km por año": 17500 kilómetros recorridos por año
    - "20000-25000 Km por año": 22500 kilómetros recorridos por año
    - ">25000 Km por año": 40000 kilómetros recorridos por año
- *Bonus*: nivel en la escala de Bonus, con valores numéricos desde 1 hasta 7
- *Make* (marca): marca del vehículo, con nueve categorías nominales

Los factores *Distance* y *Bonus* se tratan como variables numéricas y el factor *Make* como categórica nominal.

Los cálculos para aplicar el MLGBD se realizan usando la función *dbglm* de la librería *dbstats* de *R* (Boj *et al*., 2014a).

La similaridad se calcula con el índice de similaridad de Gower (2.31) teniendo en cuenta la variabilidad geométrica (*rel.gvar=1*), es decir, aplicando el modelo denominado *dbglm1* en el Apéndice A de Boj *et al.* (2012).

Las instrucciones para la lectura y preparación de datos con *R* y el ajuste del modelo *dbglm1* son:

```
R> library("dbstats")
R> require("faraway")
R> data("motorins")
R> Motor1 <- subset(motorins, Zone == 1)
R> Motor1$KmC <- rep(0,nrow(Motor1))
R> Motor1$KmC[Motor1$Kilometres == "1"] <- 750
R> Motor1$KmC[Motor1$Kilometres == "2"] <- 8000
R> Motor1$KmC[Motor1$Kilometres == "3"] <- 17500
R> Motor1$KmC[Motor1$Kilometres == "4"] <- 22500
R> Motor1$KmC[Motor1$Kilometres == "5"] <- 40000
R> Motor1$BonC <- as.numeric(Motor1$Bonus)

R> dbglm1 <- dbglm(Claims ~ KmC + BonC + factor(Make), offset =
(log(Motor1$Insured)), data = Motor1, family = poisson(link = "log"), metric = "gower",
method = "rel.gvar", rel.gvar = 1); dbglm1

Call:  dbglm(formula = Claims ~ KmC + BonC + factor(Make), data = Motor1,
    family = poisson(link = "log"), method = "rel.gvar", metric = "gower",
    rel.gvar = 1, offset = (log(Motor1$Insured)))
family: poisson
metric: gower

Degrees of Freedom: 294 Total (i.e. Null);  276 Residual
Null Deviance:      6978
```

```
Residual Deviance: 454.1

AIC:       1827.271
BIC:       1897.323
GCV:        1.7457
```

Para este modelo los grados de libertad totales son $n-1=294$ y los grados de libertad residuales son $k=276$. Esto significa que con el índice de Gower y explicando la variabilidad geométrica total del modelo se tiene un rango efectivo de 276 en el predictor lineal.

El objetivo es estimar once coeficientes de influencia: nueve para los nueve niveles desde *Make1* hasta *Make9* del factor *Make* aplicando la definición (3.1) y dos para los factores numéricos *Distance* y *Bonus* aplicando la definición (3.2).

En este sentido, se puede expresar el predictor lineal a partir de:

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= \beta_0 + \beta_1 \boldsymbol{F}_1 + \beta_2 \boldsymbol{F}_2 + \cdots + \beta_{10} \boldsymbol{F}_{10} + \boldsymbol{\varepsilon} = \\ &\beta_{Make1} + \beta_{Make2} \boldsymbol{F}_{Make2} + \beta_{Make3} \boldsymbol{F}_{Make3} + \beta_{Make4} \boldsymbol{F}_{Make4} + \\ &\beta_{Make5} \boldsymbol{F}_{Make5} + \beta_{Make6} \boldsymbol{F}_{Make6} + \beta_{Make7} \boldsymbol{F}_{Make7} + \\ &\beta_{Make8} \boldsymbol{F}_{Make8} + \beta_{Make9} \boldsymbol{F}_{Make9} + \beta_{Km} \boldsymbol{F}_{Km} + \beta_{Bon} \boldsymbol{F}_{Bon} + \boldsymbol{\varepsilon}\end{aligned}$$

Se define $\boldsymbol{f}^0 = \left(Make = 1, Km = \overline{Km}, Bon = \overline{Bon}\right) = \left(1, 9683.82, 5.58\right)$ como el individuo de referencia, siendo la clase *Make* = 1 la correspondiente al término independiente β_0.

Los resultados de los coeficientes de influencia estimados para este modelo, *dbglm1*, se recogen en la Tabla 3.1:

Tabla 3.1. Coeficientes de influencia estimados para MLGBD usando el índice de similaridad de Gower (*dbglm1*).

$\hat{\beta}_0 = \hat{\beta}_{Make1}$	-1.856603
$\hat{\beta}_1 = \hat{\beta}_{Make2}$	1.311531e-01
$\hat{\beta}_2 = \hat{\beta}_{Make3}$	-2.142091e-01
$\hat{\beta}_3 = \hat{\beta}_{Make4}$	-4.976677e-01
$\hat{\beta}_4 = \hat{\beta}_{Make5}$	1.238671e-01
$\hat{\beta}_5 = \hat{\beta}_{Make6}$	-3.880247e-01
$\hat{\beta}_6 = \hat{\beta}_{Make7}$	-1.303605e-01
$\hat{\beta}_7 = \hat{\beta}_{Make8}$	1.362958e-01
$\hat{\beta}_8 = \hat{\beta}_{Make9}$	-2.360768e-02
$\hat{\beta}_9 = \hat{\beta}_{Km}$	1.068948e-05
$\hat{\beta}_{10} = \hat{\beta}_{Bon}$	-3.732707e-02

En la siguiente Tabla 3.2 se muestran los coeficientes estimados para el modelo *dbglm4* del apéndice A de Boj *et al.* (2012), que se ha ajustado usando distancia Euclídea en el MLGBD:

Tabla 3.2. Coeficientes de influencia estimados para MLGBD usando distancia Euclídea (*dbglm4*).

$\hat{\beta}_0 = \hat{\beta}_{Make1}$	-1.640
$\hat{\beta}_1 = \hat{\beta}_{Make2}$	1.282e-01
$\hat{\beta}_2 = \hat{\beta}_{Make3}$	-2.140e-01
$\hat{\beta}_3 = \hat{\beta}_{Make4}$	-5.162e-01
$\hat{\beta}_4 = \hat{\beta}_{Make5}$	1.270e-01
$\hat{\beta}_5 = \hat{\beta}_{Make6}$	-3.976e-01
$\hat{\beta}_6 = \hat{\beta}_{Make7}$	-1.320e-01
$\hat{\beta}_7 = \hat{\beta}_{Make8}$	1.396e-01
$\hat{\beta}_8 = \hat{\beta}_{Make9}$	-3.079e-02
$\hat{\beta}_9 = \hat{\beta}_{Km}$	1.431e-05
$\hat{\beta}_{10} = \hat{\beta}_{Bon}$	-2.165e-01

La instrucción para ajustar el modelo *dbglm4* es la siguiente:

```
R> dbglm4 <- dbglm(Claims ~ KmC + BonC + factor(Make), offset =
(log(Motor1$Insured)), data = Motor1, family = poisson(link = "log"), metric =
"euclidean", method = "rel.gvar", rel.gvar = 1); dbglm4

Call:  dbglm(formula = Claims ~ KmC + BonC + factor(Make), data = Motor1,
    family = poisson(link = "log"), method = "rel.gvar", metric = "euclidean",
    rel.gvar = 1, offset = (log(Motor1$Insured)))

family: poisson
metric: euclidean

Degrees of Freedom: 294 Total (i.e. Null);  284 Residual
Null Deviance:      6978
Residual Deviance: 779.4
AIC:        2136.576
BIC:        2177.133
GCV:         2.8305
```

Los coeficientes estimados en el modelo *dbglm4* coinciden con el MLG clásico *glm1* del apéndice A de Boj *et al.* (2012). El modelo *glm1* se ha ajustado con la función *glm* de la librería *stats* de R.

La instrucción para ajustar el modelo *glm1* es la siguiente:

```
R> glm1 <- glm(Claims ~ KmC + BonC + factor(Make), offset = (log(Motor1$Insured)),
data = Motor1, family = poisson(link = "log")); glm1

Call:  glm(formula = Claims ~ KmC + BonC + factor(Make), family = poisson(link =
"log"), data = Motor1, offset = (log(Motor1$Insured)))

Coefficients:
```

```
  (Intercept)        KmC        BonC   factor(Make)2  factor(Make)3  factor(Make)4
-1.640e+00    1.431e-05   -2.165e-01    1.282e-01   -2.140e-01   -5.162e-01

  factor(Make)5  factor(Make)6  factor(Make)7  factor(Make)8  factor(Make)9
   1.270e-01    -3.976e-01    -1.320e-01     1.396e-01    -3.079e-02

Degrees of Freedom: 294 Total (i.e. Null);  284 Residual
Null Deviance:      6978
Residual Deviance: 779.4        AIC: 2137
```

En Boj *et al.* (2012) se indican los *p*-valores de los predictores. Los valores más altos (por tanto los menos significativos) son 0.02508, 0.10762 y 0.17618, para β_{Make7}, β_{Make8} y β_{Make9}, respectivamente. Como la mayoría de niveles de *MakeC* son altamente significativos se elige mantener la variable en el modelo.

En los Gráficos 3.1 y 3.2 se muestran las trayectorias de dos individuos virtuales representando dos predictores continuos observados en el MLGBD con el índice de similaridad de Gower, el modelo *dbglm1*. Estos gráficos también se puede consultar en Boj *et al.* (2015a).

En el Gráfico 3.1 se observa una evolución cercana a la lineal mientras que el Gráfico 3.2 refleja una no linealidad visible; si bien es cierto que si se hubiera decidido elegir un individuo virtual con una referencia de coordenadas mayor que 7 (en la escala horizontal) la pendiente habría tenido el signo opuesto. Para una modelo lineal las trayectorias serían líneas rectas.

Gráfico 3.1. Predictores lineales estimados para 201 valores de (3.3), una secuencia desde –1 hasta 1 con incrementos de 0.01 para el factor de riesgo *Distance* en el modelo *dbglm1*.

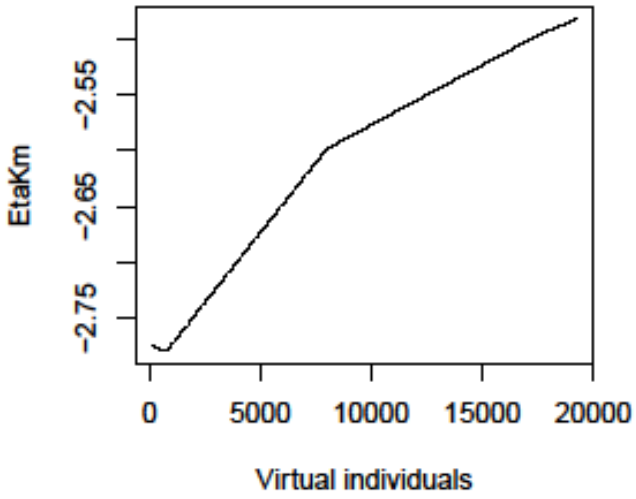
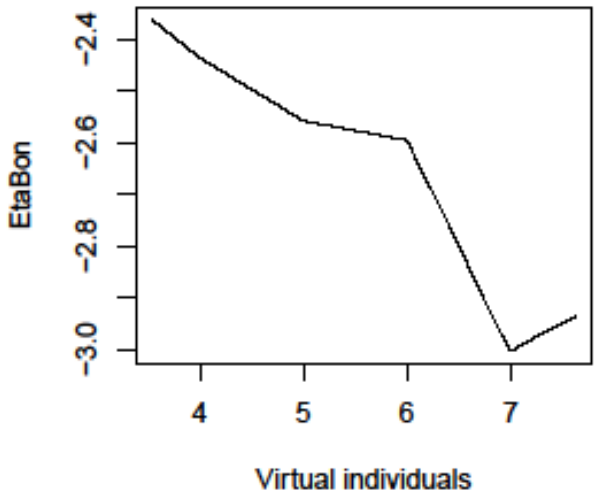


Gráfico 3.2. Predictores lineales estimados para 201 valores de (3.3), una secuencia desde –1 hasta 1 con incrementos de 0.01 para el factor de riesgo *Bonus* en el modelo *dbglm1*.



Se estiman los errores estándar en el modelo *dbglm1* con la ecuación (3.4) usando *bootstrapping pairs* con una muestra de tamaño $B = 1000$. Se ha dividido la muestra en dos subconjuntos de tamaño $B_1 = B_2 = 500$ y los resultados obtenidos se muestran en la siguiente Tabla 3.3:

Tabla 3.3. Errores estándar del *bootstrap* para dos muestras de tamaños $B_1 = B_2 = 500$, por lo que para la muestra total el tamaño es $B = 1000$, para MLGBD usando distancia de Gower (*dbglm1*).

	$B_1 = 500$	$B_2 = 500$	$B = 1000$
$\hat{\beta}_0 = \hat{\beta}_{Make1}$	1.659268e-03	1.655607e-03	1. 656611e-03
$\hat{\beta}_1 = \hat{\beta}_{Make2}$	3.207050e-03	3.206454e-03	3. 205148e-03
$\hat{\beta}_2 = \hat{\beta}_{Make3}$	6.467344e-03	6.463418e-03	6.462146e-03
$\hat{\beta}_3 = \hat{\beta}_{Make4}$	2.668724e-03	2.672244e-03	2.669149e-03
$\hat{\beta}_4 = \hat{\beta}_{Make5}$	3.360581e-03	3.347443e-03	3.352351e-03
$\hat{\beta}_5 = \hat{\beta}_{Make6}$	2.383127e-03	2.377103e-03	2.37893e-03
$\hat{\beta}_6 = \hat{\beta}_{Make7}$	5.395997e-03	5.412965e-03	5.401807e-03
$\hat{\beta}_7 = \hat{\beta}_{Make8}$	1.504232e-02	1.508169e-02	1.505449e-02
$\hat{\beta}_8 = \hat{\beta}_{Make9}$	1.317354e-03	1.311761e-03	1.313906e-03
$\hat{\beta}_9 = \hat{\beta}_{Km}$	9.231468e-08	9.298712e-08	9.260552e-08
$\hat{\beta}_{10} = \hat{\beta}_{Bon}$	1.791164e-03	1.792390e-03	1.790881e-03

Comparando las tres columnas de la Tabla 3.3 se observa que los resultados con una muestra de tamaño 500 son similares a los de una muestra de tamaño 1000 en los dos primeros decimales, de lo que se deduce que un tamaño de muestra de 500 ya resultaría suficiente.

Para el mismo modelo *dbglm1* se pueden construir intervalos simples de confianza del *bootstrap* usando los errores estándar calculados con el tamaño total de la muestra *B* y aplicando las expresiones (3.6) y (3.7).

Si se utiliza un nivel de confianza del 95% en la construcción de los intervalos para contrastar la hipótesis $H_0 : \beta_j = 0$, para $j = 0,...,10$ se deriva que todos los coeficientes son significativos. Los resultados obtenidos se presentan en la siguiente Tabla 3.4, que también se puede consultar en Boj *et al.* (2015a):

Tabla 3.4. Intervalos de confianza y errores estándar calculados para una muestra de tamaño $B = 1000$ y asumiendo $\alpha = 0.05$ para MLGBD usando distancia de Gower (*dbglm1*).

	Intervalo de confianza con $\alpha = 0.05$ usando (3.6)	Media de *bootstrap* (3.5) $\bar{\hat{\beta}}_j^*$	Intervalo de confianza con $\alpha = 0.05$ usando (3.7)
$\hat{\beta}_0 = \hat{\beta}_{Make1}$	[-1.860e+0,-1.853e+0]	-1.857e+0	[-1.860e+0,-1.853e+0]
$\hat{\beta}_1 = \hat{\beta}_{Make2}$	[1.249e-01,1.374e-01]	1.312e-01	[1.249e-01,1.374e-01]
$\hat{\beta}_2 = \hat{\beta}_{Make3}$	[-2.269e-01,-2.015e-01]	-2.150e-01	[-2.260e-01,-2.007e-01]
$\hat{\beta}_3 = \hat{\beta}_{Make4}$	[-5.029e-01,-4.924e-01]	-4.977e-01	[-5.029e-01,-4.924e-01]
$\hat{\beta}_4 = \hat{\beta}_{Make5}$	[1.173e-01,1.304e-01]	1.238e-01	[1.173e-01,1.305e-01]
$\hat{\beta}_5 = \hat{\beta}_{Make6}$	[-3.927e-01,-3.834e-01]	-3.881e-01	[-3.926e-01,-3.833e-01]
$\hat{\beta}_6 = \hat{\beta}_{Make7}$	[-1.410e-01,-1.198e-01]	-1.306e-01	[-1.407e-01,-1.195e-01]
$\hat{\beta}_7 = \hat{\beta}_{Make8}$	[1.068e-01,1.658e-01]	1.357e-01	[1.074e-01,1.664e-01]
$\hat{\beta}_8 = \hat{\beta}_{Make9}$	[-2.618e-02,-2.103e-02]	-2.363e-02	[-2.616e-02,-2.101e-02]
$\hat{\beta}_9 = \hat{\beta}_{Km}$	[1.051e-05,1.087e-05]	1.068e-05	[1.051e-05,1.087e-05]
$\hat{\beta}_{10} = \hat{\beta}_{Bon}$	[-4.084e-02,-3.382e-02]	-3.722e-02	[-4.095e-02,-3.392e-02]

Se completa el ejemplo con una medida cuantitativa de la significación de los predictores usando el estadístico modificado del test de Wald. Se construyen los correspondientes intervalos de confianza del percentil de *t* (3.12) al nivel de confianza del 95%. En la siguiente Tabla 3.5 se muestran los resultados de estos intervalos, junto con los valores de los coeficientes estimados y los cuantiles 97.5 y 2.5 de la distribución de t^* del *bootstrap* dada por los 1000 valores de (3.9). Estos resultados también se pueden consultar en Boj y Costa (2015a).

Tabla 3.5. Coeficientes estimados, cuantiles 97.5 y 2.5 de la distribución de t^* e intervalos de confianza del percentil de t asumiendo $\alpha = 0.05$ para MLGBD usando distancia de Gower (*dbglm1*).

	$\hat{\beta}_j$	$t^*_{1-\alpha/2}$ (97.5%)	$t^*_{\alpha/2}$ (2.5%)	Intervalo de confianza del percentil de t con $\alpha = 0.05$
$\hat{\beta}_0 = \hat{\beta}_{Make1}$	-1.856603	1.962086	-1.908919	[-1.860e+0, -1.853e+0]
$\hat{\beta}_1 = \hat{\beta}_{Make2}$	1.311531e-01	1.882888	-1.849325	[1.251e-01, 1.371e-01]
$\hat{\beta}_2 = \hat{\beta}_{Make3}$	-2.142091e-01	1.828144	-1.976043	[-2.260e-01, -2.014e-01]
$\hat{\beta}_3 = \hat{\beta}_{Make4}$	-4.976677e-01	1.930592	-2.144194	[-5.028e-01, -4.919e-01]
$\hat{\beta}_4 = \hat{\beta}_{Make5}$	1.238671e-01	1.81118	-1.968902	[1.178e-01, 1.305e-01]
$\hat{\beta}_5 = \hat{\beta}_{Make6}$	-3.880247e-01	1.916472	-2.03158	[-3.926e-01, -3.832e-01]
$\hat{\beta}_6 = \hat{\beta}_{Make7}$	-1.303605e-01	1.968073	-1.909681	[-1.410e-01, -1.200e-01]
$\hat{\beta}_7 = \hat{\beta}_{Make8}$	1.362958e-01	1.872734	-1.998162	[1.081e-01, 1.664e-01]
$\hat{\beta}_8 = \hat{\beta}_{Make9}$	-2.360768e-02	1.799623	-2.038099	[-2.597-02, -2.093e-02]
$\hat{\beta}_9 = \hat{\beta}_{Km}$	1.068948e-05	2.022189	-1.953856	[1.050e-05, 1.087e-05]
$\hat{\beta}_{10} = \hat{\beta}_{Bon}$	-3.732707e-02	2.113138	-1.685741	[-4.111e-02, -3.431e-02]

Como conclusión a partir de los resultados, se observa que no aparece el valor 0 en ninguno de los intervalos de confianza del percentil de t y esto significa que todos los coeficientes son significativos en el modelo *dbglm1*.

Finalmente, en los gráficos siguientes (del Gráfico 3.3 al 3.13) se muestra la distribución de t^* para los distintos coeficientes de influencia en el modelo *dbglm1*. Estos gráficos se pueden encontrar en Boj *et al.* (2015a).

Gráfico 3.3. Distribución *bootstrap* t^* de $\hat{\beta}_{Make1}$.

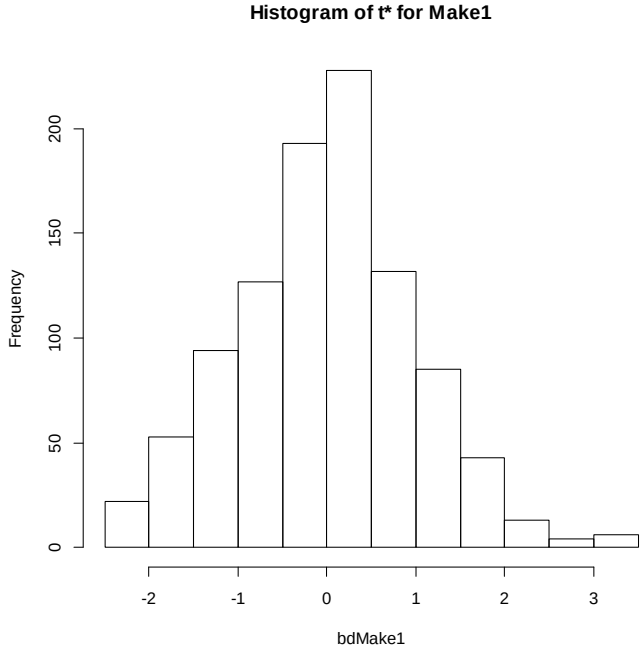


Gráfico 3.4. Distribución *bootstrap* t^* de $\hat{\beta}_{Make2}$.

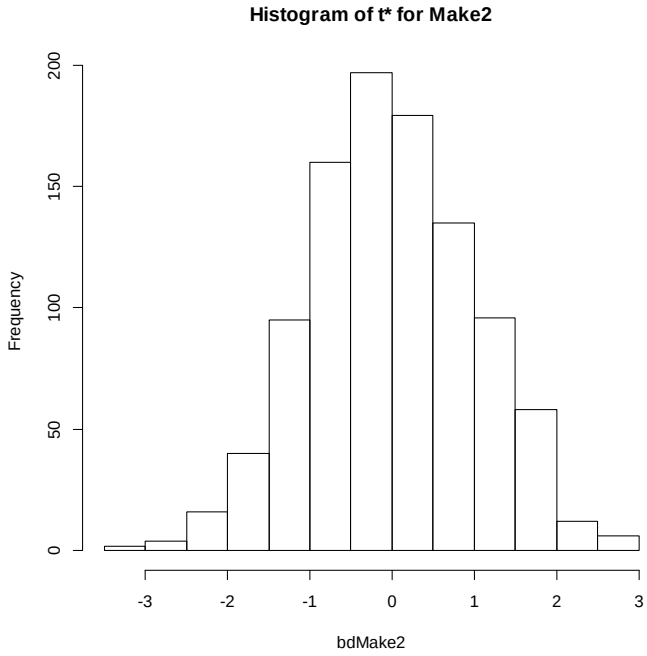


Gráfico 3.5. Distribución *bootstrap* t^* de $\hat{\beta}_{Make3}$.

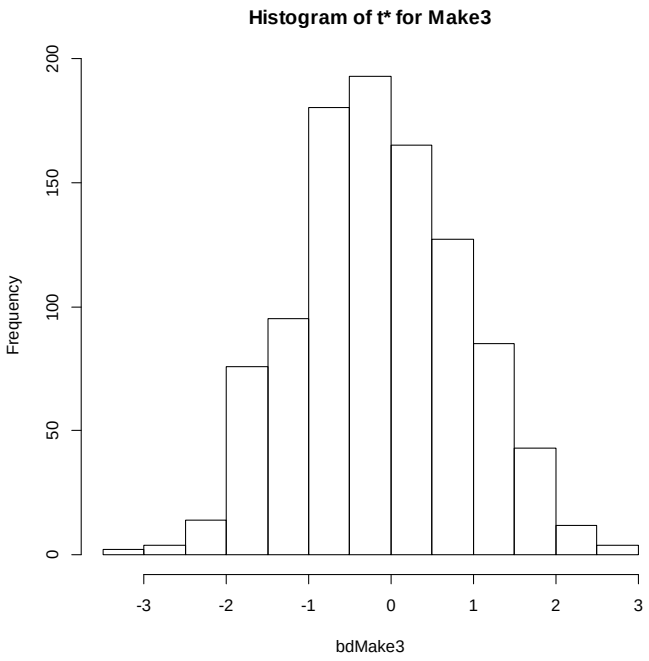


Gráfico 3.6. Distribución *bootstrap* t^* de $\hat{\beta}_{Make4}$.

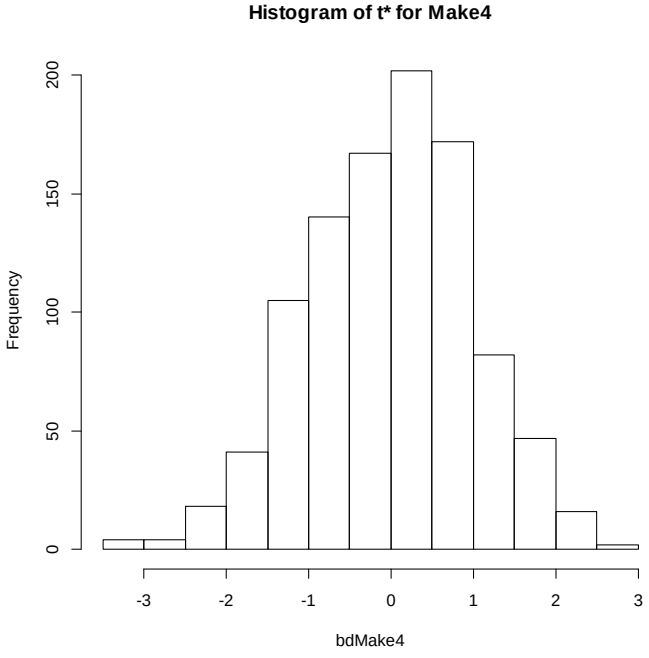


Gráfico 3.7. Distribución *bootstrap* t^* de $\hat{\beta}_{Make5}$.

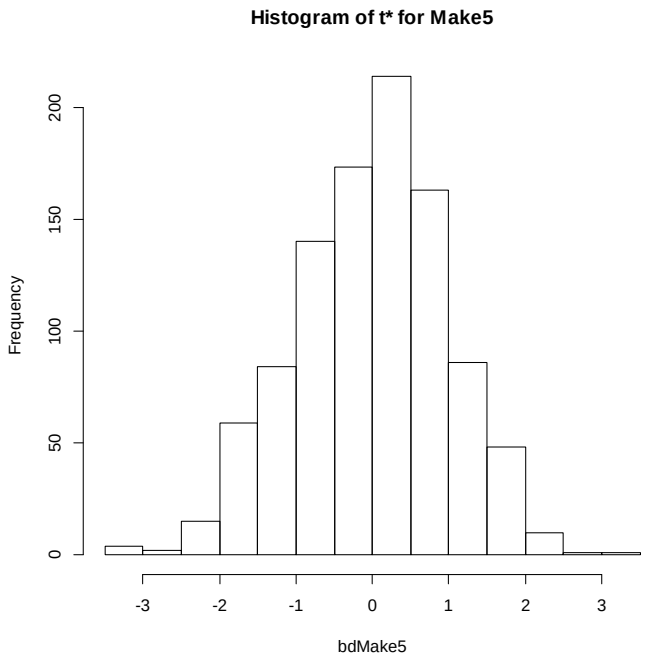


Gráfico 3.8. Distribución *bootstrap* t^* de $\hat{\beta}_{Make6}$.

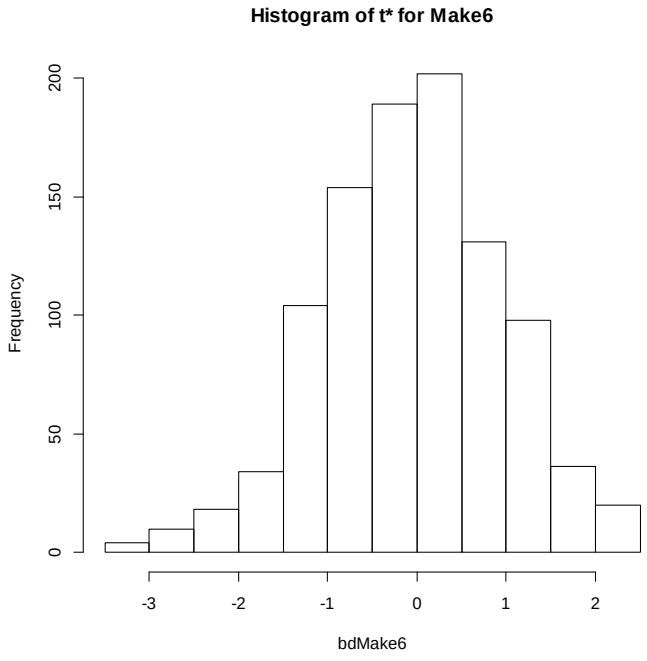


Gráfico 3.9. Distribución *bootstrap* t^* de $\hat{\beta}_{Make7}$.

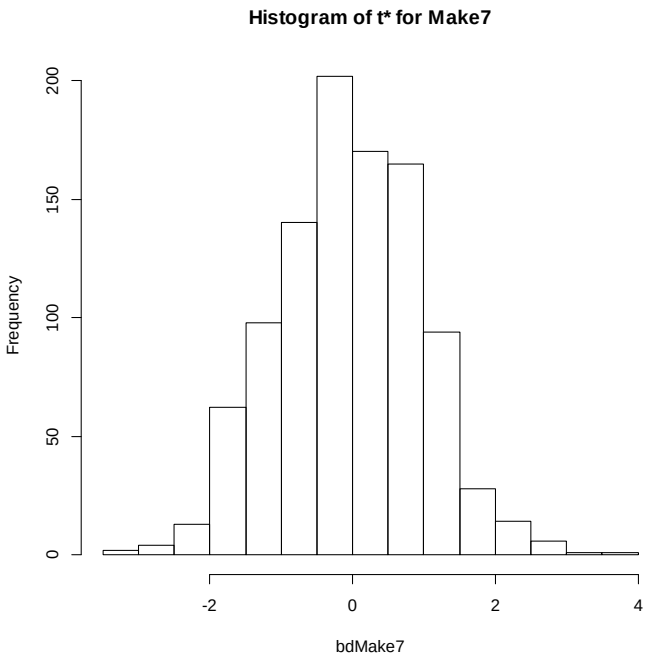


Gráfico 3.10. Distribución *bootstrap* t^* de $\hat{\beta}_{Make8}$.

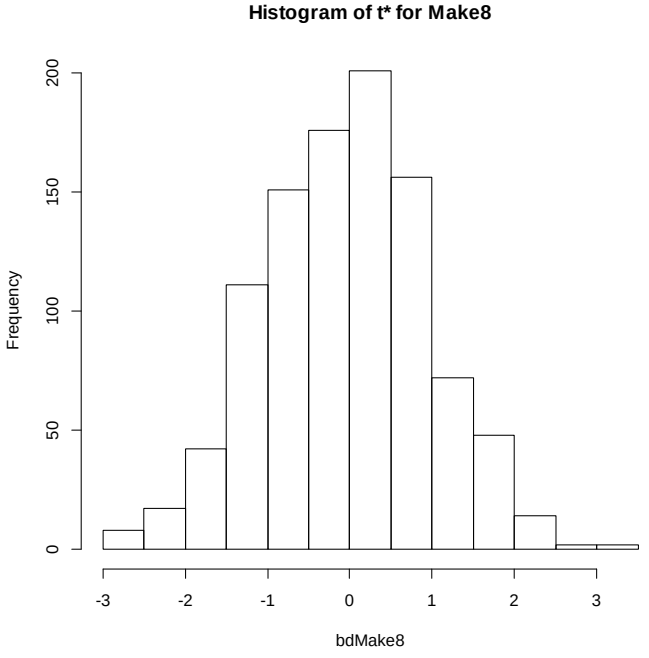


Gráfico 3.11. Distribución *bootstrap* t^* de $\hat{\beta}_{Make9}$.

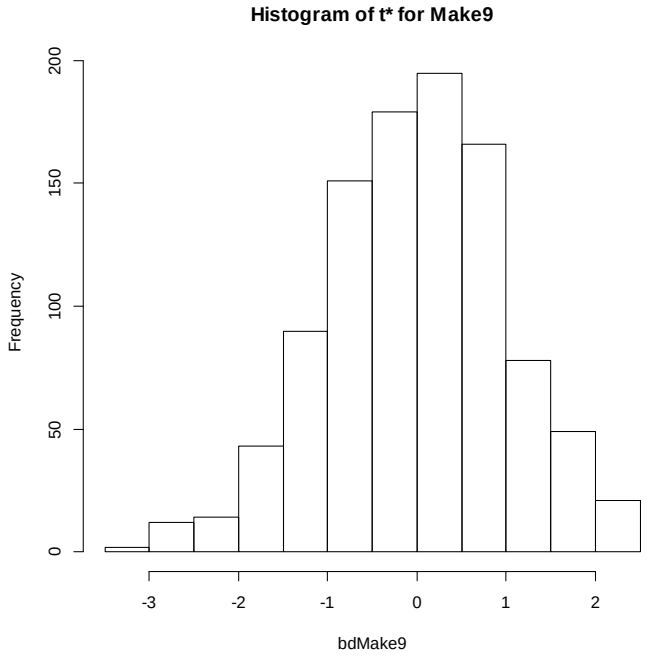


Gráfico 3.12. Distribución *bootstrap* t^* de $\hat{\beta}_{Km}$.

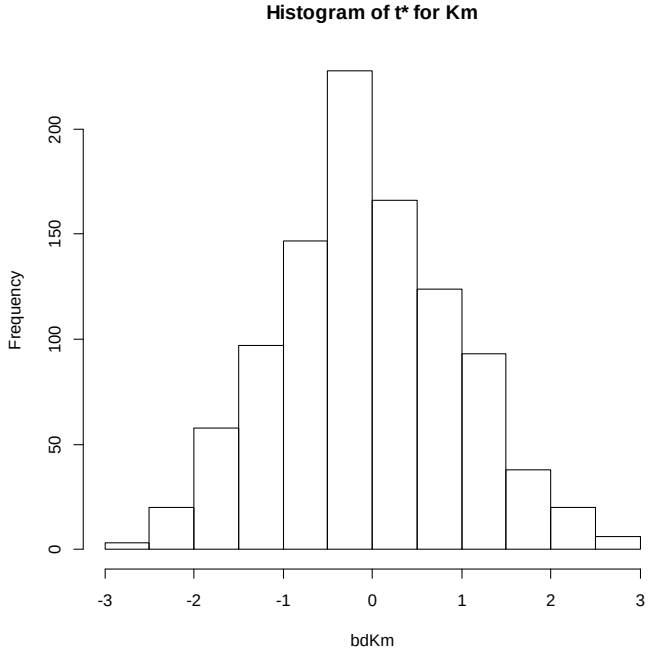
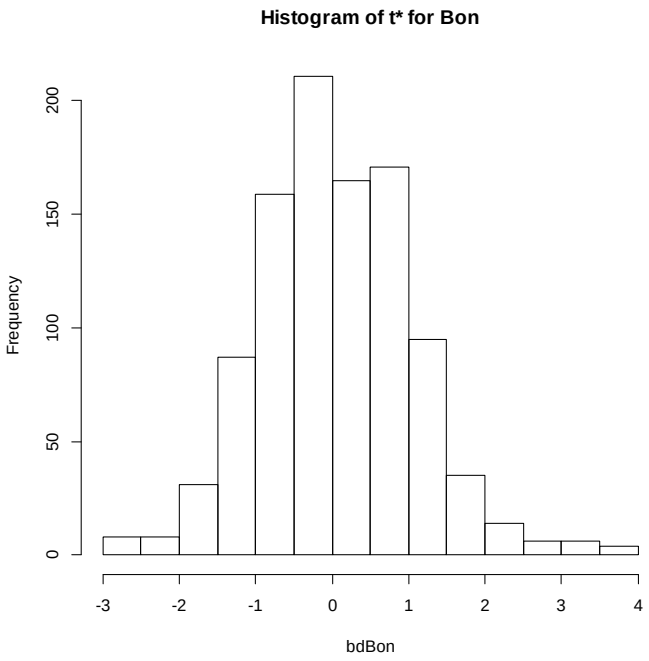


Gráfico 3.13. Distribución *bootstrap* t^* de $\hat{\beta}_{Bon}$.



Capítulo 4

Riesgo de crédito: cálculo de puntuaciones mediante regresión logística basada en distancias

En este capítulo se estudia la aplicación del modelo de regresión logística BD al problema del *credit scoring*. Para poder comparar distintas metodologías, se utilizan como criterios las probabilidades de mala clasificación de los individuos y los costes del error en la mala clasificación. Se proponen distintas alternativas para elegir el punto de corte a partir del cual se considera que un individuo es un mal riesgo de crédito.

Se han presentado en distintos congresos nacionales e internacionales el estudio de la aplicación del modelo de regresión logística BD en *credit scoring*:

- En el *4th Workshop on Risk Management and Insurance RISK2011*, celebrado en Sevilla del 20 al 21 de Octubre de 2011, se ha presentado "Aplicación de un modelo de regresión logística basado en distancias en el problema del *credit scoring*", de Boj, E.; Fortiana, J.; Esteve, A; Claramunt, M.M. y T. Costa.

- En la *Barcelona International Conference on Advances in Statistics*, celebrada en Barcelona del 18 al 22 de Junio de 2012, se ha presentado "ROC curves in distance-based credit risk models", de Boj, E.; Fortiana, J. y T. Costa.

- En las *XX Jornadas ASEPUMA y VIII Encuentro Internacional de Profesores Universitarios de Matemáticas para la Economía y la Empresa*, celebrada en Barcelona del 19 al 20 de Julio de 2012, se ha presentado "Metodologías de elección

del punto de corte para regresión logística no paramétrica y basada en distancias en el problema del *credit scoring*", de Costa, T.; Boj, E y J. Fortiana.

Se destaca también un artículo publicado con contenidos de este capítulo: "Bondad de ajuste y elección del punto de corte en regresión logística basada en distancias. Aplicación al problema de *credit scoring*" de Costa, T.; Boj, E. y J. Fortiana, en la revista *Anales del Instituto de Actuarios Españoles*, Tercera Epoca, 18, 19-40, en 2012.

Este capítulo se estructura en tres apartados.

En el primer apartado, 4.1., se describe el modelo de regresión logística BD, un caso de MLGBD, en su aplicación al cálculo de *scorings* para el problema del riesgo de crédito.

En el segundo apartado, 4.2., se presentan criterios para seleccionar un modelo en *credit scoring*. En primer lugar, se estiman las probabilidades de mala clasificación de los individuos, es decir, la probabilidad de mala clasificación de los buenos riesgos, de los malos riesgos y la probabilidad global. En segundo lugar, se consideran los costes del error derivados de una mala clasificación de los individuos. Se utilizan unos datos de riesgo de crédito de dos entidades financieras para comparar la regresión logística BD con otras metodologías empleadas en la literatura actuarial.

En el tercer apartado, 4.3., se proponen distintas maneras de elegir un punto de corte en regresión logística BD para clasificar a un individuo como mal riesgo de crédito. En primer lugar, se estudian el coeficiente Kolmogorov-Smirnov (K-S) y el índice de Gini, junto con la representación gráfica de la curva ROC. En segundo lugar, se analizan los criterios definidos en el apartado 4.2 cuando varía el punto de corte entre 0 y 1. Se utilizan los mismos datos de riesgo de crédito de las dos entidades financieras para elegir un punto de corte óptimo para estos datos al aplicar la regresión logística BD.

4.1 Modelización del riesgo de crédito

La evaluación del riesgo asociado a la concesión de créditos se ha consolidado con una de las

aplicaciones más exitosas en la estadística y la investigación operativa, el *credit scoring* (puntuación de crédito). El *credit scoring* es el conjunto de modelos y sus técnicas subyacentes que ayudan a las entidades financieras en la concesión de créditos.

Dentro de los seguros no vida, el seguro de crédito cubre el riesgo que los deudores resulten insolventes y no abonen las cantidades adeudadas (Pérez, 2001).

Las primas por riesgo de crédito se calculan haciendo uso de las probabilidades de insolvencia de los riesgos a partir de un modelo de *credit scoring*, por lo que la elección del modelo de *scoring* es un paso clave para la solvencia de una entidad financiera o de una compañía de seguros (Boj *et al.*, 2009).

El *credit scoring* constituye un problema de clasificación en el que, si se dispone de un conjunto de observaciones de las cuales se conoce *a priori* que pertenecen a una clase determinada, debe encontrarse una regla que permita clasificar a las nuevas observaciones en el grupo de individuos que podrán hacer frente a sus obligaciones crediticias con alta probabilidad (buenos riesgos) o en el grupo de individuos que resultarán fallidos (malos riesgos).

Para poder clasificar a los individuos que solicitan un crédito es necesario analizar sus características así como las características de la operación financiera. En el riesgo de crédito es usual considerar conjuntos de características que provengan de la misma fuente. Como ejemplos, se pueden considerar: Características del crédito (duración, importe, propósito,…); Características sociales del deudor (edad, estado civil, sexo, nacionalidad,….); Características económicas del deudor (antigüedad laboral, propiedades, vivienda, tipo de trabajo, personas a su cargo, cuentas bancarias, situación actual de las cuentas bancarias,…); Relación del deudor con la entidad financiera (historial crediticio, número de créditos activos con la entidad financiera,…). Todas estas características pueden ser, por tanto, variables de tipo cuantitativo, categóricas o binarias.

En las últimas décadas se han introducido varios métodos para modelizar el *credit scoring*, entre ellos los más ampliamente usados son la regresión logística, los árboles de clasificación,

el análisis discriminante lineal y las redes neuronales.

En este capítulo se estudia la viabilidad de un modelo de regresión logística BD para constituir una metodología alternativa en el cálculo de *scorings*. La metodología propuesta es adecuada en este contexto, ya que se trata de una metodología no paramétrica, que permite de modo natural una mezcla de variables numéricas y categóricas.

El modelo de regresión logística BD (ver Boj *et al.*, 2008, para una primera versión con aplicación a datos funcionales, y Costa *et al*., 2012) es una versión de la regresión logística clásica (Hosmer y Lemeshow, 2000) en el ámbito de las distancias. Se trata de un caso particular del MLGBD cuando se asume una distribución del error Binomial y una función de enlace *logit*, por tanto el modelo está construido suponiendo que es un MLG en el sentido de la familia exponencial de McCullagh y Nelder (2.5).

La variable respuesta *Y* es una variable binaria, en el caso de riesgo de crédito se construye codificando con 1 a los individuos que han resultado insolventes en el periodo de estudio y con 0 a los que no.

Se dispone de un conjunto de n individuos de una población dada $\Omega = \{\Omega_1, \ldots, \Omega_n\}$. A partir de los predictores observados $\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_p$ se calcula la matriz $\boldsymbol{\Delta}$ de distancias al cuadrado $\delta^2(\Omega_i, \Omega_j)$ con la propiedad Euclídea.

Se asume que la distribución de *Y* es una distribución Binomial, es decir, $Y_i \sim Binomial(\pi_i)$, y que la función de enlace es *logit*, de manera que:

$$\eta_i = logit(\mu_i) = log\left(\frac{\mu_i}{1+\mu_i}\right). \tag{4.1}$$

De la matriz de distancias con la propiedad Euclídea se calcula una configuración Euclídea latente, que juega el papel de predictor lineal en el modelo. A partir de aquí se propone realizar los mismos supuestos que en la regresión logística clásica. La diferencia con la

regresión logística clásica es que en el proceso iterativo de estimación por mínimos cuadrados ponderados se hace uso del MLBD en lugar del modelo lineal de regresión clásica.

En la regresión logística BD, los elementos sobre los que se itera hasta no tener cambios en la predicción del predictor lineal, son:

$$z_k = \hat{\eta}_k + (y - \hat{\mu}_k)\frac{1}{\hat{\mu}_k \cdot (1-\hat{\mu}_k)}, \tag{4.2}$$

con pesos:

$$w_k = \hat{\mu}_k \cdot (1-\hat{\mu}_k) \cdot \boldsymbol{w}, \tag{4.3}$$

siendo $\boldsymbol{w}$ unos posibles pesos *a priori* de los datos originales. El proceso iterativo propone ajustar z_k sobre la matriz de distancias con pesos w_k hasta no obtener cambios en el predictor lineal estimado $\hat{\eta}_{k+1}$ con la fórmula anterior.

En el caso de utilizar la distancia Euclídea entre predictores continuos, se obtiene el caso particular de regresión logística clásica.

Finalmente, para un individuo de la población, ω, se estima la probabilidad de insolvencia como:

$$\hat{\pi}(\omega) = \hat{\mu}(\omega) = \frac{e^{\hat{\eta}(\omega)}}{1+e^{\hat{\eta}(\omega)}}. \tag{4.4}$$

4.2 Criterios de selección de modelo en *credit scoring*

En este apartado se describen dos criterios de selección de modelo en *credit scoring*, siguiendo el trabajo de West (2000) y Boj *et al*. (2009).

En primer lugar, se estiman las probabilidades de mala clasificación de los individuos, tanto para la población de buenos riesgos y malos riesgos como para la población total. Se consideran buenos riesgos aquellos clientes que es probable que paguen sus obligaciones

financieras y malos riesgos aquellos clientes a los que debería denegarse un crédito debido a una alta probabilidad de impago de sus obligaciones financieras (West, 2000).

Para poder calcular las probabilidades necesarias debe obtenerse, en primer lugar, la matriz de confusión, que se describe en la Figura 4.1:

Figura 4.1. Matriz de confusión.

		Estimada		
		Buenos riesgos	Malos riesgos	Total
Real	Buenos riesgos	n_{11}	n_{21}	$n_{11} + n_{21}$
	Malos riesgos	n_{12}	n_{22}	$n_{12} + n_{22}$
	Total	$n_{11} + n_{12}$	$n_{21} + n_{22}$	n

En la matriz se dispone, por filas, la clasificación real de los individuos y por columnas la clasificación obtenida con la predicción estimada.

A partir de los individuos que forman esta matriz de confusión se calculan las siguientes probabilidades:

- Para el grupo de buenos riesgos, la probabilidad de mala clasificación es la proporción de solicitantes que son solventes pero que son clasificados como malos riesgos:

$$\frac{n_{21}}{n_{11} + n_{21}}$$

- Para el grupo de malos riesgos, la probabilidad de mala clasificación es la proporción de solicitantes que no son solventes pero se clasifican incorrectamente como buenos riesgos:

$$\frac{n_{12}}{n_{12}+n_{22}}$$

- Para la población total, la probabilidad de mala clasificación es la proporción de todos aquellos solicitantes que están clasificados incorrectamente:

$$\frac{n_{21}+n_{12}}{n}$$

La probabilidad de equivocarse en conceder créditos a malos riesgos es realmente importante y tampoco es bueno clasificar mal a los buenos riesgos, ya que si no se conceden créditos a buenos clientes, en términos esperados no se podrán compensar las pérdidas de los siniestros. Por todo ello, debe elegirse una técnica predictiva que mantenga un equilibrio entre las tres probabilidades.

Es evidente que la exactitud en la clasificación de cada grupo de individuos (buenos y malos riesgos) puede variar ampliamente según el modelo aplicado. En las aplicaciones de *credit scoring*, en general, se acepta que el coste de conceder un crédito a un solicitante que sea mal riesgo es significativamente mayor que denegar un crédito a un solicitante que sea un buen riesgo.

Si se denomina C_{12} al coste de conceder un crédito a un mal riesgo y C_{21} al coste de denegar un crédito a un buen riesgo, se puede definir la función de coste del error de West (2000):

$$Coste = C_{12}\pi_2 \frac{n_{12}}{n_{11}+n_{12}} + C_{21}\pi_1 \frac{n_{21}}{n_{21}+n_{22}}, \tag{4.5}$$

donde π_1 es la probabilidad *a priori* de ser un buen riesgo y π_2 es la probabilidad *a priori* de ser un mal riesgo en la cartera de solicitantes del modelo de *credit scoring*. Estas probabilidades *a priori* pueden estimarse a partir de datos recopilados de insolvencia.

4.2.1 Aplicación práctica

En este apartado se aplica la regresión logística BD a dos conjuntos de datos reales de riesgo de crédito. Ambos conjuntos son carteras de entidades financieras, los primeros de una entidad australiana y los segundos de una entidad alemana y pueden descargarse gratuitamente del repositorio *Statlog*: los datos australianos en la dirección electrónica http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval), y los datos alemanes en http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

Los cálculos de este apartado se realizan utilizando la librería *dbstats* (Boj *et al.*, 2014a) de *R*. En concreto, la función *dbglm* permite estimar el MLGBD y la regresión logística BD es un caso particular de dicho modelo cuando se asume distribución Binomial y función de enlace *logit.*

Los datos de Australia hacen referencia al riesgo asociado a tarjetas de crédito de una entidad financiera. Para mantener la confidencialidad, el autor no cedió los nombres de los factores de riesgo ni lo que significan sus clases y valores. La base de datos es de especial interés porque el conjunto de predictores es de tipo mixto y el número de datos faltantes es reducido. Tal y como se explica en Boj *et al.* (2009), para las variables continuas los datos faltantes son reemplazados por la media de la variable correspondiente, y para las variables categóricas y binarias éstos son reemplazados por la moda. En total contiene n = 690 individuos, de los cuales 307 han sido buenos riesgos y 383 han sido malos riesgos. Los factores potenciales de riesgo son 14, de los cuales 6 son continuos, 4 categóricos y 4 binarios.

A partir de los 14 predictores mixtos, se calcula la matriz de distancias al cuadrado, *D2*, como la suma pitagórica teniendo en cuenta el índice de similaridad de Gower (2.31) para cada una de las variables individualizadas. La instrucción con *dbstats* es la siguiente:

```
R> dbglmAus<-dbglm(Delta, y, family = binomial ( link = "logit"), maxiter = 50, eps1 =
0.05, eps2 = 0.05, method = "rel.gvar", rel.gvar = 0.99); dbglmAus

Call:  dbglm(D2 = Delta, y = y, family = binomial(link = "logit"), method = "rel.gvar",
```

```
   maxiter = 50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99)

family: binomial

Degrees of Freedom: 689 Total (i.e. Null);  556 Residual
Null Deviance:      948.2
Residual Deviance: 248.4

AIC:        516.3833
BIC:        1124.3
GCV:        0.5524
```

Como resultado, la matriz de confusión del modelo ajustado de regresión logística BD que se obtiene es:

$$\begin{bmatrix} 282 & 25 \\ 21 & 362 \end{bmatrix}.$$

En cuanto a los datos alemanes, la cartera contiene datos cedidos en fecha 17-11-1994. En total contiene n = 1000 individuos, de los cuales 700 han sido buenos riesgos y 300 malos riesgos. Los factores potenciales de riesgo considerados son 20, de los cuales 7 son continuos, 11 categóricos y 2 binarios.

En estos datos se realiza una agrupación previa en cuatro conjuntos de predictores, de la misma manera que en Boj *et al.* (2009) para poder comparar después los resultados obtenidos. Concretamente, los conjuntos de datos por los que se ha optado son:

- Conjunto 1: Características del crédito. Se incluyen dos variables continuas y una categórica nominal. Los factores que se consideran son:
    - Factor 2: Duración en meses (numérica)
    - Factor 5: Importe del crédito (numérica)
    - Factor 4: Propósito (categórica nominal con 11 niveles)

- Conjunto 2: Características sociales del deudor (beneficiario del crédito). Se incluyen un total de cinco factores, de los cuales dos son variables continuas, una es categórica nominal y dos son binarias:
    - Factor 11: Residencia actual desde (numérica)
    - Factor 13: Edad en años (numérica)
    - Factor 9: situación personal y sexo (categórica nominal con 5 niveles)
    - Factor 19: Teléfono (binaria)
    - Factor 20: Trabajador extranjero (binaria)

- Conjunto 3. Características económicas del deudor. Consta de cinco variables cuantitativas y cuatro categóricas nominales. Los factores que se incluyen son:
    - Factor 1: Situación actual de la cuenta corriente (categórica ordinal). Esta variable era originariamente cuantitativa y en este caso se discretiza utilizando las marcas de clase de los intervalos en los que estaba definida. Para más detalle se puede consultar Boj *et al.* (2009)
    - Factor 6: Cuenta de ahorros (categórica ordinal). Se aplica el mismo procedimiento que en la variable anterior
    - Factor 7: Tiene empleo desde (categórica ordinal). Se discretiza siguiendo el mismo tratamiento que en los dos casos anteriores
    - Factor 7b: Indicador de empleo (binaria). Se crea para tener en cuenta la categoría del Factor 7 de no disponer de empleo. En Boj *et al.* (2009) se indica el procedimiento aplicado
    - Factor 8: Cuota del crédito en porcentaje de la renta disponible (numérica)
    - Factor 18: Número de personas mantenidas (numérica)
    - Factor 12: Propiedades (categórica nominal con 4 niveles)
    - Factor 14: Otros planes periódicos (categórica nominal con 3 niveles)
    - Factor 15: Vivienda (categórica nominal con 3 niveles)
    - Factor 17: Trabajo (categórica nominal con 4 niveles)

- Conjunto 4. Relación del deudor con el banco. Se consideran tres factores, de los

cuales uno es una variable continua y dos son categóricas nominales:

- Factor 16: Número de créditos activos en este banco (numérica)
- Factor 3: Historial crediticio (categórica nominal con 5 niveles)
- Factor 10: Otras personas en el crédito (categórica nominal con 3 niveles)

En general, si se supone que se han agrupado las variables en c conjuntos, para cada conjunto de variables se calcula la matriz de distancias Euclídeas asociada:

$$\boldsymbol{\Delta}_{[s]} \text{ con } s = 1,\ldots,c \text{ de tamaño } n \times n.$$

Se puede construir la matriz $\boldsymbol{\Delta}$ como la suma pitagórica de las matrices de los conjuntos, asumiendo en este caso implícitamente la independencia entre los factores, y se tiene que:

$$\boldsymbol{\Delta} = \sum_{s=1}^{c} \boldsymbol{\Delta}_{[s]}. \tag{4.6}$$

Para incluir la dependencia entre los factores, una de las maneras es construir distancias con familias paramétricas, cuyo detalle teórico se puede consultar en Esteve (2003). Un tipo de familias son las que se obtienen como combinación lineal convexa de las diferentes matrices:

$$\boldsymbol{\Delta}(\lambda) = \sum_{s=1}^{c} \lambda_s \boldsymbol{\Delta}_{[s]}, \tag{4.7}$$

donde los parámetros λ_s cumplen que $\sum_{s=1}^{c} \lambda_s = 1$. Mediante esta familia se pondera *a priori* cada uno de los conjuntos de variables.

Para los datos alemanes se analizan seis casos distintos, al aplicar diferentes métricas al modelo de regresión logística BD. Se calculan, por lo tanto, seis matrices de distancias al cuadrado distintas y después la instrucción para ajustar el modelo es la misma que en el conjunto de datos australianos:

```
R> dbglmGer<-dbglm(Delta, y, family = binomial ( link = "logit"), maxiter = 50, eps1 =
0.05, eps2 = 0.05, method = "rel.gvar", rel.gvar = 0.99); dbglmGer

Call:  dbglm(D2 = Delta, y = y, family = binomial(link = "logit"), method = "rel.gvar",
    maxiter = 50, eps1 = 0.05, eps2 = 0.05, rel.gvar = 0.99)

family: binomial

Degrees of Freedom: 999 Total (i.e. Null);  860 Residual
Null Deviance:      1222
Residual Deviance: 780.1

AIC:        1060.143
BIC:        1747.229
GCV:         1.0524
```

Se obtienen los siguientes resultados:

- El caso general, igual que en los datos australianos, considerando que todos los predictores corresponden a un único conjunto de variables, por tanto, con suma pitagórica de las distancias aportadas por cada predictor, utilizando el índice de similaridad de Gower. La matriz de confusión resultante es:

$$\begin{bmatrix} 640 & 60 \\ 114 & 186 \end{bmatrix}.$$

- Por otro lado, se supone la familia convexa de distancias paramétricas, con diferentes pesos a priori:

    1) Para $\lambda = [0.25, 0.25, 0.25, 0.25]$ la matriz de confusión es:

$$\begin{bmatrix} 626 & 74 \\ 124 & 176 \end{bmatrix}.$$

2) Para $\lambda = [0.16, 0.05, 0.32, 0.47]$ la matriz de confusión es:

$$\begin{bmatrix} 626 & 74 \\ 122 & 178 \end{bmatrix}.$$

3) Para $\lambda = [0.14, 0.05, 0.36, 0.45]$ la matriz de confusión es:

$$\begin{bmatrix} 627 & 73 \\ 123 & 177 \end{bmatrix}.$$

4) Para $\lambda = [0.10, 0.10, 0.40, 0.40]$ la matriz de confusión es:

$$\begin{bmatrix} 627 & 73 \\ 123 & 177 \end{bmatrix}.$$

5) Para $\lambda = [0.40, 0.40, 0.10, 0.10]$ la matriz de confusión es:

$$\begin{bmatrix} 631 & 69 \\ 124 & 176 \end{bmatrix}.$$

Estas métricas coinciden con las aplicadas en Boj *et al.* (2009) para el modelo de análisis discriminante basado en distancias (ADBD). Se utilizan las mismas métricas puesto que éstas se adaptan bien a los datos y, además, permiten realizar una comparativa adecuada de resultados respecto de ambos modelos basados en distancias.

Para las dos bases de datos se estiman las probabilidades de mala clasificación de buenos y malos riesgos y la probabilidad global y se comparan los ajustes obtenidos con regresión logística BD con los obtenidos mediante otras metodologías de la literatura. Las metodologías comparadas ya han sido propuestas por varios autores para dar solución al problema de cálculo de *scorings*. En función de si son no paramétricas o paramétricas se pueden considerar los siguientes métodos:

- Métodos no-paramétricos:
    - ADBD
    - Redes neuronales
    - Método de los *k* vecinos más próximos
    - Método de la estimación núcleo de la densidad
    - Árboles de clasificación *classification and regression trees* (CART)
- Métodos paramétricos:
    - Análisis discriminante lineal
    - Regresión logística.

Cabe destacar que dentro de las técnicas no paramétricas hay dos metodologías que también son basadas en distancias: el ADBD y el método de los *k* vecinos más próximos.

Respecto de las metodologías alternativas que aparecen en las tablas comparativas de este apartado, los resultados numéricos están extraídos de West (2000) y de Boj *et al*. (2009) y, por lo tanto, para obtener información de los procesos de estimación de los métodos y las hipótesis en que se basan deben consultarse ambas referencias.

Respecto al punto de corte considerado, tanto para el caso de regresión logística clásica en West (2000) como para el caso de regresión logística BD, se ha utilizado el valor de 0.5, cuya aplicación es una práctica estándar en la literatura.

Los resultados de probabilidades estimadas de mala clasificación para cada grupo, buenos y malos riesgos, y global para los datos de crédito australianos utilizando las diferentes metodologías de *credit scoring* quedan recogidos en la Tabla 4.1.

Para estos datos se observa que la metodología con menores probabilidades de mala clasificación, tanto para los malos riesgos como global, es la regresión logística BD, con probabilidades 0.055 y 0.067, respectivamente. La siguiente técnica con menor probabilidad

de mala clasificación de malos riesgos es CART, con 0.120, seguida de la red neuronal MOE, con 0.124. Centrándose en la probabilidad total, la siguiente técnica con menor probabilidad es la regresión logística, con 0.127, seguida de la red neuronal RBF con un valor de 0.128.

Tabla 4.1. Probabilidades de mala clasificación para los buenos riesgos, para los malos riesgos y total aplicando distintas metodologías de *credit scoring* a los datos australianos.

	Buenos riesgos	Malos riesgos	Total
Modelos no paramétricos			
Regresión logística BD (suma pitagórica)	0.081	0.055	0.067
ADBD (suma pitagórica)	0.094	0.162	0.132
Red neuronal MOE	0.145	0.124	0.133
Red neuronal RBF	0.131	0.127	0.128
Red neuronal MLP	0.154	0.132	0.141
Red neuronal LVQ	0.171	0.171	0.170
Red neuronal FAR	0.256	0.238	0.246
k vecinos más próximos	0.153	0.133	0.142
Estimación núcleo de la densidad	0.185	0.151	0.166
Árbol de clasificación CART	0.192	0.120	0.156
Modelos paramétricos			
Análisis discriminante lineal	0.078	0.190	0.140
Regresión logística	0.110	0.140	0.127

En la siguiente Tabla 4.2 se muestran las probabilidades estimadas de mala clasificación para cada grupo, buenos y malos riesgos, y la probabilidad global para los datos alemanes utilizando diferentes metodologías de *credit scoring*.

Tabla 4.2. Probabilidades de mala clasificación para los buenos riesgos, para los malos riesgos y total aplicando distintas metodologías de *credit scoring* a los datos alemanes.

	Buenos riesgos	Malos riesgos	Total
Modelos no paramétricos			
Regresión logística BD (suma pitagórica)	0.086	0.38	0.174
Regresión logística BD λ=[0.25,0.25,0.25,0.25]	0.105	0.413	0.198
Regresión logística BD λ=[0.16,0.05,0.32,0.47]	0.105	0.407	0.196
Regresión logística BD λ=[0.14,0.05,0.36,0.45]	0.104	0.41	0.196
Regresión logística BD λ=[0.10,0.10,0.40,0.40]	0.104	0.41	0.196
Regresión logística BD λ=[0.40,0.40,0.10,0.10]	0.099	0.413	0.193
ADBD (suma pitagórica)	0.223	0.627	0.344
ADBD λ=[0.25,0.25,0.25,0.25]	0.350	0.287	0.331
ADBD λ=[0.16,0.05,0.32,0.47]	0.437	0.243	0.379
ADBD λ=[0.14,0.05,0.36,0.45]	0.419	0.253	0.369
ADBD λ=[0.10,0.10,0.40,0.40]	0.400	0.27	0.361
ADBD λ=[0.40,0.40,0.10,0.10]	0.341	0.353	0.345
Red neuronal MOE	0.142	0.477	0.243
Red neuronal RBF	0.134	0.529	0.254
Red neuronal MLP	0.135	0.575	0.267
Red neuronal LVQ	0.249	0.481	0.316
Red neuronal FAR	0.403	0.488	0.427
k vecinos más próximos	0.225	0.553	0.324
Estimación núcleo de la densidad	0.155	0.630	0.308
Árbol de clasificación CART	0.206	0.545	0.304
Modelos paramétricos			
Análisis discriminante lineal	0.277	0.266	0.274
Regresión logística	0.188	0.513	0.237

En este caso, para la probabilidad de mala clasificación en los malos riesgos, la metodología con menor probabilidad es el ADBD con $\lambda = [0.16, 0.05, 0.32, 0.47]$, con un valor de 0.243, seguida de la misma metodología pero con la métrica λ = [0.14,0.05,0.36,0.45], con un valor de 0.253. La siguiente técnica con menor probabilidad es el análisis discriminante lineal, con una probabilidad de 0.266. Se observa por lo tanto que el análisis discriminante (el paramétrico clásico y el basado en distancias) parece ser la técnica más adecuada para minimizar esta probabilidad. Si se comparan los resultados para la regresión logística, la que proporciona un mejor resultado es la basada en distancias, con un valor de 0.38 (suma pitagórica) en comparación en la probabilidad de 0.513 de la clásica. Cabe destacar que después del análisis discriminante (tanto clásico como basado en distancias) la siguiente técnica con menor probabilidad es la regresión logística BD (como suma pitagórica).

En la probabilidad global, el método con un menor error de clasificación es la regresión logística BD (suma pitagórica) con una probabilidad de 0.174. Las siguientes probabilidades con un valor menor vienen proporcionadas por el resto de casos de regresión logística BD (entre 0.193 y 0.198) y por la regresión logística clásica, con una probabilidad de 0.237.

Para ilustrar el cálculo del coste del error definido en West (2000) se hacen las siguientes consideraciones:

- Con respecto al coste de conceder un crédito a un mal riesgo, se supone $C_{12} = 5$ y con respecto al coste de denegar un crédito a un mal riesgo, se supone $C_{21} = 2$, tal como propone el Dr. Hans Hofmann, responsable de la recopilación de los datos de crédito alemanes. Este mismo supuesto se tiene en cuenta en West (2000) y en Boj *et al.* (2009).

- Con respecto a las probabilidades *a priori* de ser un buen riesgo, π_1, o de ser un mal riesgo, π_2, se considera la propuesta de West (2000) de fijar un valor mínimo de $\pi_2 = 0.144$ y un valor máximo de $\pi_2 = 0.249$. En este sentido, se están considerando dos escenarios y se obtiene entre qué valores puede oscilar el coste del error si se

dieran ambas situaciones.

Estos mismos supuestos se tienen en cuenta en Boj *et al.* (2009) en la aplicación del método de ADBD.

Los resultados del coste de error para los datos de crédito australianos utilizando las diferentes metodologías de *credit scoring* quedan recogidos en la siguiente Tabla 4.3:

Tabla 4.3. Costes del error estimados suponiendo π_2=0.144 y π_2=0.249 aplicando distintas metodologías de *credit scoring* a los datos australianos.

	π_2=0.144	π_2=0.249
Modelos no paramétricos		
Regresión logística BD (suma pitagórica)	0.105	0.135
ADBD (suma pitagórica)	0.202	0.289
Red neuronal MOE	0.196	0.243
Red neuronal RBF	0.194	0.245
Red neuronal MLP	0.198	0.243
Red neuronal LVQ	0.237	0.300
Red neuronal FAR	0.319	0.388
k vecinos más próximos	0.227	0.281
Estimación núcleo de la densidad	0.267	0.328
Árbol de clasificación CART	0.251	0.294
Modelos paramétricos		
Análisis discriminante lineal	0.204	0.296
Regresión logística	0.196	0.258

Se observa que los menores costes para los dos escenarios se obtienen con regresión logística BD, con valores 0.105 y 0.135, respectivamente. Las siguientes metodologías con menores costes, cuando la probabilidad *a priori* es de 0.144, son la red neuronal RBF, con 0.194,

seguidas de la red MOE y la regresión logística, que tienen el mismo valor de 0.196. En el otro escenario, en que la probabilidad *a priori* es de 0.249, los siguientes menores costes son para la red neuronal MOE y la red neuronal MLP con el mismo valor de 0.243, seguidos por la red neuronal RBF, con 0.245.

En cuanto a los costes del error suponiendo los dos escenarios, π_2=0.144 y π_2=0.249, para los datos alemanes según las diferentes metodologías de *credit scoring*, los resultados quedan recogidos en la Tabla 4.4.

Se observa que los menores costes para los dos escenarios se obtienen con los modelos de regresión logística BD. En concreto, el menor se obtiene en el caso de suma pitagórica de distancias, con un valor de 0.318 cuando π_2=0.144 y de 0.371 cuando π_2=0.249. Este modelo coincide también con el de menor probabilidad global, tal como se observa en los resultados de la Tabla 4.2. En cuanto al resto de metodologías, la que presenta un menor coste es el análisis discriminante lineal, con un valor de 0.429 en el escenario en que la probabilidad *a priori* de la población de malos riesgos es π_2=0.144 y con un valor de 0.540 en el escenario en que π_2=0.249.

Tabla 4.4. Costes del error estimados suponiendo π_2=0.144 y π_2=0.249 aplicando distintas metodologías de *credit scoring* a los datos alemanes.

	π_2=0.144	π_2=0.249
Modelos no paramétricos		
Regresión logística BD (suma pitagórica)	0.318	0.371
Regresión logística BD λ=[0.25,0.25,0.25,0.25]	0.372	0.428
Regresión logística BD λ=[0.16,0.05,0.32,0.47]	0.369	0.423
Regresión logística BD λ=[0.14,0.05,0.36,0.45]	0.368	0.424
Regresión logística BD λ=[0.10,0.10,0.40,0.40]	0.368	0.423
Regresión logística BD λ=[0.40,0.40,0.10,0.10]	0.359	0.416
ADBD (suma pitagórica)	0.683	0.756
ADBD λ=[0.25,0.25,0.25,0.25]	0.562	0.591
ADBD λ=[0.16,0.05,0.32,0.47]	0.604	0.625
ADBD λ=[0.14,0.05,0.36,0.45]	0.598	0.621
ADBD λ=[0.10,0.10,0.40,0.40]	0.596	0.622
ADBD λ=[0.40,0.40,0.10,0.10]	0.607	0.647
Red neuronal MOE	0.432	0.653
Red neuronal RBF	0.469	0.707
Red neuronal MLP	0.483	0.758
Red neuronal LVQ	0.501	0.714
Red neuronal FAR	0.668	0.942
k vecinos más próximos	0.592	0.858
Estimación núcleo de la densidad	0.587	0.901
Árbol de clasificación CART	0.569	0.834
Modelos paramétricos		
Análisis discriminante lineal	0.429	0.540
Regresión logística	0.471	0.728

4.3 Elección del punto de corte en regresión logística basada en distancias

En las aplicaciones realizadas en Boj *et al.* (2009) y en Boj *et al.* (2011) con ADBD y regresión logística BD se considera un punto de corte de 0.5, ya que se trata del valor usualmente utilizado en las metodologías de *credit scoring* de la literatura (se aplica también en West, 2000).

En este apartado se considera que el punto de corte, *s*, puede ser un valor comprendido entre 0 y 1. La justificación de este supuesto es que no siempre puede resultar adecuado utilizar el punto de corte de 0.5 si se cuenta con datos reales no balanceados, en *credit scoring* en el caso en que los individuos insolventes no supongan, aproximadamente, la mitad de la cartera. Teniendo en cuenta este supuesto, en primer lugar, dado un punto de corte *s*, la matriz de confusión se calcula tal y como se indica en la Figura 4.2:

Figura 4.2. Matriz de confusión para un punto de corte *s*.

		Estimada: Buenos riesgos	Estimada: Malos riesgos	Total
Real	Buenos riesgos	n_{11}^s	n_{21}^s	$n_{11}^s + n_{21}^s$
	Malos riesgos	n_{12}^s	n_{22}^s	$n_{12}^s + n_{22}^s$
	Total	$n_{11}^s + n_{12}^s$	$n_{21}^s + n_{22}^s$	n

Con estos resultados se definen los criterios de calidad de ajuste con los que elegir un punto de corte “óptimo” para unos datos determinados.

4.3.1 Calidad del modelo: Cálculo del coeficiente Kolmogorov-Smirnov e índice de Gini. Representación gráfica de la curva ROC.

Para medir la calidad del modelo de *credit scoring* es usual utilizar índices cuantitativos como el coeficiente Kolmogorov-Smirnov (K-S) o el índice de Gini, que se basan en la función de distribución o probabilidades acumuladas.

El coeficiente K-S varía entre 0 y 1 y se obtiene como la máxima diferencia entre las distribuciones acumuladas de malos riesgos y buenos riesgos. Además de medir la calidad del ajuste, identifica el valor del *score* para el cual se maximiza dicho coeficiente, y es ”ideal” si el punto de corte “esperado” es cercano a dicho *score* (Řezáč y Řezáč, 2011).

El procedimiento de cálculo del punto de corte óptimo $\hat{s}^*$ que maximiza el coeficiente K-S es el siguiente (Íñiguez y Morales, 2009):

- Ordenar los valores de los puntos de corte, s, de manera ascendente.
- Calcular la proporción de buenos y malos riesgos que comparten el mismo punto de corte, $p_b(s)$ y $p_m(s)$, siendo:

$$p_b(s) = \frac{n_{11}^s + n_{12}^s}{\sum_s n_{11}^s + n_{12}^s} \qquad \text{y} \qquad p_m(s) = \frac{n_{21}^s + n_{22}^s}{\sum_s n_{21}^s + n_{22}^s}.$$

- Calcular la proporción acumulada de buenos y malos riesgos $P_b(s)$ y $P_m(s)$:

$$P_b(s) = \sum_{S \le s} p_b(S) \qquad \text{y} \qquad P_m(s) = \sum_{S \le s} p_m(S).$$

- Calcular las diferencias entre proporciones acumuladas por punto de corte entre buenos y malos riesgos: $\left|P_m(s) - P_b(s)\right|$.
- Identificar el punto de corte $\hat{s}^*$ que proporciona la máxima diferencia absoluta del coeficiente K-S:

$$\text{K-S} = \max_s \left\{\left|P_m(s) - P_b(s)\right|\right\}. \tag{4.8}$$

El índice de Gini es una medida global de calidad del modelo, que se puede calcular a partir de la expresión (Íñiguez y Morales, 2009):

$$Gini = 1 - \sum_{i=1}^{n} \left(P_m(s_i) - P_m(s_{i-1}) \right) \left(P_b(s_i) + P_b(s_{i-1}) \right), \tag{4.9}$$

con $P_m(s_0) = 0,\ P_b(s_0) = 0$, donde:

- $P_m(s_i)$ es la proporción acumulada de malos riesgos para un *score* s_i,
- $P_m(s_{i-1})$ es la proporción acumulada de malos riesgos para un *score* anterior a s_i,
- $P_b(s_i)$ es la proporción acumulada de buenos riesgos para un *score* s_i,
- $P_b(s_{i-1})$ es la proporción acumulada de buenos riesgos para un *score* anterior a s_i.

El modelo "ideal", es decir, que predice exactamente los buenos y malos riesgos, tiene un índice de Gini igual a 1; en caso contrario, aquel modelo que asigna un *score* aleatorio al cliente tiene un índice de Gini igual a 0 (Řezáč y Řezáč, 2011). El índice de Gini puede resultar útil para comparar los obtenidos con distintos modelos para un mismo conjunto de datos.

En el caso de riesgo de crédito, para la determinación del punto de corte o valor del *score s* a partir del cual decidir si un cliente es un mal riesgo de crédito también se puede analizar gráficamente la denominada curva ROC (*Receiver Operating Characteristic*). La curva ROC fue desarrollada inicialmente por ingenieros para la estimación de errores en la transmisión de mensajes y se ha aplicado posteriormente en áreas como la medicina y la estadística.

Una curva ROC describe el comportamiento predictivo de un sistema clasificador. En un espacio ROC se pueden representar los intercambios entre verdaderos positivos (eje de ordenadas) y falsos positivos (eje de abcisas).

En el caso de la aplicación del modelo de regresión logística BD en *credit scoring*, a partir de la matriz de confusión obtenida con cada valor del punto de corte, *s*, se calculan las razones

de verdaderos positivos y falsos positivos. En el caso que se estudia, los verdaderos positivos son aquellos malos riesgos predichos como malos (en la matriz de confusión el elemento n_{22}^s) y los falsos positivos son aquellos buenos riesgos predichos como malos (en la matriz de confusión el elemento n_{21}^s). Por tanto, según se van variando los puntos de corte o frontera, *s*, se obtienen los distintos pares de puntos que conforman la curva ROC.

Otra interpretación posible de la curva ROC es la de representar la Sensibilidad (eje de ordenadas) frente a (1 – Especificidad) (eje de abcisas), como se indica en Reyes *et al.* (2007). De esta manera, en función de los elementos de la matriz de confusión, se tiene que:

- Sensibilidad: $\frac{n_{22}^s}{n_{12}^s + n_{22}^s}$, proporción de malos riesgos predichos como malos
- Especificidad: $\frac{n_{11}^s}{n_{11}^s + n_{21}^s}$, proporción de buenos riesgos predichos como buenos.

Además, en el gráfico donde se represente una curva ROC se puede identificar al punto de corte óptimo $\hat{s}^*$ que maximiza el K-S, que se corresponde con el punto en la curva cuya distancia horizontal al eje de abcisas es máxima (Balzarotti y Castelpoggi, 2009).

4.3.1.1 Aplicación práctica

En este apartado, en primer lugar, se calcula el coeficiente K-S utilizando distintos puntos de corte entre 0 y 1, variando en incrementos de 0.05, para localizar con qué valor del punto de corte se consigue maximizar este coeficiente, aplicando la metodología de regresión logística BD.

Los resultados numéricos de este apartado obtenidos para los datos de riesgo de crédito australianos y alemanes se incluyen en Costa *et al.* (2012).

En la siguiente Tabla 4.5 se muestran los resultados obtenidos para los datos de riesgo de crédito australianos:

Tabla 4.5. Proporción de buenos riesgos acumulados, de malos riesgos acumulados y coeficiente K-S para distintos puntos de corte $s \in (0,1)$ aplicando regresión logística BD a los datos australianos.

Punto de corte s	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	Coeficiente K-S
0.05	0.0152	0.0392	0.0240
0.1	0.0341	0.0755	0.0414
0.15	0.0557	0.1095	0.0538
0.2	0.0783	0.1428	0.0645
0.25	0.1024	0.1750	0.0726
0.3	0.1278	0.2060	0.0782
0.35	0.1537	0.2368	0.0831
0.4	0.1801	0.2671	0.0870
0.45	0.3163	0.4151	0.0988
0.5	0.4563	0.5603	0.1040
0.55	0.6015	0.7013	0.0998
0.6	0.7504	0.8394	0.0890
0.65	0.7817	0.8658	0.0841
0.7	0.8139	0.8914	0.0775
0.75	0.8468	0.9166	0.0698
0.8	0.8815	0.9403	0.0588
0.85	0.9178	0.9627	0.0449
0.9	0.9566	0.9831	0.0265
0.95	1	1	0

Como puede observarse en la Tabla 4.5, los valores más elevados del coeficiente K-S para los datos australianos se obtienen entre los puntos de corte de 0.45 y 0.55. A continuación se calcula el coeficiente K-S para los puntos de corte del intervalo $[0.45, 0.55]$, variando en incrementos de 0.01, para analizar con qué valor de s se obtiene el mayor coeficiente K-S.

Tabla 4.6. Proporción de buenos riesgos acumulados, de malos riesgos acumulados y coeficiente K-S para distintos puntos de corte $s \in [0.45, 0.55]$ aplicando regresión logística BD a los datos australianos.

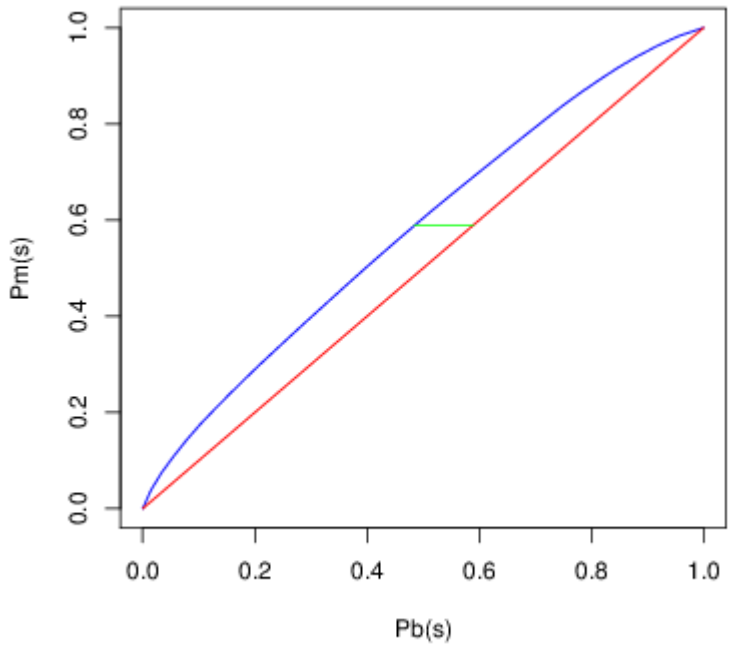
Punto de corte s	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	Coeficiente K-S
0.45	0.3163	0.4151	0.0988
0.46	0.3439	0.4445	0.1006
0.47	0.3717	0.4737	0.1020
0.48	0.3996	0.5028	0.1032
0.49	0.4279	0.5316	0.1037
0.5	0.4563	0.5603	0.1040
0.51	0.4848	0.5889	0.1041
0.52	0.5134	0.6174	0.1040
0.53	0.5426	0.6455	0.1029
0.54	0.5720	0.6734	0.1014
0.55	0.6015	0.7013	0.0998

Como se desprende de los resultado numéricos de la Tabla 4.6, el punto de corte óptimo que maximiza el coeficiente K-S es $\hat{s}^* = 0.51$, al aplicar regresión logística BD a los datos de riesgo de crédito australianos.

En el siguiente Gráfico 4.1 se representa la curva ROC y el punto de corte que maximiza el

coeficiente K-S, que se corresponde con el punto en la curva ROC cuya distancia horizontal al eje es máxima.

Gráfico 4.1. Curva ROC para los datos de riesgo de crédito australianos.



Finalmente, se puede calcular el índice de Gini para los datos australianos, en cuyo caso son necesarias las proporciones acumuladas de buenos y malos riesgos, es decir, los mismos datos que se han utilizado para el cálculo del coeficiente K-S. En los datos de riesgo de crédito australianos el valor obtenido para el índice de Gini es 0.16.

En la Tabla 4.7 se muestran los resultados del coeficiente K-S utilizando distintos puntos de corte entre 0 y 1, variando en incrementos de 0.05, para localizar con qué valor del punto de corte se consigue maximizar este coeficiente, aplicando la metodología de regresión logística BD a los datos de riesgo de crédito alemanes.

Analizado los resultados obtenidos se puede comprobar que el coeficiente K-S toma sus valores más elevados entre los puntos de corte de 0.4 y 0.5. Seguidamente, en la Tabla 4.8, se

calcula el coeficiente K-S para los puntos de corte del intervalo $[0.4, 0.5]$, variando en incrementos de 0.01, para analizar con qué valor de s se obtiene el mayor coeficiente K-S.

Tabla 4.7. Proporción de buenos riesgos acumulados, de malos riesgos acumulados y coeficiente K-S para distintos puntos de corte $s \in (0,1)$ aplicando regresión logística BD a los datos alemanes.

Punto de corte s	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	Coeficiente K-S
0.05	0.0075	0.0828	0.0753
0.1	0.0199	0.1531	0.1332
0.15	0.0356	0.2150	0.1794
0.2	0.0546	0.2681	0.2135
0.25	0.0761	0.3150	0.2389
0.3	0.0994	0.3571	0.2577
0.35	0.1244	0.3948	0.2704
0.4	0.1510	0.4286	0.2776
0.45	0.2872	0.5885	0.3013
0.5	0.4317	0.7271	0.2954
0.55	0.5856	0.8416	0.2560
0.6	0.7445	0.9432	0.1987
0.65	0.7781	0.9588	0.1807
0.7	0.8125	0.9721	0.1596
0.75	0.8481	0.9826	0.1345
0.8	0.8847	0.9906	0.1059
0.85	0.9223	0.9959	0.0736
0.9	0.9609	0.9987	0.0378
0.95	1	1	0

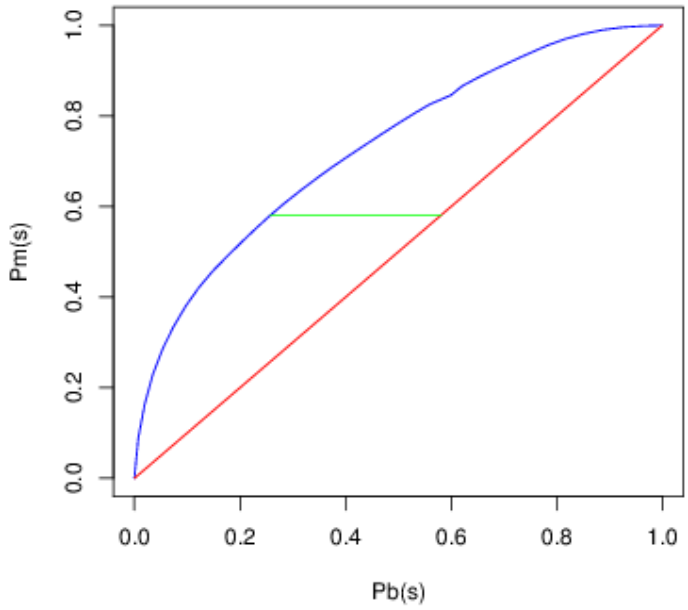
Tabla 4.8. Proporción de buenos riesgos acumulados, de malos riesgos acumulados y coeficiente K-S para distintos puntos de corte $s \in [0.4, 0.5]$ aplicando regresión logística BD a los datos alemanes.

Punto de corte s	Proporción de buenos riesgos acumulados	Proporción de malos riesgos acumulados	Coeficiente K-S
0.40	0.1510	0.4286	0.2776
0.41	0.1776	0.4623	0.2847
0.42	0.2045	0.4950	0.2905
0.43	0.2319	0.5268	0.2949
0.44	0.2544	0.5581	0.2987
0.45	0.2872	0.5885	0.3013
0.46	0.3153	0.6183	0.3030
0.47	0.3437	0.6473	0.3036
0.48	0.3725	0.6753	0.3028
0.49	0.4018	0.7020	0.3002
0.50	0.4317	0.7271	0.2954

Por tanto, al aplicar regresión logística BD a los datos de riesgo de crédito alemanes, el punto de corte óptimo que maximiza el coeficiente K-S es $\hat{s}^* = 0.47$.

Gráficamente, se puede representar la curva ROC e identificar donde se maximiza el coeficiente K-S, tal como se muestra en el Gráfico 4.2.

Gráfico 4.2. Curva ROC para los datos de riesgo de crédito alemanes.



Por último, el índice de Gini que se obtiene para los datos de riesgo de crédito alemanes es igual a 0.44.

4.3.2 Estudio de las probabilidades de mala clasificación y de los costes del error en función del punto de corte

En este apartado se estudia, en primer lugar, el comportamiento de las probabilidades de mala clasificación de buenos riesgos, de malos riesgos y global para los distintos puntos de corte, s, entre 0 y 1.

A partir de los individuos que forman la matriz de confusión, dado un valor de s, se estiman las siguientes probabilidades:

- Para el grupo de buenos riesgos, la probabilidad de mala clasificación es la proporción de solicitantes que son solventes pero que son clasificados como malos riesgos:

$$\frac{n_{21}^s}{n_{11}^s + n_{21}^s}$$

- Para el grupo de malos riesgos, la probabilidad de mala clasificación es la proporción de solicitantes que no son solventes pero se clasifican incorrectamente como buenos riesgos:

$$\frac{n_{12}^s}{n_{12}^s + n_{22}^s}$$

- Para la población total, la probabilidad de mala clasificación es la proporción de todos aquellos solicitantes que están clasificados incorrectamente:

$$\frac{n_{21}^s + n_{12}^s}{n}$$

Por otro lado, se puede definir la función de coste definida en West (2000) para un determinado punto de corte, s, mediante la siguiente expresión:

$$Coste^s = C_{12}\pi_2 \frac{n_{12}^s}{n_{11}^s + n_{12}^s} + C_{21}\pi_1 \frac{n_{21}^s}{n_{21}^s + n_{22}^s}, \tag{4.10}$$

siendo C_{12} el coste de conceder un crédito a un mal riesgo, C_{21} el coste de denegar un crédito a un buen riesgo, π_1 la probabilidad *a priori* de ser un buen riesgo y π_2 la probabilidad *a priori* de ser un mal riesgo en la cartera de solicitantes del modelo de *credit scoring*.

4.3.2.1 Aplicación práctica

En este apartado se estiman las probabilidades de mala clasificación y los costes del error teniendo en cuenta distintos valores para el punto de corte entre 0 y 1 en la regresión logística BD para los datos de riesgo de crédito australianos y alemanes. Los resultados que se presentan se incluyen en Costa *et al.* (2012).

En la siguiente Tabla 4.9 se muestran los resultados de las probabilidades de mala clasificación para los datos de riesgo de crédito australianos utilizando distintos puntos de corte entre 0 y 1, variando en incrementos de 0.05.

Tabla 4.9. Probabilidades de mala clasificación para los buenos riesgos, para los malos riesgos y total para distintos puntos de corte $s \in (0,1)$ aplicando regresión logística BD a los datos australianos.

Punto de corte s	Buenos riesgos	Malos riesgos	Total
0.05	0.479	0.005	0.216
0.1	0.349	0.005	0.195
0.15	0.261	0.010	0.122
0.2	0.225	0.010	0.106
0.25	0.179	0.013	0.087
0.3	0.143	0.021	0.075
0.35	0.134	0.026	0.074
0.4	0.127	0.037	0.077
0.45	0.098	0.044	0.068
0.5	0.081	0.055	0.067
0.55	0.065	0.073	0.070
0.6	0.062	0.084	0.074
0.65	0.046	0.107	0.080
0.7	0.036	0.123	0.087
0.75	0.029	0.138	0.090
0.8	0.026	0.180	0.112
0.85	0.020	0.227	0.135
0.9	0.013	0.292	0.168
0.95	0	0.407	0.226

Analizando la probabilidad de mala clasificación global el mejor resultado, es decir, la menor probabilidad de mala clasificación, se encuentra en el punto de corte $s = 0.5$.

En el punto de corte $\hat{s}^* = 0.51$, donde se maximiza el coeficiente K-S, se obtienen unas probabilidades de mala clasificación de buenos riesgos, malos riesgos y global de 0.081, 0.06 y 0.07, respectivamente.

En la siguiente Tabla 4.10 se muestran los costes del error estimados en los dos escenarios para los datos de riesgo de crédito australianos utilizando distintos puntos de corte entre 0 y 1, variando en incrementos de 0.05.

Tabla 4.10. Costes del error estimados suponiendo π_2=0.144 y π_2=0.249 para distintos puntos de corte $s \in (0,1)$ aplicando regresión logística BD a los datos australianos.

Punto de corte s	π_2=0.144	π_2=0.249
0.05	0.247	0.224
0.1	0.195	0.177
0.15	0.162	0.152
0.2	0.144	0.136
0.25	0.122	0.120
0.3	0.111	0.116
0.35	0.111	0.119
0.4	0.118	0.134
0.45	0.106	0.129
0.5	0.105	0.135
0.55	0.110	0.151
0.6	0.116	0.163
0.65	0.122	0.182
0.7	0.130	0.201
0.75	0.131	0.208
0.8	0.156	0.252
0.85	0.178	0.294
0.9	0.207	0.347
0.95	0.243	0.419

Se observa que, en el escenario en que la probabilidad *a priori* es de 0.144, el mínimo coste se obtiene en el punto de corte $s = 0.5$, con un valor de 0.105, mientras que en el otro escenario el mínimo coste es de 0.116 y se consigue cuando el punto de corte es $s = 0.3$.

Los costes en los dos escenarios para el punto de corte de $\hat{s}^* = 0.51$ son de 0.110 y 0.143 respectivamente.

Tabla 4.11. Probabilidades de mala clasificación para los buenos riesgos, para los malos riesgos y total para distintos puntos de corte $s \in (0,1)$ aplicando regresión BD los datos alemanes.

Punto de corte s	Buenos riesgos	Malos riesgos	Total
0.05	0.733	0.010	0.516
0.1	0.567	0.030	0.406
0.15	0.463	0.063	0.343
0.2	0.361	0.110	0.286
0.25	0.287	0.143	0.244
0.3	0.239	0.183	0.222
0.35	0.186	0.203	0.191
0.4	0.159	0.263	0.189
0.45	0.127	0.303	0.180
0.5	0.086	0.380	0.174
0.55	0.059	0.420	0.167
0.6	0.043	0.477	0.173
0.65	0.031	0.563	0.191
0.7	0.017	0.607	0.194
0.75	0.014	0.690	0.217
0.8	0.010	0.763	0.236
0.85	0.006	0.840	0.256
0.9	0.001	0.913	0.275
0.95	0	0.957	0.287

Tabla 4.12. Costes del error estimados suponiendo π_2=0.144 y π_2=0.249 para distintos puntos de corte $s \in (0,1)$ aplicando regresión logística BD a los datos alemanes.

Punto de corte s	π_2=0.144	π_2=0.249
0.05	0.554	0.495
0.1	0.515	0.469
0.15	0.493	0.462
0.2	0.466	0.451
0.25	0.433	0.428
0.3	0.414	0.421
0.35	0.371	0.385
0.4	0.369	0.397
0.45	0.349	0.386
0.5	0.318	0.371
0.55	0.279	0.343
0.6	0.264	0.339
0.65	0.267	0.357
0.7	0.230	0.330
0.75	0.249	0.360
0.8	0.256	0.377
0.85	0.257	0.389
0.9	0.234	0.378
0.95	0.209	0.362

En la Tabla 4.11 anterior se muestran los resultados obtenidos en las probabilidades de mala clasificación para los datos de riesgo de crédito alemanes cuando se aplica regresión logística BD, considerando puntos de corte que varían entre 0 y 1 con incrementos de 0.05.

Si se considera la probabilidad de mala clasificación global, el valor más pequeño se obtiene para el punto de corte $s = 0.55$.

Las probabilidades de mala clasificación de buenos riesgos, malos riesgos y global para el punto de corte $\hat{s}^* = 0.47$, donde se maximiza el coeficiente K-S para estos datos, son 0.116, 0.323 y 0.178, respectivamente.

Por último, en la Tabla 4.12 se muestran los resultados obtenidos en los costes del error de la mala clasificación para los datos de riesgo de crédito alemanes cuando se aplica regresión logística BD, considerando puntos de corte que varían entre 0 y 1 con incrementos de 0.05.

Se obtiene que los puntos de corte que minimizan el coste son $s = 0.95$ en el primer escenario y $s = 0.7$ en el segundo.

Los costes en los dos escenarios para el punto de corte $\hat{s}^* = 0.47$ son 0.342 y 0.383, respectivamente.

Capítulo 5

Cálculo de provisiones en los seguros no vida

En este capítulo se estudian las provisiones técnicas en los seguros no vida y su cálculo aplicando diferentes métodos estadísticos. Se describen algunos métodos estocásticos, que permiten estimar el error cometido en la predicción de los pagos futuros que debe afrontar una entidad aseguradora. Por último, se contextualiza el problema del cálculo de provisiones teniendo en cuenta la Directiva Solvencia II, donde se especifica que se debe incluir un margen de riesgo.

La aplicabilidad del MLG en el cálculo de provisiones en los seguros no vida se ha presentado en diferentes congresos internacionales:

- En la *International Conference on Risk Analysis ICRA6/Risk 2015*, celebrada en Barcelona del 26 al 29 de Mayo de 2015, se ha presentado "Claim reserving: calendar year reserves for the GLM", de Boj, E. y T. Costa.

- En la *16th Applied Stochastic Models and Data Analysis International Conference*, celebrada en Piraeus (Grecia) del 30 de Junio al 4 de Julio de 2015, se ha presentado "Claim reserving including risk margins", de Boj, E. y T. Costa.

- En el *1st Workshop on Pensions and Insurance*, celebrado en Barcelona del 1 al 2 de Julio de 2015, se ha presentado "Claim reserving with generalized linear models including risk margins", de Boj, E. y T. Costa.

En otros congresos internacionales se ha presentado la aplicación del MLGBD en el cálculo de provisiones en seguros no vida:

- En la *2nd Conference of the International Society of NonParametric Statisitcs*, celebrada en Cádiz del 12 al 16 de Junio de 2014, se ha presentado "Claim reserving using distance-based generalized linear models", de Boj, E. y T. Costa.

- En la *Conference of the International Federation of Classification Societies*, celebrada en Bologna (Italia) del 6 al 8 de Julio de 2015, se ha presentado "Prediction error in distance-based generalized linear models", de Boj, E.; Fortiana, J. y T. Costa.

También cabe resaltar dos artículos publicados con contenidos que se incluyen en este capítulo:

- "Provisiones técnicas por años de calendario mediante el modelo lineal generalizado. Una aplicación con RExcel" de Boj, E.; Costa, T. y J. Espejo, en la revista *Anales del Instituto de Actuarios Españoles*, Tercera Epoca, 20, 83-116, en 2014.

- "Provisions for claims outstanding, incurred but not reported, with generalized linear models: prediction error formulation by calendar years" de Boj, E. y T. Costa, en la revista *Cuadernos de gestión*, que ha sido aceptado en 2015.

Este capítulo se estructura en tres apartados y un anexo:

En el primer apartado, 5.1, se describe el concepto de provisiones técnicas y el contexto legal en el que se enmarca su cálculo, haciendo mención a la Directiva Solvencia II.

En el segundo apartado, 5.2, se indican las distintas provisiones técnicas y se describen algunos aspectos de la provisiones de siniestros pendientes.

El apartado 5.3 está dedicado a los métodos estadísticos que se pueden emplear en el cálculo de provisiones de siniestros pendientes. Se hace una distinción entre los métodos deterministas, aplicados tradicionalmente, y los métodos estocásticos, que son los que

actualmente se adaptan a las exigencias de Solvencia II. Entre los métodos deterministas se dedica especial atención al método de Chain-ladder, ampliamente estudiado en la literatura actuarial y generalizado por varios autores desde el punto de vista estocástico. Dentro de los métodos estocásticos se resalta la aplicabilidad del MLG en el cálculo de provisiones, especialmente el caso en que se asume la distribución Poisson sobredispersa y la función de enlace logarítmica, debido a que generaliza el método de Chain-ladder clásico. Como novedad, se propone la aplicación del MLGBD para el cálculo de provisiones, que generaliza el MLG al campo de las distancias y, además, contiene como caso particular el método de Chain-ladder clásico. En este capítulo se obtienen las expresiones que estiman los errores de predicción en el MLG para los pagos futuros por años de calendario, completando las formulaciones elaboradas por otros autores para las provisiones por año de origen y para la provisión total. Aplicando la metodología *bootstrap* se definen las expresiones que estiman los errores de predicción para los pagos futuros por años de calendario en el MLG y en el MLGBD. Se ilustra numéricamente la aplicación de los distintos modelos tanto deterministas como estocásticos presentados en el capítulo, utilizando unos datos de cuantías de siniestros con los que se estiman los importes de los pagos futuros con los distintos métodos. Finalmente, se definen distintas maneras de considerar márgenes de riesgo con sentido estadístico en el cálculo de provisiones, siguiendo las recomendaciones de Solvencia II, y se incluyen algunos ejemplos numéricos basados en los mismos datos que se utilizan a lo largo de este capítulo.

Por último, en el anexo se incluyen funciones en *R* que se han programado para realizar algunos de los cálculos numéricos.

5.1 Definición y contexto legal

Las provisiones técnicas deben reflejar en el balance de las entidades aseguradoras el importe de las obligaciones asumidas que se derivan de los contratos de seguros y reaseguros. (Ley 20/2015, de 14 de julio, de ordenación, supervisión y solvencia de las entidades aseguradoras

y reaseguradoras y Reglamento de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras (pendiente de aprobación)).

Se deben constituir y mantener por un importe suficiente para garantizar, atendiendo a criterios prudentes y razonables, todas las obligaciones derivadas de los referidos contratos, así como para mantener la necesaria estabilidad de la entidad aseguradora frente a oscilaciones aleatorias o cíclicas de la siniestralidad o frente a posibles riesgos especiales.

Es relevante enmarcar el estudio de las provisiones técnicas en el contexto de la Directiva Solvencia II, a la que las entidades aseguradoras deben ir adaptándose antes del año 2016. En esta directiva se considera que los principios y las metodologías actuariales y estadísticas correspondientes al cálculo de las citadas provisiones técnicas deben armonizarse en toda la Comunidad, con objeto de lograr una mayor comparabilidad y transparencia.

El cálculo de las provisiones, así como la corrección en la metodología utilizada para el mismo y su adecuación a las bases técnicas de la entidad aseguradora y al comportamiento real de las magnitudes que las definen, son funciones que corresponden a un actuario de seguros.

Según la Directiva Solvencia II, algunas de las funciones actuariales eficaces de las compañías de seguros y reaseguros son:

a) Coordinar el cálculo de las provisiones técnicas.
b) Cerciorarse de la adecuación de las metodologías y los modelos de base utilizados, así como de las hipótesis empleadas en el cálculo de las provisiones técnicas.
c) Evaluar la suficiencia y calidad de los datos utilizados en el cálculo de las provisiones técnicas.
d) Cotejar las mejores estimaciones con la experiencia anterior.
e) Informar al órgano de administración, dirección o supervisión sobre la fiabilidad y la adecuación del cálculo de las provisiones técnicas.

Las provisiones técnicas deben calcularse de forma prudente, fiable y objetiva. El valor de las provisiones técnicas se corresponde con el importe actual que las entidades aseguradoras y reaseguradoras tendrían que pagar si transfirieran sus obligaciones de seguros y reaseguros de manera inmediata a otra entidad aseguradora o reaseguradora. Para calcular las provisiones técnicas, las entidades deben segmentar sus obligaciones de seguro y reaseguro en grupos de riesgo homogéneos y, como mínimo, por líneas de negocio.

5.2 Clases de provisiones técnicas

Las provisiones técnicas son las siguientes:

a) De primas
b) De seguros de vida.
c) De participación en beneficios y para extornos.
d) De siniestros pendientes.
e) Del seguro de decesos.
f) Del seguro de enfermedad.
g) De desviaciones en las operaciones de capitalización por sorteo.

En este capítulo se estudian las provisiones de siniestros pendientes, que deben representar el importe total de las obligaciones pendientes del asegurador derivadas de los siniestros ocurridos con anterioridad a la fecha de cierre del ejercicio e incluye tanto los gastos externos como internos de gestión y tramitación de los expedientes.

La provisión de siniestros pendientes es igual a la diferencia entre su coste total estimado o cierto y el conjunto de importes ya pagados por razón de tales siniestros. Para determinar el importe de la provisión los siniestros se clasifican por años de ocurrencia y su cálculo se realiza, al menos, por ramos de seguros.

Cada siniestro es objeto de valoración individual con independencia que, adicionalmente, la entidad pueda utilizar métodos estadísticos para el cálculo de la provisión de siniestros pendientes.

La provisión de siniestros pendientes está integrada por:

- Provisión de siniestros pendientes de liquidación y pago: Incluye el importe de todos aquellos siniestros ocurridos y declarados antes del cierre del ejercicio. Forman parte de ella los gastos de carácter externos inherentes a la liquidación de siniestros y, en su caso, los intereses de demora y las penalizaciones legalmente establecidas en las que haya incurrido la entidad.
- Provisión de siniestros pendientes de declaración: Debe recoger el importe estimado de los siniestros ocurridos antes del cierre del ejercicio y no declarados en esa fecha.
- Provisión de gastos internos de liquidación de siniestros: Debe dotarse por importe suficiente para afrontar los gastos internos de la entidad necesarios para la total finalización de los siniestros, que han de incluirse en la provisión de siniestros pendientes.

5.3 Métodos estadísticos de cálculo de la provisión de siniestros pendientes

Tal como se indica en la Directiva Solvencia II, el valor de las provisiones técnicas debe ser igual a la suma de la mejor estimación y de un margen de riesgo.

El cálculo de la mejor estimación debe basarse en información actualizada y fiable y en hipótesis realistas y realizase con arreglo métodos actuariales estadísticos que sean adecuados, aplicables y pertinentes.

Según el *Claims Reserving Manual* (Institute and Faculty of Actuaries, 1997), el cálculo de provisiones debe seguir los siguientes pasos:

1. Construir un modelo del proceso, recogiendo las suposiciones hechas.
2. Ajustar el modelo, utilizando observaciones pasadas.
3. Contrastar el ajuste del modelo y las hipótesis, rechazarlo o ajustarlo.
4. Utilizar el modelo para hacer predicciones sobre datos futuros de interés.
5. Aplicar el juicio profesional y la experiencia para elegir un número.

Existen diversos criterios para clasificar los modelos de cálculo de provisiones de siniestros pendientes, este capítulo se centra en la distinción entre métodos deterministas y métodos estocásticos.

Los pagos futuros por siniestros predichos por un modelo de cálculo de provisiones son sucesos aleatorios, ya que no se sabe con certeza cuáles serán estos importes. Lo mejor que puede hacer un modelo es obtener un valor estimado de estos pagos.

Los métodos deterministas sólo hacen suposiciones sobre el valor esperado de los pagos futuros. Los modelos estocásticos también permiten modelizar la variación de los pagos futuros. Proporcionan, por lo tanto, estimaciones no sólo del valor esperado de los pagos futuros sino también de la variación respecto al valor esperado.

En el grupo de los modelos deterministas, se destacan los siguientes (para el detalle de estos métodos y de algunos otros, ver Van Eeghen, 1981):

- Método de Chain-ladder
- Método de mínimos cuadrados de De Vylder
- Método de separación aritmética

Estos métodos se incluyen en el apartado 5.3.1 de este capítulo.
En el grupo de los modelos estocásticos, se analizan:

- Modelo de Mack (Mack, 1993)
- MLG (ver, por ejemplo, England y Verrall, 1999)
- MLGBD

Estos métodos se incluyen en el apartado 5.3.2 de este capítulo.

Para poder calcular las provisiones deben observarse siniestros que se sabe que se han producido pero cuyo importe eventual es desconocido en el momento en que se calculan las provisiones.

Se describen a continuación los datos sobre los siniestros que son necesarios para el cálculo de las provisiones y en qué formato suelen presentarse.

En el cálculo de las provisiones se pueden utilizar diversos tipos de datos:

- Importe de los siniestros: Las cuantías pagadas de los siniestros son, por definición, las cantidades centrales para los objetivos de las provisiones.
- Ingresos de primas: Ofrecen una medida esencial del volumen de negocio a partir del cual se están pagando los siniestros y se indica en referencia al año de origen del negocio.
- Ratio de pérdida: Se puede definir como el cociente entre el importe final del siniestro y la prima total para una clase dada de negocio.
- Número de siniestros: Miden la frecuencia de siniestralidad y cuando se combinan con los datos de cuantías de los siniestros permiten obtener la media del coste total por póliza.

En primer lugar, se considera el año de origen, i, como el año en el que ha ocurrido el siniestro. Por otro lado, el año de desarrollo, j, indica el número de años transcurridos desde el año de origen hasta el año de pago del siniestro. Finalmente, el año de calendario en el que se paga un siniestro, se obtiene a partir de $i+j$.

Si se considera la información disponible de un número determinado de años de experiencia pasada, los datos se presentan en un triángulo (triángulo *run-off*), de manera que los valores a lo largo de una fila muestran el patrón de desarrollo para cada año de ocurrencia, mientras que analizando los datos por columnas se observa el patrón de tendencia desde un año de origen hasta el siguiente. Finalmente, las diagonales permiten analizar la situación en sucesivos años de calendario, correspondiendo la última diagonal al año de calendario más reciente disponible.

Por tanto, se considera una familia de variables aleatorias $\{c_{ij}\}_{i,j\in\{0,1,\ldots k\}}$ donde c_{ij} es el importe de los siniestros del año de origen i que se pagan al cabo de j años desde su ocurrencia, es decir, en el año de desarrollo j y por tanto en el año de calendario $i+j$. Se denomina c_{ij} al

importe de los siniestros ocurridos en el año de origen i que se han pagado en el año de desarrollo j, siendo $i, j = 0,1,...,k$.

Se asume que los pagos de siniestros c_{ij} son observables para los años de calendario $i + j \leq k$ y el triángulo que recoge los datos de experiencias pasadas es el triángulo *run-off* o de desarrollo, que se representa como en la Figura 5.1:

Figura 5.1. Triángulo *run-off* con cuantías no acumuladas.

	Año de desarrollo								
Año de origen	0	1	...	j	…	$k-i$	…	$k-1$	k
0	c_{00}	c_{01}	...	c_{0j}	...	c_{0k-i}	...	c_{0k-1}	c_{0k}
1	c_{10}	c_{11}	...	c_{1j}	...	c_{1k-i}	...	c_{1k-1}	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
i	c_{i0}	c_{i1}	...	c_{ij}	...	c_{ik-i}			
⋮	⋮	⋮	⋮	⋮	⋮				
$k-j$	c_{k-j0}	c_{k-j1}	...	c_{k-jj}					
⋮	⋮	⋮	⋮						
$k-1$	c_{k-10}	c_{k-11}							
k	c_{k0}								

Pueden presentarse también los datos de forma acumulada, sumando los valores a lo largo de cada fila. Para modelizar una cartera con importes acumulados, se considera una familia de variables aleatorias $\{C_{ij}\}_{i,j=0,1,...,k}$ donde C_{ij} es el importe de los siniestros del año de origen i que se han pagado hasta el año de desarrollo j. Se denomina C_{ij} al importe acumulado de los siniestros del año de origen i y año de desarrollo j.

Se asume que los pagos acumulados de siniestros C_{ij} son observables para los años de calendario $i+j \leq k$ y, por lo tanto, el triángulo *run-off* en este caso se representa como en la Figura 5.2:

Figura 5.2. Triángulo *run-off* con cuantías acumuladas.

	Año de desarrollo								
Año de origen	0	1	...	j	…	$k-i$	…	$k-1$	k
0	C_{00}	C_{01}	...	C_{0j}	...	C_{0k-i}	...	C_{0k-1}	C_{0k}
1	C_{10}	C_{11}	...	C_{1j}	...	C_{1k-i}	...	C_{1k-1}	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
i	C_{i0}	C_{i1}	...	C_{ij}	...	C_{ik-i}			
⋮	⋮	⋮	⋮	⋮	⋮				
$k-j$	C_{k-j0}	C_{k-j1}	...	C_{k-jj}					
⋮	⋮	⋮	⋮						
$k-1$	C_{k-10}	C_{k-11}							
k	C_{k0}								

Por supuesto, los importes acumulados C_{ij} se pueden obtener a partir de los importes no acumulados c_{ij} haciendo:

$$C_{ij} = \sum_{h=0}^{j} c_{ih} . \tag{5.1}$$

En estos casos se asume que el patrón de desarrollo dura k años, es decir, que la totalidad de los siniestros se pagan después de k años desde la ocurrencia de los mismos. En algunos casos, los siniestros pueden estar abiertos más de k años, entonces deben hacerse

predicciones por años de desarrollo más allá de k, de los cuales el triángulo *run-off* no ofrece datos.

El objetivo de los métodos de cálculo de provisiones es hacer predicciones sobre los siniestros que serán pagados en años de calendario futuros y, por tanto, completar el triángulo en un cuadrado. Para ello, se pueden aplicar diferentes métodos, que reflejan la influencia de factores externos. En el sentido del año de origen, la variación del tamaño de la cartera tiene influencia sobre los datos de siniestralidad. En cambio, para el patrón de desarrollo deben tenerse en cuenta cambios en los procedimientos de tratamiento de los siniestros o la velocidad de finalización de los mismos. Los datos de las diagonales, que corresponden a los pagos en un año de calendario particular, pueden cambiar debido a la inflación, pero también por cambios en la jurisprudencia o incrementos de la siniestralidad.

Los importes obtenidos en la parte derecha del triángulo que completa el cuadrado constituyen el total de los siniestros que deberán pagarse en el futuro con las primas ingresadas en los años de origen que recoge el triángulo. A partir de estos importes futuros estimados se pueden calcular:

- Las provisiones para los distintos años de origen, P_i, $i \in \{1,...,k\}$, se obtienen sumando los importes de los pagos futuros en la correspondiente fila del cuadrado:

$$P_i = \sum_{j=k-i+1}^{k} \hat{c}_{ij}. \tag{5.2}$$

- Los pagos futuros para los distintos años de calendario, PF_t, $t \in \{k+1,...,2k\}$ se obtienen sumando los importes de los importes que están ubicados en la misma diagonal t:

$$PF_t = \sum_{j=t-k}^{k} \hat{c}_{t-j\,j}. \tag{5.3}$$

En la práctica actuarial es usual tener en cuenta el valor temporal del dinero y se calcula el valor actual de los pagos futuros que se van a realizar en distintos años de calendario. Este objetivo está incluido en el artículo 77.2 de la Directiva Solvencia II.

- La provisión total, P, es decir, la suma de los importes de los pagos futuros, que a su vez coincide con la suma de las provisiones para los distintos años de origen y con la suma de los pagos futuros para los distintos años de calendario:

$$P = \sum_{i=1}^{k} \sum_{j=k-i+1}^{k} \hat{c}_{ij} = \sum_{i=1}^{k} P_i = \sum_{t=k+1}^{2k} PF_t. \tag{5.4}$$

5.3.1 Métodos deterministas

En este apartado se describen algunos métodos estadísticos deterministas que pueden considerarse como clásicos y que se han utilizado en la práctica por su sencillez (ver, por ejemplo, Van Eeghen *et al*., 1981, para sus características más importantes, y Claramunt y Costa, 2003 para su aplicabilidad).

5.3.1.1 Método de Chain-ladder

Este método es un caso de los denominados métodos *link ratio*, que relacionan los pagos realizados hasta un año de desarrollo con los pagos realizados hasta el siguiente año de desarrollo. En el triángulo *run-off* se disponen los importes acumulados pagados por año de origen i hasta el año de desarrollo j, que se indican como C_{ij}.

La idea que hay detrás del método de Chain-ladder es que en cualquier año de desarrollo se paga el mismo porcentaje total de los siniestros de cada año de origen. En otras palabras, en el triángulo *run-off*, las columnas son proporcionales.

En el método de Chain-ladder, como en otros métodos, las estimaciones de los importes futuros se obtienen bajo la hipótesis que existe un patrón de desarrollo. En general, la estimación del patrón de desarrollo puede basarse en distintas fuentes de información:

- Información interna: cualquier información que está completamente incluida en el triángulo *run-off* de la cartera considerada.
- Información externa: cualquier información que es completamente independiente del triángulo *run-off* de la cartera considerada. Se puede obtener, por ejemplo, de estadísticas del mercado o de otras carteras similares a la considerada. También medidas de volumen, como primas o número de contratos de la cartera considerada presentan información externa, ya que no están contenidas en el triángulo *run-off*.
- También pueden combinarse diferentes fuentes de información y, en ese caso, la estimación se basa en información mixta.

En el método de Chain-ladder se utiliza solamente información interna y lo único que se puede captar es el patrón de desarrollo, dado que todos los otros factores que tienen influencia sobre la proporción de siniestros pagados permanecen invariables a lo largo del tiempo.

Se dice que un vector $m = (m_1, ..., m_k)$ de parámetros es un patrón de desarrollo para factores si la relación:

$$m_j = \frac{E\left[C_{ij}\right]}{E\left[C_{ij-1}\right]},$$

se cumple para todo $j \in \{1, ...k\}$ y para todo $i \in \{0, 1, ...k\}$. Entonces, un patrón de desarrollo para factores existe si, y solo si, para cada año de desarrollo $j \in \{1, ...k\}$ los factores individuales:

$$m_{ij} = \frac{E\left[C_{ij}\right]}{E\left[C_{ij-1}\right]},$$

son idénticos para todos los años de origen.

En el caso de un triángulo *run-off* de importes pagados es razonablemente usual asumir además que se cumple que $m_j > 1$ para todo $j \in \{1, ...k\}$.

Se estima el patrón de desarrollo basándose en los factores individuales empíricos. Se asume que $m=(m_1,....,m_k)$ es un patrón de desarrollo para factores. Entonces, para cada año de desarrollo $j\in\{1,...k\}$, cada uno de los factores individuales:

$$\hat{m}_{ij}=\frac{C_{ij}}{C_{ij-1}},$$

con $i\in\{0,1,...,k-j\}$ es un estimador razonable de m_j y esto es cierto también para cada media ponderada:

$$\hat{m}_j=\sum_{h=0}^{k-j} w_{hj}\hat{m}_{hj},$$

con variables aleatorias (o constantes) que satisfagan $\sum_{h=0}^{k-j} w_{hj}=1$.

El estimador que se usa en el método Chain-ladder es:

$$\hat{m}_j=\frac{\sum_{h=0}^{k-j} C_{hj}}{\sum_{h=0}^{k-j} C_{hj-1}}=\sum_{h=0}^{k-j}\frac{C_{hj-1}}{\sum_{n=0}^{k-j} C_{nj-1}}\hat{m}_{hj}. \tag{5.5}$$

Una vez conocidos los valores de $\hat{m}_j$ para $j=1,2,...,k$ pueden estimarse los pagos futuros que completan el triángulo hasta obtener el cuadrado. Los predictores de Chain-ladder de los importes acumulados C_{ij} con $i+j>k$ se definen como:

$$\hat{C}_{ij}=C_{ik-i}\prod_{h=k-i+1}^{j}\hat{m}_h. \tag{5.6}$$

5.3.1.2 Método de los mínimos cuadrados de De Vylder

El método de los mínimos cuadrados de De Vylder (ver De Vylder, 1978) parte de los importes de los siniestros ocurridos en el año de origen i que se pagan en el año de desarrollo j, es decir, las cuantías c_{ij}, y se asume que estos importes dependen de dos efectos:

- Primero, un parámetro que caracteriza el año de origen, proporcional al tamaño de la cartera en ese año.
- Segundo, un parámetro que determina qué proporción de los siniestros se paga en el año de desarrollo.

De esta manera, los importes c_{ij} pueden aproximarse por los productos $x_i p_j$, donde:

- x_i es el importe total de los siniestros del año de origen i
- p_j es la proporción fija del importe x_i que se paga en el año de desarrollo j.

Se impone, además, la condición que $\sum_{j=0}^{k} p_j = 1$.

Los valores de los estimadores $\hat{x}_i, \hat{p}_j$ se obtienen minimizando:

$$\sum_{\substack{i,j \\ i+j\le k}} \left(c_{ij} - \hat{x}_i \hat{p}_j\right)^2,$$

es decir, por mínimos cuadrados.

Las soluciones se obtienen solucionando los sistemas de ecuaciones:

$$\begin{cases} \hat{x}_i = \dfrac{\sum_j c_{ij}\hat{p}_j}{\sum_j \hat{p}_j^2} \\ \hat{p}_j = \dfrac{\sum_i c_{ij}\hat{x}_i}{\sum_i \hat{x}_i^2} \end{cases}. \tag{5.7}$$

5.3.1.3 Método de separación aritmética

En los modelos de separación se asume que en cada año de desarrollo se paga un porcentaje fijo y que hay efectos adicionales que operan en el sentido de las diagonales en el triángulo *run-off*. Así, este modelo describe mejor la situación en la que hay inflación en los datos de siniestros o cuando el riesgo se incrementa por otras causas. Este incremento se representa por

un factor índice para cada año de calendario, que es constante para las observaciones paralelas a la diagonal. Se supone que las variables aleatorias s_{ij} son datos de cuantías medias, de manera que la cuantía total c_{ij} se divide por el número de siniestros del año de origen i, n_i. Así, para el año de origen i y año de desarrollo j, se tiene que:

$$s_{ij} = \frac{c_{ij}}{n_i}. \tag{5.8}$$

Se hace la hipótesis que:

$$s_{ij} = r_j \lambda_t \quad \text{para } t = i + j,$$

donde:

- r_j representa la proporción fija del importe que se paga en el año de desarrollo j
- λ_t recoge el efecto de la inflación del año de calendario $t = i + j$.

Si se expresa el triángulo *run-off* con cuantías medias a partir de los productos de $r_j \lambda_t$, se tienen las expresiones plasmadas en la Figura 5.3.

El método de separación aritmética fue descrito en Verbeek (1972) y se asume que $\sum_{j=0}^{k} r_j = 1$, es decir, que los siniestros se pagan totalmente después de k años de desarrollo.

Para obtener los estimadores de los parámetros $\hat{r}_j$ y $\hat{\lambda}_t$, se suman los importes de las cuantías medias del triángulo, s_{ij}, por columnas y por diagonales, de manera que:

- d_t es la suma de la diagonal t: $d_t = \sum_{i=0}^{t} s_{i\,t-i}$,
- v_j es la suma de la columna j: $v_j = \sum_{i=1}^{n-j+1} s_{ij}$.

A partir de estas sumas por diagonales y por columnas se pueden hallar los estimadores de los parámetros del modelo, $\hat{r}_j, \hat{\lambda}_t$, aplicando las siguientes expresiones:

$$\hat{r}_j = \frac{v_j}{\sum_{t=j}^{k} \hat{\lambda}_k} \qquad j = 0,1,...,k, \tag{5.9}$$

$$\hat{\lambda}_t = \frac{d_t}{1 - \sum_{j=t+1}^{k} \hat{r}_j} \qquad t = 0,1,2,...,k. \tag{5.10}$$

Conocidos los estimadores $\hat{r}_j$ para $j = 0,1,...,k$ y $\hat{\lambda}_t$ para $t = 0,1,...,k$, se pueden estimar las cuantías medias $\hat{s}_{ij} = \hat{r}_j \hat{\lambda}_t$ para $t = i + j \le k$, es decir, los datos del triángulo *run-off*.

Figura 5.3. Triángulo de desarrollo con cuantías medias no acumuladas.

	Año de desarrollo								
Año de origen	0	1	...	j	…	$k-i$	…	$k-1$	k
0	$r_0\lambda_0$	$r_1\lambda_1$	...	$r_j\lambda_j$	...	$r_{k-i}\lambda_{k-i}$	...	$r_{k-1}\lambda_{k-1}$	$r_k\lambda_k$
1	$r_0\lambda_1$	$r_1\lambda_2$	...	$r_j\lambda_{j+1}$	...	$r_{k-i}\lambda_{k-i+1}$	...	$r_{k-1}\lambda_k$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
i	$r_0\lambda_i$	$r_1\lambda_{i+1}$	...	$r_j\lambda_{j+i}$	...	$r_{k-i}\lambda_k$			
⋮	⋮	⋮	⋮	⋮	⋮				
$k-j$	$r_0\lambda_{k-j}$	$r_1\lambda_{k-j+1}$	...	$r_j\lambda_k$					
⋮	⋮	⋮	⋮						
$k-1$	$r_0\lambda_{k-1}$	$r_1\lambda_k$							
k	$r_0\lambda_k$								

Para completar el cuadrado se necesitan los valores de los parámetros $\hat{\lambda}_{k+1}, \hat{\lambda}_{k+2}, ..., \hat{\lambda}_{2k}$ para multiplicarlos por el correspondiente estimador $\hat{r}_j$. Para calcular los valores para estos parámetros se extendiendo la secuencia $\hat{\lambda}_1, ..., \hat{\lambda}_k$. Esto puede hacerse de varias maneras, por ejemplo, con una extrapolación loglineal, o bien considerando una inflación futura constante.

5.3.1.4 Aplicación práctica

Para ilustrar numéricamente los diferentes métodos deterministas de cálculo de provisiones explicados se usan los datos disponibles en Taylor y Ashe (1983), un triángulo con 55 valores de importes pagados, que han sido utilizados por diferentes autores en sus ilustraciones numéricas dentro de la literatura actuarial, por ejemplo, England y Verrall (1999), England (2002) y Renshaw (1989,1994).

Los datos del triángulo *run-off*, expresados como cuantías no acumuladas, se pueden consultar en Kaas *et al.* (2008) y se recogen en la Figura 5.4.

Se puede obtener, aplicando (5.1) a los datos de la Figura 5.4, el triángulo *run-off* con los importes acumulados calculados en la Figura 5.5.

Figura 5.4. Triángulo *run-off* de cuantías no acumuladas de Taylor y Ashe (1983).

	Año de desarrollo									
Año de origen	0	1	2	3	4	5	6	7	8	9
0	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
1	352118	884021	933894	1183289	445745	320996	527804	266172	425046	
2	290507	1001799	926219	1016654	750816	146923	495992	280405		
3	310608	1108250	776189	1562400	272482	352053	206286			
4	443160	693190	991983	769488	504851	470639				
5	396132	937085	847498	805037	705960					
6	440832	847361	1131398	1063269						
7	359480	1061648	1443370							
8	376686	986608								
9	344014									

Figura 5.5. Triángulo *run-off* de cuantías acumuladas de Taylor y Ashe (1983).

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	357848	1124788	1735330	2218270	2745596	33199994	3466336	3606286	3833515	3901463
1	352118	1236139	2170033	3353322	3799067	4120063	4647867	4914039	5339085	
2	290507	1292306	2218525	3235179	3985995	4132918	4628910	4909315		
3	310608	1418858	2195047	3757447	4029929	4381982	4588268			
4	443160	1136350	2128333	2897821	3402672	3873311				
5	396132	1333217	2180715	2985752	3691712					
6	440832	1288463	2419861	3483130						
7	359480	1421128	2864498							
8	376686	1363294								
9	344014									

En este apartado se muestran los resultados para distintos métodos deterministas que se han estimado haciendo uso del software *R*. En el Anexo 5.1 de este capítulo se incluye el código en *R* para los distintos casos.

Resultados con el método de Chain-ladder

Si se aplica el método de Chain-ladder se obtienen, en primer lugar en la Tabla 5.1, los estimadores de los factores de desarrollo (5.5), es decir, las proporciones entre las columnas del triángulo *run-off* de cuantías acumuladas.

Se completa el triángulo de cuantías acumuladas para obtener el cuadrado aplicando (5.6) y se recogen los resultados en la Figura 5.6.

Para poder calcular las provisiones por años de origen (5.2), la provisión total (5.4) y los pagos futuros por años de calendario (5.3), se desacumulan los importes de las cuantías que se han estimado y se recogen los valores en la Figura 5.7.

Tabla 5.1. Factores de desarrollo estimados
con el método de Chain-ladder.

$\hat{m}_1$	3.490607
$\hat{m}_2$	1.747333
$\hat{m}_3$	1.457413
$\hat{m}_4$	1.173852
$\hat{m}_5$	1.103824
$\hat{m}_6$	1.086269
$\hat{m}_7$	1.053874
$\hat{m}_8$	1.076555
$\hat{m}_9$	1.017725

Figura 5.6. Cuadrado de cuantías acumuladas de Taylor y Ashe (1983)
con el método de Chain-ladder.

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	357848	1124788	1735330	2218270	2745596	33199994	3466336	3606286	3833515	3901463
1	352118	1236139	2170033	3353322	3799067	4120063	4647867	4914039	5339085	5433719
2	290507	1292306	2218525	3235179	3985995	4132918	4628910	4909315	528148	5378826
3	310608	1418858	2195047	3757447	4029929	4381982	4588268	4835458	5205637	5297906
4	443160	1136350	2128333	2897821	3402672	3873311	4207459	4434133	4773589	4858200
5	396132	1333217	2180715	2985752	3691712	4074999	4426546	4665023	5022155	5111171
6	440832	1288463	2419861	3483130	4088678	4513179	4902528	5166649	5562182	5660771
7	359480	1421128	2864498	4174756	4900545	5409337	5875997	6192562	6666635	6784799
8	376686	1363294	2382128	3471744	4075313	4498426	4886502	5149760	5544000	5642266
9	344014	1200818	2098228	3057984	3589620	3962307	4304132	4536015	4883270	4969825

Figura 5.7. Cuadrado de cuantías no acumuladas de Taylor y Ashe (1983) con el método de Chain-ladder.

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	357848	766940	610542	482940	527326	574398	146342	139950	227229	67948
1	352118	884021	933894	1183289	445745	320996	527804	266172	425046	94633.8
2	290507	1001799	926219	1016654	750816	146923	495992	280405	375833.5	93677.8
3	310608	1108250	776189	1562400	272482	352053	206286	247190	370179.3	92268.5
4	443160	693190	991983	769488	504851	470639	334148.1	226674.1	339455.9	84610.6
5	396132	937085	847498	805037	705960	383286.6	351547.5	238477.3	357131.7	89016.3
6	440832	847361	1131398	1063269	605548.1	414501	389349.1	264120.5	395533.7	98588.2
7	359480	1061648	1443370	1310258.2	725788.5	508791.9	466660	316565.5	474072.7	118164.3
8	376686	986608	1018834.1	1089616	603568.6	423113.4	388076.4	263257.2	394240.8	98265.9
9	344014	856803.5	897410.1	959756.3	531635.7	372687	348125.7	231182.4	347255.4	86554.6

Tal como se ha indicado, se pueden obtener los diferentes importes:

- Provisiones por años de origen, Tabla5.2:

Tabla 5.2. Provisiones por años de origen con el método de Chain-ladder.

Año de origen	Provisión
1	94633.8
2	469511.3
3	709637.8
4	984888.6
5	1419459.5
6	2177640.6
7	3920301.0

8	4278972.2
9	4625810.7
Total	18680856

- Pagos futuros por años de calendario, Tabla 5.3:

Tabla 5.3. Pagos futuros por años de calendario con el método de Chain-ladder.

Año de calendario	Pago
10	5226535.8
11	4179394.4
12	3131667.5
13	2127271.9
14	1561878.9
15	1177743.7
16	744287.4
17	445521.3
18	86554.6
Total	18680856

- Provisión total. La provisión total asciende a 18680856.

Resultados con el método de los mínimos cuadrados de De Vylder

Al aplicar el método de los mínimos cuadrados de De Vylder, en primer lugar, se obtienen en las Tablas 5.4 y 5.5 los estimadores de los parámetros $\hat{x}_i$ y $\hat{p}_j$ con la solución del sistema de ecuaciones (5.7).

Tabla 5.4. Estimadores de los parámetros $\hat{x}_i$ con el método de De Vylder.

$\hat{x}_0$	3656852
$\hat{x}_1$	5432728
$\hat{x}_2$	5415000
$\hat{x}_3$	5762511
$\hat{x}_4$	4658091
$\hat{x}_5$	4915940
$\hat{x}_6$	5581835
$\hat{x}_7$	7032206
$\hat{x}_8$	5773044
$\hat{x}_9$	5137471

Tabla 5.5. Estimadores de los parámetros $\hat{p}_j$ con el método de De Vylder.

$\hat{p}_0$	0.06696174
$\hat{p}_1$	0.17022541
$\hat{p}_2$	0.18132334
$\hat{p}_3$	0.19831451
$\hat{p}_4$	0.10449431
$\hat{p}_5$	0.06970886
$\hat{p}_6$	0.06903220
$\hat{p}_7$	0.04814073
$\hat{p}_8$	0.07321787
$\hat{p}_9$	0.01858101

Para obtener el cuadrado con las cuantías no acumuladas, se calculan los productos $\hat{x}_i \cdot \hat{p}_j$. En este caso, los valores originales del triángulo *run-off* también se recalculan y se obtienen los siguientes resultados recogidos en la Figura 5.8:

Figura 5.8. Cuadrado de cuantías no acumuladas de Taylor y Ashe (1983) con el método de mínimos cuadrados de De Vylder.

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	244869.2	622489.1	663072.6	725206.8	382120.2	254915	252440.5	176043.5	267746.9	67948
1	363784.9	924788.3	985080.4	1077388.8	567689.1	378709.3	375033.2	261535.5	397772.8	100945.6
2	362597.8	921770.6	981865.9	1073873	565836.7	377473.5	373809.4	260682.1	396474.8	100616.2
3	385867.8	980925.7	1044877.7	1142789.5	602149.6	401698.1	397798.8	277411.5	421918.8	107073.3
4	311913.9	792925.5	844620.7	923767.1	486744	324710.2	321558.3	224243.9	341055.5	86552
5	329179.9	936817.9	891374.6	974902.2	513687.7	342684.6	339358.2	236656.9	359934.7	91343.1
6	373769.4	950170.1	1012116.9	1106958.8	583269.9	389103.3	385326.3	268713.6	408690.1	103716.1
7	470888.8	1197060.2	12754103.1	1394588.5	734825.5	490207.1	485448.7	338535.5	514883.2	130665.5
8	368573.1	982718.7	1046787.6	1144878.3	603250.2	402432.3	398525.9	277918.5	422690	107269
9	344014	874528.1	931543.4	1018835	536836.5	358127.2	354650.9	247321.6	376154.7	95459.4

Los importes que se calculan a partir de los datos del cuadrado anterior son:

- Provisiones por año de origen de origen, Tabla 5.6:

Tabla 5.6. Provisiones por años de origen con el método de De Vylder.

Año de origen	Provisión
1	100945.4
2	497090.4
3	806402.5

4	973409.8
5	1369978.3
6	2138821.0
7	4089153.1
8	4403751.5
9	4793453.8
Total	19173006

- Pagos futuros por años de calendario, Tabla 5.7:

Tabla 5.7. Pagos futuros por años de calendario con el método de De Vylder.

Año de calendario	Pago
10	5338247.9
11	4286487.1
12	3182404.2
13	2139918.5
14	1595221.9
15	1251167.8
16	800676.3
17	483423.0
18	95459.2
Total	19173006

- Provisión total. La provisión total asciende a 19173006.

Resultados con el método de separación aritmética

Para aplicar el método de separación a los datos de Taylor y Ashe (1983) es necesario disponer del número de siniestros para cada año de origen *i*, n_i, con el fin de poder calcular las cuantías medias s_{ij}. El número de siniestros de Taylor y Ashe (1983) se reproducen en la Tabla 5.8:

Tabla 5.8. Número de siniestros por años de origen de Taylor y Ashe (1983).

Año de origen	Número de siniestros
0	606
1	721
2	697
3	321
4	600
5	552
6	543
7	503
8	435
9	420

El método se aplica a los datos del triángulo *run-off* expresados como cuantías medias (5.8). Los resultados se recogen en la Figura 5.9. De este modo se obtienen, a partir de estas cuantías medias, los estimadores de $\hat{r}_j$ y $\hat{\lambda}_t$, para $j = 0,1,...,9$ y $t = 0,1,...,9$ con las expresiones (5.9) y (5.10), respectivamente. Los resultados de los estimadores se recogen en las Tablas 5.9 y 5.10.

Figura 5.9. Triángulo de cuantías medias no acumuladas de Taylor y Ashe (1983).

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	590.51	1265.58	1007.5	796.93	870.18	947.85	241.49	230.94	374.97	112.13
1	488.37	1226.10	1295.28	1641.18	618.23	455.21	732.04	369.17	589.52	
2	416.80	1437.30	1328.87	1458.61	1077.21	210.79	711.61	402.30		
3	967.63	3452.49	2418.03	4867.30	848.85	1096.74	642.64			
4	738.60	1155.32	1653.31	1282.48	841.42	784.40				
5	717.63	1697.62	1535.32	1458.40	1278.91					
6	811.85	1561.02	2083.61	1958.14						
7	714.67	2210.63	2869.52							
8	865.94	2268.06								
9	819.08									

Tabla 5.9. Estimadores de los parámetros $\hat{r}_j$

con el método de separación aritmética.

$\hat{r}_0$	0.084076930
$\hat{r}_1$	0.207913124
$\hat{r}_2$	0.197691284
$\hat{r}_3$	0.202838020
$\hat{r}_4$	0.092432444
$\hat{r}_5$	0.070194215
$\hat{r}_6$	0.056646655
$\hat{r}_7$	0.034264975
$\hat{r}_8$	0.044379178
$\hat{r}_9$	0.009563176

Tabla 5.10. Estimadores de los parámetros $\hat{\lambda}_t$
con el método de separación aritmética.

$\hat{\lambda}_0$	7023.428
$\hat{\lambda}_1$	6006.890
$\hat{\lambda}_2$	5412.490
$\hat{\lambda}_3$	6493.875
$\hat{\lambda}_4$	10231.596
$\hat{\lambda}_5$	8554.888
$\hat{\lambda}_6$	11838.181
$\hat{\lambda}_7$	7521.868
$\hat{\lambda}_8$	10008.196
$\hat{\lambda}_9$	11724.704

Figura 5.10. Cuadrado de cuantías no acumuladas de Taylor y Ashe (1983)
con el método de separación aritmética.

Año de origen	Año de desarrollo									
	0	1	2	3	4	5	6	7	8	9
0	357848	756840.2	648421.3	798226.1	573113.3	363905.2	406379.6	156188.4	269158.2	67948
1	364134.4	811361.3	925607.2	1496332.2	570130.2	599130.7	307209.9	247252.9	375159.9	82055.1
2	317180.7	941062.8	1409820.1	1209473.8	762679.7	368010.2	395150.8	280017.5	368112	80513.5
3	175261.2	682857.9	542883.8	770795.8	223180	225508.1	213197.1	130895.1	172075.2	37636.3
4	516144.7	1067204.1	1404183.1	915432.5	555049.2	493803.8	404476.6	248333.7	326460.4	71403.5
5	397036.3	1358644.9	820828.3	1120583.5	598226.2	461114	377700.3	231894	304848.7	66676.6
6	540457.6	849195.1	1074343.5	1291371.2	597299.6	460399.8	377115.3	231534.8	304376.5	66573.3
7	318105	1046660.1	1165889.5	1214186.2	561599.1	432881.8	354575.2	217696	286184	62594.2
8	366034.4	1060408.1	1023398.4	1065792.4	492962.3	379976.4	311240.2	191089.9	251207.6	54944.2
9	414026.4	1039200	1002930.4	1044476.5	483103.1	372376.9	305015.4	187268.1	246183.4	53845.3

En este ejemplo numérico, para estimar los valores de los parámetros $\hat{\lambda}_{10}, \hat{\lambda}_{11}, ..., \hat{\lambda}_{18}$, se supone una inflación futura anual constante del 1,5%, de manera que:

$$\hat{\lambda}_{k+t} = \hat{\lambda}_k 1.015^{t-k}.$$

Finalmente, se pueden estimar los importes futuros de las cuantías totales no acumuladas, $\hat{c}_{ij} = n_i \hat{r}_j \hat{\lambda}_t$ para $t = i + j = 10,12,...,18$ y completar el cuadrado. Los resultados se recogen en la Figura 5.10.

A partir de las cuantías no acumuladas estimadas, se pueden calcular:

- Provisiones por años de origen, Tabla 5.11:

Tabla 5.11. Provisiones por años de origen
con el método de separación aritmética.

Año de origen	Provisión
1	82055.1
2	448625.5
3	340606.6
4	1050674.2
5	1442233.6
6	2037299.3
7	3129716.5
8	3770611.4
9	4734399.1
Total	17036221

- Pagos futuros por años de calendario, Tabla 5.12:

Tabla 5.12. Pagos futuros por años de calendario
con el método de separación aritmética.

Año de calendario	Pago
10	5320736.9
11	3969344.4
12	2943426.6
13	1825441.7
14	1272366.2
15	848862.6
16	501069.9
17	301127.6
18	53845.3
Total	17036221

- Provisión total. La provisión total asciende a 17036221.

5.3.2 Métodos estocásticos

En el problema de cálculo de provisiones en seguros no vida es de interés poder modelizar el comportamiento de los siniestros a partir de la información disponible y estimar los parámetros de este modelo óptimamente para construir buenos predictores de las observaciones desconocidas.

En las últimas décadas se han propuesto modelos estocásticos para generar el proceso subyacente en el triángulo *run-off* mediante los cuales se obtienen predicciones óptimas y que permiten, además, estimar el error de predicción cometido con la metodología utilizada para

la estimación de los pagos futuros de la entidad e incluso obtener una distribución predictiva de cuantías futuras a partir de la cual se pueden calcular percentiles y otras medidas estadísticas, como su valor en riesgo a un nivel de confianza fijado.

Uno de estos métodos es el denominado modelo de Mack (Mack, 1993), que tiene como caso particular el método de Chain-ladder (Mack y Venter, 2000), ya que la estimación de las provisiones coincide en ambos métodos. Por lo tanto, el modelo de Mack puede considerarse una generalización estocástica del método de Chain-ladder.

Muchos métodos actuariales tradicionales y usualmente utilizados para completar el triángulo *run-off* se pueden describir a través de un MLG, entre ellos los métodos deterministas antes descritos: Chain-ladder, mínimos cuadadros de De Vylder y separación aritmética.

Estos modelos estocásticos generalizan los métodos clásicos, pues proporcionan la misma estimación de las provisiones, pero además añaden las correspondientes formulaciones sobre errores de predicción y otras medidas que amplían la información que no se obtiene con los métodos deterministas.

5.3.2.1 Modelo de Mack

Uno de los primeros modelos estocásticos que reproduce las estimaciones de Chain-ladder es el modelo de Mack (Mack, 1993), que hace suposiciones limitadas en cuanto a la distribución de los datos subyacentes, especificando simplemente sólo los dos primeros momentos ordinarios.

Se considera una familia de variables aleatorias $\{C_{ij}\}_{i,\, j\in\{0,1,\ldots k\}}$ donde C_{ij} es el importe de los siniestros del año de origen i que se han pagado hasta el año de desarrollo j, es decir, el importe acumulado de los siniestros del año de origen i y año de desarrollo j, siendo $i, j = 0,1,...,k$.

Las hipótesis de este modelo son:

1. $E\left[\frac{C_{ij}}{C_{ij-1}}\middle|C_{ij-1}\right]=m_j \qquad i=0,1,...,k-1 \qquad j=1,...,k$

   o, de forma equivalente,

   $E\left[C_{ij}\middle|C_{ij-1}\right]=m_j\cdot C_{ij-1} \qquad i=0,1,...,k-1 \qquad j=1,...,k\,.$

2. $V\left[\frac{C_{ij}}{C_{ij-1}}\middle|C_{ij-1}\right]=\sigma_j^2 \qquad i=0,1,...,k-1 \qquad j=1,...,k$

   o, de forma equivalente,

   $V\left[C_{ij}\middle|C_{ij-1}\right]=C_{ij-1}\cdot\sigma_j^2 \qquad i=0,1,...,k-1 \qquad j=1,...,k\,.$

3. $\{C_{i1},...,C_{ik}\},\{C_{j1},...,C_{jk}\}$ son variables aleatorias independientes para $i\neq j$.

El modelo proporciona estimadores de los parámetros m_j y σ_j^2**:**

$$\hat{m}_j=\sum_{i=0}^{k-j}\frac{C_{ij-1}}{\sum_{i=0}^{k-j}C_{ij-1}}\cdot\frac{C_{ij}}{C_{ij-1}}=\frac{\sum_{i=0}^{k-j}C_{ij}}{\sum_{i=0}^{k-j}C_{ij-1}} \qquad j=1,...,k. \tag{5.11}$$

El estimador que se obtiene para m_j coincide con el del método de Chain-ladder.

$$\hat{\sigma}_j^2=\frac{1}{k-j+1}\sum_{i=0}^{k-j}C_{ij-1}\left(\frac{C_{ij}}{C_{ij-1}}-\hat{m}_j\right)^2 \qquad j=1,...,k-1. \tag{5.12}$$

Falta un estimador para $j=k$ y se proponen diversos métodos para su estimación:

- Si se espera que la totalidad de los importes se hayan pagado a cabo de *k* años de desarrollo, entonces $\hat{\sigma}_k^2 = 0$.
- Si no se cumple lo anterior deberá extrapolarse el valor $\hat{\sigma}_k^2$ a partir de la secuencia $\left\{\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_{k-1}^2\right\}$, que suele ser una sucesión de valores decrecientes exponencialmente.

Suponiendo que las hipótesis 1., 2. y 3. también se cumplen para los valores futuros de C_{ij} con $i + j > k$, se pueden estimar los importes acumulados mediante la aplicación de la expresión (5.6) descrita anteriormente en el método de Chain-ladder y obtener las provisiones, ya sea por años de origen o total , así como los pagos futuros por años de calendario.

Con este método se puede estimar el error de predicción de las provisiones por años de origen, a partir del error cuadrático medio (mean squared error, MSE):

$$MSE(P_i) = \hat{C}_{i,k}^2 \sum_{s=k+1-i}^{k-1} \frac{\hat{\sigma}_s^2}{\hat{m}_s^2} \left(\frac{1}{\hat{C}_{i,s}} + \frac{1}{\sum_{q=1}^{k-s} C_{q,s}} \right) \quad i = 1, ..., k. \tag{5.13}$$

Se propone también un estimador para el error cuadrático medio de la provisión total:

$$MSE(P) = \sum_{i=1}^{k} \left(SE\left(P_i\right) \right)^2 + \hat{C}_{i,k}^2 \left(\sum_{j=i+1}^{k} \hat{C}_{jk} \right) \sum_{j=k+1-i}^{k-1} \frac{2\hat{\sigma}_k^2}{\hat{m}_k^2} \left(\frac{1}{\sum_{q=1}^{k-j} C_{qj}} \right). \tag{5.14}$$

Haciendo la raíz cuadrada del error cuadrático medio se obtiene el error de predicción, tanto para las provisiones por años de origen como para la provisión total.

5.3.2.2 Modelo lineal generalizado

En el capítulo 2 se ha descrito el MLG y en este apartado se detalla su aplicación en el cálculo de las provisiones en los seguros no vida a partir de los datos de un triángulo *run-off*.

Como ya se ha indicado, las variables c_{ij}, para $i, j = 0,1,...,k$, denotan los datos de siniestralidad para el año de origen i y el año de desarrollo j, lo que significa que los siniestros se pagan en el año de calendario $i + j$.

Se asume un MLG para modelizar las cuantías no acumuladas del triángulo *run-off*. Se asume la familia paramétrica de distribuciones de error (2.5) para la función de varianza (2.10), que depende del parámetro ξ. Esta familia paramétrica tiene como casos particulares: la distribución Normal cuando $\xi = 0$; la distribución Poisson cuando $\xi = 1$; la distribución Gamma cuando $\xi = 2$; y la distribución Inversa Gaussiana cuando $\xi = 3$.

Con estas hipótesis, la media y la varianza del MLG son:

$$\mu_{ij} = E\left[c_{ij}\right] \tag{5.15}$$

$$V\left[c_{ij}\right] = \frac{\phi}{w_{ij}} V\left(\mu_{ij}\right) = \frac{\phi}{w_{ij}} \mu_{ij}^{\xi}, \tag{5.16}$$

donde ϕ es el parámetro de dispersión y w_{ij} son pesos *a priori* de los datos, que se asumen igual a uno, $w_{ij} = 1$, para las cuantías no acumuladas de un triángulo *run-off*.

Se asume para el MLG la función de enlace logarítmica:

$$\log \mu_{ij} = \eta_{ij}. \tag{5.17}$$

De esta manera, se puede definir:

$$\log\left(\mu_{ij}\right) = c_0 + \alpha_i + \beta_j, \tag{5.18}$$

un MLG en el que las respuestas c_{ij} se modelizan como variables aleatorias con función de varianza (5.16), con una función de enlace logarítmica (5.17) y con un predictor lineal:

$$\eta_{ij} = c_0 + \alpha_i + \beta_j, \tag{5.19}$$

donde α_i es el factor correspondiente al año de origen $i = 1,...,k$ y β_j es el factor correspondiente al año de desarrollo $j = 1,...,k$. El valor c_0 es el término correspondiente al año de origen 0 y año de desarrollo 0.

Las predicciones $\hat{c}_{ij}$ necesarias para calcular las provisiones por años de origen (5.2), la provisión total (5.4) y los pagos futuros por años de calendario (5.3) se estiman a partir de:

$$\hat{c}_{ij} = \exp\left(\hat{c}_0 + \hat{\alpha}_i + \hat{\beta}_j\right). \tag{5.20}$$

Algunos métodos deterministas que se han descrito en este capítulo pueden obtenerse a partir de un MLG, como se muestra a continuación.

Caso particular: método de mínimos cuadrados de De Vylder a partir de la distribución Normal.

Si se utiliza el MLG para modelizar los datos c_{ij} del triángulo *run-off* y se considera el caso en que la distribución del error es Normal, junto con la función de enlace logarítmica, se obtiene como caso particular el modelo de mínimos cuadrados de De Vylder, es decir, la estimación de las cuantías futuras coincide en ambos modelos.

En este caso, tenemos que:

$$E\left[c_{ij}\right] = \mu_{ij} \qquad V\left[c_{ij}\right] = \phi = \sigma^2.$$

Caso particular: método de Chain-ladder a partir de la distribución Poisson sobre-dispersa

Si se utiliza el MLG para modelizar los datos c_{ij} del triángulo *run-off* y se considera el caso en que la distribución del error es Poisson sobre-dispersa, junto con la función de enlace canónica logarítmica, se obtiene como caso particular el modelo de Chain-ladder, es decir, la estimación de las cuantías futuras coincide en ambos modelos. Este caso se puede consultar con detalle en Boj y Costa (2015c); Boj *et al.* (2014b); England y Verrall (1999, 2002, 2006); England (2002); Haberman y Renshaw (1996); Verrall(2000); y Verrall y England (2000).

La distribución Poisson sobre-dispersa difiere de la distribución de Poisson en que la varianza no es igual a la media pero, en cambio, es proporcional a la media.

En el modelo Poisson sobre-disperso se cumple que:

$$E\left[c_{ij}\right]=\mu_{ij} \qquad V\left[c_{ij}\right]=\phi\mu_{ij} \qquad \phi>1.$$

La sobre-dispersión se introduce a través del parámetro ϕ, que es desconocido y se estima a partir de los datos disponibles.

Caso particular: Distribución Gamma

En el MLG, si se considera que la distribución del error es una Gamma, se puede deducir que:

$$E\left[c_{ij}\right]=\mu_{ij} \qquad V\left[c_{ij}\right]=\phi\mu_{ij}^2.$$

En este caso, la varianza es proporcional al cuadrado de la media y no proporcional a la media como ocurre en el caso Poisson sobre-disperso.

Si se asume la función de enlace logarítmica (5.17), como en los dos casos particulares anteriores, se pueden estimar las cuantías futuras para completar el cuadrado mediante la expresión (5.20). Este caso se ha estudiado, por ejemplo, en Boj y Costa (2015c) y England y Verrall (1999, 2002).

5.3.2.3 Modelo lineal generalizado basado en distancias

El objetivo de este apartado es proponer el MLGBD como una metodología alternativa para aplicar en el problema de cálculo de provisiones.

Se han analizado algunos métodos estocásticos, entre ellos el MLG, que reproducen los resultados que proporcionan algunos métodos deterministas cuando se asumen determinadas hipótesis sobre la distribución del error y la función de enlace. En concreto, si se considera la distribución de error Poisson sobre-dispersa junto con la función de enlace logarítmica, el MLG proporciona los mismos resultados numéricos que el método de Chain-ladder.

El MLGBD, cuando usa la métrica ℓ^2 denominada Euclídea, reproduce el MLG ordinario y, por tanto, se puede considerar que el MLGBD es una generalización del MLG en el contexto del análisis basado en distancias. Si se usa la métrica Euclídea y se asume la distribución Poisson sobre-dispersa del error, junto con la función de enlace logarítmica, el MLGBD reproduce el modelo determinista de Chain-ladder.

Se propone la aplicación del MLGBD en el problema del cálculo de provisiones asumiendo estas hipótesis, como una alternativa que generaliza el método clásico determinista de Chain-ladder en un marco estocástico.

Con el MLGBD se tienen los mismos casos particulares que con el MLG clásico cuando se usa la métrica Euclídea pero, además, se pueden tener más herramientas de análisis si se considera el uso de otras métricas diferentes de la Euclídea.

5.3.2.4 Errores de predicción en el modelo lineal generalizado

El objetivo del cálculo de provisiones es hacer una predicción de cuánto falta por pagar de los siniestros ocurridos en el pasado.

Los errores de predicción permiten conocer algo más sobre la incertidumbre de los pagos futuros que deben afrontarse en los próximos periodos y, por lo tanto, posibilita añadir a las

provisiones márgenes de riesgo con sentido estadístico. Dichos errores pueden ser calculados a partir de formulaciones analíticas y también haciendo uso de metodología *bootstrap.*

En England y Verrall (1999) y England (2002) se describe un método para obtener estimaciones del error de predicción en el MLG. Basándose en el método delta, dan una aproximación que puede obtenerse a partir de las varianzas y covarianzas de los predictores y valores ajustados.

Si se considera una variable aleatoria c_{ij} y un valor predicho $\hat{c}_{ij}$, el error cuadrático medio de predicción es:

$$MSE(c_{ij}) = E\left[\left(c_{ij} - \hat{c}_{ij}\right)^2\right] = E\left[\left(\left(c_{ij} - E\left[c_{ij}\right]\right) - \left(\hat{c}_{ij} - E\left[c_{ij}\right]\right)\right)^2\right]. \qquad (5.21)$$

Haciendo la aproximación $E\left[c_{ij}\right] \approx E\left[\hat{c}_{ij}\right]$, y teniendo en cuenta que las cuantías pasadas y predichas son variables aleatorias independientes, y por lo tanto la varianza de su diferencia es exactamente la suma de sus varianzas, se puede aproximar el error como:

$$\begin{aligned} MSE(c_{ij}) &= E\left[\left(c_{ij} - \hat{c}_{ij}\right)^2\right] \approx E\left[\left(c_{ij} - E\left[c_{ij}\right]\right)^2\right] \\ &-2E\left[\left(c_{ij} - E\left[c_{ij}\right]\right)\left(\hat{c}_{ij} - E\left[\hat{c}_{ij}\right]\right)\right] + E\left[\left(\hat{c}_{ij} - E\left[\hat{c}_{ij}\right]\right)^2\right] \end{aligned}$$

$$MSE(\hat{c}_{ij}) = E\left[\left(c_{ij} - \hat{c}_{ij}\right)^2\right] \approx E\left[\left(c_{ij} - E\left[c_{ij}\right]\right)^2\right] + E\left[\left(\hat{c}_{ij} - E\left[\hat{c}_{ij}\right]\right)^2\right]. \qquad (5.22)$$

El error cuadrático medio de la predicción (insesgada) $\hat{c}_{ij}$ puede descomponerse, aproximadamente, en una parte de varianza de estimación $E\left[\left(c_{ij} - E\left[c_{ij}\right]\right)^2\right] = V\left[c_{ij}\right]$ y otra parte de varianza del proceso $E\left[\left(\hat{c}_{ij} - E\left[\hat{c}_{ij}\right]\right)^2\right] = V\left[\hat{c}_{ij}\right]$. Por tanto:

$$MSE\left(c_{ij}\right) = E\left[\left(c_{ij} - \hat{c}_{ij}\right)^2\right] \approx V\left[c_{ij}\right] + V\left[\hat{c}_{ij}\right]. \qquad (5.23)$$

Utilizando el método delta, se deriva que:

$$V\left[c_{ij}\right] \approx \left|\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right|^2 V\left[\eta_{ij}\right]$$

Por tanto, el error cuadrático medio de predicción del pago futuro $\hat{c}_{ij}$ se puede aproximar por:

$$MSE\left(c_{ij}\right) = E\left[\left(c_{ij} - \hat{c}_{ij}\right)^2\right] \approx \frac{\phi}{w_{ij}} \mu_i^{\xi} + \left|\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right|^2 V\left[\eta_{ij}\right]. \tag{5.24}$$

Para la varianza de la estimación, cabe notar que $\mu_{ij} = e^{\eta_{ij}}$ en el caso de función de enlace logarítmica y, por lo tanto, $\frac{\partial \mu_{ij}}{\partial \eta_{ij}} = e^{\eta_{ij}} = \mu_{ij}$. Estas dos últimas relaciones conducen a la siguiente aproximación para el error cuadrático medio de predicción del pago futuro $\hat{c}_{ij}$:

$$MSE\left(c_{ij}\right) \cong \frac{\phi}{w_{ij}} \mu_{ij}^{\xi} + \mu_{ij}^2 V\left[\eta_{ij}\right]. \tag{5.25}$$

cuando se asume la familia paramétrica de distribuciones (2.5).

También se pueden calcular los errores cuadráticos medios de predicción para las provisiones por años de origen (5.2), para la provisión total (5.4) y para los pagos futuros por años de calendario (5.3). En estos casos se obtienen expresiones en las que aparecen las covarianzas estimadas de los diferentes predictores lineales, ya que si $\hat{c}_{ij}$ y $\hat{c}_{kl}$ son diferentes estimadores de pagos futuros, entonces:

$$Cov\left[\hat{c}_{ij}, \hat{c}_{kl}\right] \approx \hat{\mu}_{ij} \hat{\mu}_{kl} Cov\left[\hat{\eta}_{ij}, \hat{\eta}_{kl}\right]$$

Los errores cuadráticos medios para las provisiones por años de origen son:

$$MSE\left(P_i\right)=E\left[\left(P_i-\hat{P}_i\right)^2\right]\approx\sum_{\substack{j=1,..,k\\ i+j>k}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\mu_i^T V\left[\eta_i\right]\mu_i=$$

$$=\sum_{\substack{j=1,..,k\\ i+j>k}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\sum_{\substack{j=1,..,k\\ i+j>k}}\mu_{ij}{}^2V\left[\eta_{ij}\right]+2\sum_{\substack{j_1,j_2=1,..,k\\ j_2>j_1\\ i+j_1>k,i+j_2>k}}\mu_{ij_1}\mu_{ij_2}Cov\left[\eta_{ij_1},\eta_{ij_2}\right]. \qquad (5.26)$$

$$i=1,...,k$$

Los errores cuadráticos medios para los pagos futuros por años de calendario son:

$$MSE\left(PF_t\right)=E\left[\left(PF_t-\widehat{PF_t}\right)^2\right]\approx\sum_{\substack{i,j=1,..,k\\ i+j=t}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\mu_t^T V\left[\eta_t\right]\mu_t=$$

$$=\sum_{\substack{i,j=1,..,k\\ i+j=t}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\sum_{\substack{i,j=1,..,k\\ i+j=t}}\mu_{ij}{}^2V\left[\eta_{ij}\right]+2\sum_{\substack{i_1,i_2,j_1,j_2=1,..,k\\ i_1j_1\neq i_2j_2\\ i_1+j_1=t,i_2+j_2=t}}\mu_{i_1j_1}\mu_{i_2j_2}Cov\left[\eta_{i_1j_1},\eta_{i_2j_2}\right]. \qquad (5.27)$$

$$t=k+1,...,2k$$

Por último, el error cuadrático medio para la provisión total es:

$$MSE\left(P\right)=E\left[\left(P-\hat{P}\right)^2\right]\approx\sum_{\substack{i,j=1,...,k\\ i+j>k}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\mu^T V\left[\eta\right]\mu=$$

$$=\sum_{\substack{i,j=1,...,k\\ i+j>k}}\frac{\phi}{w_{ij}}\mu_{ij}^{\xi}+\sum_{\substack{i,j=1,...,k\\ i+j>k}}\mu_{ij}{}^2\left[\eta_{ij}\right]+2\sum_{\substack{i_1,j_1,i_2,j_2=1,...,k\\ i_1+j_1>k,i_2+j_2>k\\ i_1j_1\neq i_2j_2}}\mu_{i_1j_1}\mu_{i_2j_2}Cov\left[\eta_{i_1j_1},\eta_{i_2j_2}\right]. \qquad (5.28)$$

Hay que destacar que en England y Verrall (1999) y England (2002) se obtiene la estimación del error de predicción para las provisiones por años de origen (5.26) y para la provisión total (5.28) en el MLG. En este trabajo se ha obtenido la estimación del error de predicción para los pagos futuros por años de calendario (5.27) en el MLG, que también se puede encontrar en Boj y Costa (2015c).

Finalmente, el error de predicción (PE) para cada predicción $\hat{c}_{ij}$**,** para las provisiones por años de origen, para los pagos futuros por año de calendario y para la provisión total**,** se

pueden calcular como la raíz cuadrada de los errores cuadráticos medios (5.25), (5.26), (5.27) y (5.28), respectivamente.

Para calcular los errores de predicción se puede usar también la metodología *bootstrapping residuals* descrita en el capítulo 2 de este trabajo, que permite crear estimadores por muestreo (con reemplazo) de los residuos observados en las observaciones pasadas para obtener un gran conjunto de pseudo-datos. Concretamente, en el caso del MLG y el cálculo de provisiones es usual aplicar *bootstrapping residuals* basado en los residuos de Pearson, como se hace en Boj *et al.* (2014b), Boj y Costa (2015c), England y Verrall (1999) , England (2002) y Pinheiro *et al.* (2003).

Cuando se asume (5.16), los residuos de Pearson (2.20) siguen la siguiente expresión:

$$r_{ij}^{P} = \frac{c_{ij} - \hat{\mu}_{ij}}{\sqrt{\frac{\hat{\mu}_{ij}^{\xi}}{w_{ij}}}}, \tag{5.29}$$

por lo tanto $c_{ij} = r_{ij}^{P}\sqrt{\frac{\hat{\mu}_{ij}^{\xi}}{w_{ij}}} + \hat{\mu}_{ij}$.

Los residuos de Pearson son no escalados en el sentido que no incluyen el parámetro de escala ϕ. Para estimar el parámetro de escala (2.23), en este caso se aplica:

$$\hat{\phi}^{P} = \frac{1}{n-2k-1}\sum_{\substack{i,j=1,\ldots,k \\ i+j\leq k}} w_{ij}\frac{\left(c_{ij} - \hat{\mu}_{ij}\right)}{\hat{\mu}_{ij}^{\xi}}, \tag{5.30}$$

donde $n = \frac{k\left(k+1\right)}{2}$ es el número de observaciones pasadas y $p = 2k+1$ es el número de parámetros estimados . Se utiliza $n-p$ en el denominador en lugar de n para reducir el sesgo.

En el proceso de *bootstrap* se ajustan los residuos por los grados de libertad:

$$r_{ij}^{P'} = \sqrt{\frac{n}{n-2k-1}} r_{ij}^{P}, \tag{5.31}$$

de la misma manera que en el parámetro de escala $\hat{\phi}^P$ para que sea compatible.

El procedimiento que se sigue al aplicar la metodología *bootstrapping residuals* en este caso que se está considerando se resume en los siguientes pasos:

- En primer lugar se aplica el MLG elegido al triángulo *run-off* de los datos de cuantías observadas y se calculan los residuos de Pearson (5.29).
- Se remuestrean *B* veces los residuos ajustados (5.31), con reemplazo, y utilizando este conjunto de residuos $r_{ij}^{P'*}$ y los valores estimados de $\hat{\mu}_{ij}$ se crean *B* nuevas muestras de triángulos *run-off* de cuantías no acumuladas aplicando la expresión:

$$c_{ij}^{*} = r_{ij}^{P'*} \sqrt{\hat{\mu}_{ij}^{\xi}} + \hat{\mu}_{ij}.$$

- Se estima el MLG elegido a cada una de las *B* muestras y se calculan las cuantías correspondientes.

Los *B* valores de las estimaciones proporcionan la distribución predictiva de una cuantía $\hat{c}_{ij}^{boot}$ (para cada valor), de las provisiones por años de origen $\hat{P}_i^{boot}$ (para cada año de origen), de la provisión total $\hat{P}^{boot}$, y de los pagos futuros por años de calendario $\widehat{PF}_t^{boot}$ (para cada año de calendario).

Además, con la varianza del conjunto de estimadores obtenidos de esta manera se puede calcular un estimador *bootstrap* del error de predicción, que se denota por PE^{boot}.

En concreto, en primer lugar, se deduce la fórmula en el caso en el que $V\left[\hat{c}_{ij}\right]$ de (5.23) se estima mediante la varianza de la distribución predictiva de c_{ij}:

$$PE^{boot}\left(c_{ij}\right) \approx \sqrt{\frac{\hat{\phi}^P}{w_{ij}}\hat{c}_{ij}^{\xi} + V\left[\hat{c}_{ij}^{boot}\right]}, \quad i, j = 1, ..., k. \tag{5.32}$$

Teniendo todo esto en cuenta, a continuación se deducen las fórmulas de las estimaciones *bootstrap* del error de predicción para las provisiones por años de origen, para los pagos futuros por años de calendario y para la provisión total.

Las estimaciones *bootstrap* del error de predicción para las provisiones por año de origen son:

$$PE^{boot}\left(P_i\right) \approx \sqrt{\sum_{\substack{j=1,...,k \\ i+j>k}} \frac{\hat{\phi}^P}{w_{ij}}\hat{c}_{ij}^{\xi} + V\left[\hat{P}_i^{boot}\right]}, \quad i = 1, \ldots, k. \tag{5.33}$$

Las estimaciones *bootstrap* del error de predicción para los pagos futuros por años de calendario son:

$$PE^{boot}(PF_t) \approx \sqrt{\sum_{\substack{i,\, j=1,..,k \\ i+j=t}} \frac{\hat{\phi}^P}{w_{ij}}\hat{c}_{ij}^{\xi} + V\left[\widehat{PF}_t^{boot}\right]}, \quad t = k+1, ..., 2k. \tag{5.34}$$

La estimación *bootstrap* del error para la provisión total es:

$$PE^{boot}\left(P\right) \approx \sqrt{\sum_{\substack{i,\, j=1,...,k \\ i+j>k}} \frac{\hat{\phi}^P}{w_{ij}}\hat{c}_{ij}^{\xi} + V\left[\hat{P}^{boot}\right]}. \tag{5.35}$$

Para poder comparar los resultados con el error de predicción definido de forma analítica para el MLG no ha sido necesario hacer un ajuste que tenga en cuenta los grados de libertad, porque se asume que en el proceso de *bootstraping* se están utilizando los residuos de Pearson ajustados (5.31). Si se utilizan directamente los residuos de Pearson (5.29) es necesario corregir el error estándar multiplicando por $\frac{n}{n-p}$, donde $p = 2k+1$ es el número de parámetros del modelo.

Además, hay que añadir un estimador de la varianza del proceso según la distribución supuesta en (2.5).

Con esta metodología se puede estimar la distribución predictiva de los importes individuales $\hat{c}_{ij}$, de las provisiones por años de origen (5.2), de los pagos futuros por años de calendario (5.3) y de la provisión total (5.4). A partir de estas distribuciones predictivas se pueden calcular estadísticos como la media, la desviación estándar, cuantiles (como puede ser el valor en riesgo), asimetría, curtosis, realizar histogramas, etc.

En el MLGBD, a diferencia del MLG clásico, la metodología *bootstrap* que se aplica es *bootstrapping pairs*.

5.3.2.5 Aplicación práctica

Para ilustrar los distintos métodos estocásticos de cálculo de provisiones se usan los mismos datos de Taylor y Ashe (1983) que ya han sido utilizados en los métodos deterministas.

En estos métodos, además de obtener el importe de las provisiones y de los pagos futuros por años de calendario, también es posible calcular los errores de estimación que se hayan cometido.

Los resultados se han obtenido haciendo uso del software *R*. En concreto, para obtener los resultados del modelo de Mack se ha usado la función *MackChainLadder* de la librería *Chain-Ladder* de *R* (Gesmann *et al.*, 2015) y para ajustar los MLG se ha usado la función *glm* de la librería *stats* de *R*. En el caso de aplicarse el MLGBD se hace uso de la función *dbglm* de la librería *dbstats* de Boj *et al.* (2014a).

Resultados con el modelo de Mack

Las instrucciones para ajustar el modelo de Mack son las siguientes:

```
R> library(ChainLadder)
R> cij<- c(357848, 766940, 610542, 482940, 527326, 574398, 146342, 139950, 227229, 67948, 352118, 884021, 933894, 1183289, 445745, 320996, 527804, 266172)
R> cij<-c(cij, 425046, 290507,1001799, 926219, 1016654, 750816, 146923, 495992, 280405, 310608, 1108250, 776189, 1562400, 272482, 352053, 206286, 443160, 693190, 991983, 769488, 504851)
R> cij<-c(cij, 470639, 396132, 937085, 847498, 805037, 705960, 440832, 847631, 1131398, 1063269, 359480, 1061648, 1443370, 376686, 986608, 344014)
R> n<-length(cij);k<-trunc(sqrt(2*n))
R> i<-rep(1:k,k:1);i<-as.factor(i)
R> j<-sequence(k:1);j<-as.factor(j)
R> cij.l<-xtabs(cij~i+j)
R> cij.v<-as.vector(cij.l)
R> ii<-row(cij.l);jj<-col(cij.l)
R> future<-as.numeric(ii+jj-1>k)
R> cij.v.2<-cij.v
R> cij.v.2[future==1]<-NA
R> C<-matrix(cij.v.2,k,k)
R> C<-t(apply(C,1,cumsum))
R> C<-as.triangle(C)
R> MackChainLadder(C)

MackChainLadder(Triangle = C)
```

	Latest	Dev.To.Date	Ultimate	IBNR	Mack.S.E	CV(IBNR)
1	3,901,463	1.0000	3,901,463	0	0	NaN
2	5,339,085	0.9826	5,433,719	94,634	71,835	0.759
3	4,909,315	0.9127	5,378,826	469,511	119,474	0.254
4	4,588,268	0.8661	5,297,906	709,638	131,573	0.185
5	3,873,311	0.7973	4,858,200	984,889	260,530	0.265
6	3,691,712	0.7223	5,111,171	1,419,459	410,407	0.289
7	3,483,130	0.6153	5,660,771	2,177,641	557,796	0.256
8	2,864,498	0.4222	6,784,799	3,920,301	874,882	0.223

```
9    1,363,294     0.2416  5,642,266  4,278,972    970,960    0.227
10    344,014      0.0692  4,969,825  4,625,811  1,362,981    0.295

              Totals
Latest:   34,358,090.00
Dev:              0.65
Ultimate: 53,038,945.61
IBNR:     18,680,855.61
Mack.S.E   2,441,364.13
CV(IBNR):         0.13
```

Los estimadores de los parámetros $\hat{m}_j$ que se obtienen son los mismos que los que se han obtenido con el método de Chain-ladder y se recogen a continuación en la Tabla 5.13:

Tabla 5.13. Estimadores de los parámetros $\hat{m}_j$ con el modelo de Mack.

$\hat{m}_1$	3.490607
$\hat{m}_2$	1.747333
$\hat{m}_3$	1.457413
$\hat{m}_4$	1.173852
$\hat{m}_5$	1.103824
$\hat{m}_6$	1.086269
$\hat{m}_7$	1.053874
$\hat{m}_8$	1.076555
$\hat{m}_9$	1.017725

Los estimadores de los parámetros $\hat{\sigma}_j^2$ se muestran en la Tabla 5.14:

Tabla 5.14. Estimadores de los parámetros $\hat{\sigma}_j^2$ con el modelo de Mack.

$\hat{\sigma}_1^2$	160280.3275
$\hat{\sigma}_2^2$	37736.8550
$\hat{\sigma}_3^2$	41965.2130
$\hat{\sigma}_4^2$	15182.9027
$\hat{\sigma}_5^2$	13731.3239
$\hat{\sigma}_6^2$	8185.7716
$\hat{\sigma}_7^2$	446.6166
$\hat{\sigma}_8^2$	1147.3660
$\hat{\sigma}_9^2$	403.9358

Los valores estimados de los importes futuros, $\hat{c}_{ij}$ con $i+j>k$ coinciden con los obtenidos por Chain-ladder y, en consecuencia, las provisiones por años de origen, los pagos futuros por años de calendario y la provisión total van a ser iguales.

Sin embargo, al tratarse de un modelo estocástico, se pueden estimar los errores de predicción que se han cometido, tanto para las provisiones por años de origen como para la provisión total, haciendo la raíz cuadrada de (5.13) y (5.14). Estos resultados los proporciona la función *MackChainLadder* de la librería *Chain-Ladder* de *R* y, además, expresa estos errores como porcentaje de la provisión correspondiente, en lo que se denomina coeficiente de variación. Los valores obtenidos se muestran en la Tabla 5.15.

Se desprende de los valores obtenidos, por ejemplo, que para el año de origen 1 el error de predicción es del 76% de la provisión de ese año; en la provisión total, el error de predicción cometido en su estimación asciende al 13% de su importe. Estos valores resultan útiles para poder comparar entre distintos métodos estocásticos de cálculo de provisiones.

Tabla 5.15. Errores de predicción para las provisiones por años de origen y para la provisión total y coeficientes de variación con el modelo de Mack.

Año de origen	Error de predicccción	Coeficiente de variación
1	71835	75.91%
2	119474	25.45%
3	131573	18.54%
4	260530	26.45%
5	410407	28.91%
6	557796	25.61%
7	874882	22.32%
8	970960	22.69%
9	1362981	29.46%
Total	2441364.1	13.07%

Resultados con MLG con distribución Normal y función de enlace logarítmica

La instrucción para ajustar el MLG con distribución Normal y función de enlace logarítmica es la siguiente:

```
R> glm(formula = cij ~ i + j, family = gaussian(link = "log"))

Call:  glm(formula = cij ~ i + j, family = gaussian(link = "log"))
Coefficients:
(Intercept)          i2          i3          i4          i5          i6
   12.40848     0.39584     0.39257     0.45477     0.24200     0.29588
         i7          i8          i9         i10          j2          j3
    0.42291     0.65390     0.45660     0.33996     0.93300     0.99616
    j4          j5          j6          j7          j8          j9
```

```
  1.08573     0.44501     0.04021     0.03045    -0.32999     0.08932
      j10
 -1.28198

Degrees of Freedom: 54 Total (i.e. Null);  36 Residual
Null Deviance:      6.598e+12
Residual Deviance: 1.096e+12    AIC: 1500
```

Los valores de los parámetros asociados al predictor lineal del MLG que se han estimado se plasman en la Tabla 5.16.

Tabla 5.16. Estimadores de los parámetros del predictor lineal para MLG con distribución Normal y función de enlace logarítmica.

$\hat{c}_0$	12.40848
$\hat{\alpha}_1$	0.39584
$\hat{\alpha}_2$	0.39257
$\hat{\alpha}_3$	0.45477
$\hat{\alpha}_4$	0.242
$\hat{\alpha}_5$	0.29588
$\hat{\alpha}_6$	0.42291
$\hat{\alpha}_7$	0.6539
$\hat{\alpha}_8$	0.4566
$\hat{\alpha}_9$	0.33996
$\hat{\beta}_1$	0.933
$\hat{\beta}_2$	0.99616
$\hat{\beta}_3$	1.08573
$\hat{\beta}_4$	0.44501
$\hat{\beta}_5$	0.04021
$\hat{\beta}_6$	0.03045
$\hat{\beta}_7$	-0.32999

$\hat{\beta}_8$	0.08932
$\hat{\beta}_9$	-1.28198

Para poder completar el triángulo *run-off* hasta obtener el cuadrado se aplica la expresión (5.20). Por ejemplo, para el año de origen $i = 6$ y el año de desarrollo $j = 9$, se obtiene:

$$\hat{c}_{69} = \exp\left(\hat{c}_0 + \hat{\alpha}_6 + \hat{\beta}_9\right) = \exp(12.40848 + 0.42291 - 1.28198) = 103716\,,$$

que coincide con el resultado ya obtenido en el método de mínimos cuadrados de De Vylder. El importe de las provisiones por años de origen, los pagos futuros por años de calendario y la provisión total también coinciden.

Tabla 5.17. Provisiones por año de origen y provisión total, errores de predicción y coeficientes de variación para MLG con distribución Normal y función de enlace logarítmica usando fórmula analítica.

Año de origen	Provisión	Error de predicción	Coeficiente de variación
1	100945.4	1763993	1747.17%
2	497090.4	1770671	16.99 %
3	806402.5	1779059	20.57 %
4	973409.8	1771605	22.52 %
5	1369978.3	1779036	25.93 %
6	2138821.0	1796344	30.93 %
7	4089153.1	1860366	39.56 %
8	4403751.5	1951428	56.34 %
9	4793453.8	3111097	64.90%
Total	19173006	4399697	22.95%

Para el caso del MLG se pueden obtener de los errores de predicción de forma analítica para

las provisiones por años de origen y para la provisión total a partir de la raíz cuadrada de los errores cuadráticos medios (5.26) y (5.28), respectivamente. En la Tabla 5.17 se muestran estos resultados junto con los coeficientes de variación, que expresan el porcentaje que supone este error con respecto al importe de la provisión correspondiente.

De manera análoga se calcula, en la Tabla 5.18, el error de predicción para los pagos futuros por años de calendario (5.27) y los coeficientes de variación, que expresan el porcentaje que supone este error con respecto al importe del pago correspondiente.

Tabla 5.18. Pagos futuros por años de calendario, errores de predicción y coeficientes de variación para MLG con distribución Normal y función de enlace logarítmica usando fórmula analítica.

Año de calendario	Pago	Error de predicción	Coeficiente de variación
10	5338247.9	1861081	34.86%
11	4286487.1	1869066	43.60%
12	3182404.2	1872331	58.83%
13	2139918.5	1803973	84.30%
14	1595221.9	1791901	112.33%
15	1251167.8	1794356	143.41%
16	800676.3	1794002	224.06%
17	483423.0	1783502	368.93%
18	95459.2	1762682	1846.53%

Resultados con MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica

La instrucción para ajustar el MLG con distribución Poisson-sobredispersa y función de enlace logarítmica es la siguiente:

```
R> glm(formula = cij ~ i + j, family = "quasipoisson")

Call:  glm(formula = cij ~ i + j, family = "quasipoisson")

Coefficients:
(Intercept)          i2          i3          i4          i5          i6
  12.506405    0.331272    0.321119    0.305960    0.219316    0.270077
         i7          i8          i9         i10          j2          j3
   0.372208    0.553333    0.368934    0.242033    0.912526    0.958831
         j4          j5          j6          j7          j8          j9
   1.025997    0.435276    0.080057   -0.006381   -0.394452    0.009378
        j10
  -1.379907

Degrees of Freedom: 54 Total (i.e. Null);  36 Residual
Null Deviance:      10700000
Residual Deviance: 1903000      AIC: NA
```

Los valores de los parámetros asociados al predictor lineal del MLG que se han estimado se plasman en la Tabla 5.19:

Tabla 5.19. Estimadores de los parámetros del predictor lineal para MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica.

$\hat{c}_0$	12.506405
$\hat{\alpha}_1$	0.331272
$\hat{\alpha}_2$	0.321119
$\hat{\alpha}_3$	0.30596
$\hat{\alpha}_4$	0.219316
$\hat{\alpha}_5$	0.270077
$\hat{\alpha}_6$	0.372208
$\hat{\alpha}_7$	0.553333

$\hat{\alpha}_8$	0.398934
$\hat{\alpha}_9$	0.242033
$\hat{\beta}_1$	0.912526
$\hat{\beta}_2$	0.958831
$\hat{\beta}_3$	1.025997
$\hat{\beta}_4$	0.435276
$\hat{\beta}_5$	0.080057
$\hat{\beta}_6$	-0.006381
$\hat{\beta}_7$	-0.394452
$\hat{\beta}_8$	0.009378
$\hat{\beta}_9$	-1.379907

Para poder completar el triángulo *run-off* hasta obtener el cuadrado se aplica la expresión (5.20). Por ejemplo, para el año de origen $i = 9$ y el año de desarrollo $j = 4$ se obtiene:

$$\hat{c}_{94} = \exp\left(\hat{c}_0 + \hat{\alpha}_9 + \hat{\beta}_4\right) = \exp(12.506405 + 0.242033 + 0.435276) = 531635.8\,,$$

que coincide con el resultado obtenido en el método de Chain-ladder, igual que el importe de las provisiones por años de origen, los pagos futuros por años de calendario y la provisión total.

También en este MLG se pueden obtener de los errores de predicción con la expresión analítica para las provisiones por años de origen, para los pagos futuros por años de calendario y para la provisión total a partir de la raíz cuadrada de los errores cuadráticos medios (5.26), (5.27) y (5.28), respectivamente.

En la Tabla 5.20 se indican los errores de predicción para las provisiones por años de origen y para la provisión total y los coeficientes de variación correspondientes:

Tabla 5.20. Provisiones por año de origen y provisión total, errores de predicción y coeficientes de variación para MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica usando fórmula analítica.

Año de origen	Provisión	Error de predicción	Coeficiente de variación
1	94633.8	110099.6	116.34%
2	469511.3	216042.8	46.01%
3	709637.8	260871.3	36.76%
4	984888.6	303549.1	30.82%
5	1419459.5	375012.8	26.42%
6	2177640.6	495376.8	22.75%
7	3920301.0	789959.7	20.15%
8	4278972.2	1046512.6	24.46%
9	4625810.7	1980100.7	42.81%
Total	18680856	2945659	15.77%

De manera análoga se muestran en la Tabla 5.21 los errores de predicción para los pagos futuros por años de calendario y los coeficientes de variación:

Tabla 5.21. Pagos futuros por años de calendario, errores de predicción y coeficientes de variación para MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica usando fórmula analítica.

Año de calendario	Pago	Error de predicción	Coeficiente de variación
10	5226535.8	747369.6	14.30%
11	4179394.4	710144.6	16.99%
12	3131667.5	644139.5	20.57%

13	2127271.9	479125.6	22.52%
14	1561878.9	404967.7	25.93%
15	1177743.7	364294.9	30.93%
16	744287.4	294424.6	39.56%
17	445521.3	250986.8	56.34%
18	86554.6	108268.8	125.09%

Para este MLG se aplica también la metodología *bootstrapping residuals* para analizar los errores de predicción para las provisiones por año de origen (5.33), para la provisión total (5.35) y para los pagos futuros por años de calendario (5.34). Con $B = 1000$ muestras generadas también se puede obtener la media de las distribuciones predictivas en cada caso.

En primer lugar, en la Tabla 5.22, se muestran los resultados para las provisiones por año de origen y para la provisión total:

Tabla 5.22. Provisiones medias por año de origen y provisión media total, errores de predicción y coeficientes de variación para MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica usando *bootstrap* con 1000 muestras.

Año de origen	Provisión media	Error de predicción	Coeficiente de variación
1	100416.0	108422.0	114.57%
2	477357.4	213628.5	45.50%
3	727897.6	257700.5	36.31%
4	978122.3	301692.6	30.63%
5	1438384.0	369127.7	26.00%
6	2194055.0	491173.6	22.55%
7	3934897.0	787571.2	20.09%
8	4236251.0	1032951.4	24.14%
9	4711136.0	2081503.2	45.00%

Total	18757856	2882413	15.43%

En la siguiente Tabla 5.23 se indican los resultados para los pagos futuros por años de calendario:

Tabla 5.23. Pagos futuros medios por años de calendario, errores de predicción y coeficientes de variación para MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica usando *bootstrap* con 1000 muestras.

Año de calendario	Pago medio	Error de predicción	Coeficiente de variación
10	5262187.6	756563.2	14.48%
11	4206004.9	721067.3	17.25 %
12	3153556.8	649753.2	20.75 %
13	2139244.8	487995.9	22.94 %
14	1562523.4	411005.7	26.31 %
15	1178586.1	365547.7	31.04 %
16	771451.6	292974.4	39.36 %
17	455633.0	254458.2	57.11 %
18	91579.0	107988.8	124.76 %

Los resultados obtenidos, tanto aplicando la fórmula analítica como aplicando *bootstrap*, en el cálculo de los errores de predicción se pueden consultar también en Boj y Costa (2015b, 2015c).

Resultados con MLG con distribución Gamma y función de enlace logarítmica

La instrucción para ajustar el MLG con distribución Gamma y función de enlace logarítmica es la siguiente:

```
R> Call:  glm(formula = cij ~ i + j, family = Gamma(link = "log"))

Coefficients:
(Intercept)          i2          i3          i4          i5          i6
   12.55954     0.31725     0.28342     0.16543     0.23059     0.27302
         i7          i8          i9         i10          j2          j3
    0.35231     0.46192     0.30715     0.18890     0.90857     0.93156
         j4          j5          j6          j7          j8          j9
    0.99753     0.41453     0.11082    -0.05421    -0.44967    -0.05944
        j10
   -1.43304

Degrees of Freedom: 54 Total (i.e. Null);  36 Residual
Null Deviance:      20.1
Residual Deviance: 4.023        AIC: 1501
```

Los valores de los parámetros asociados al predictor lineal del MLG que se han estimado se plasman en la Tabla 5.24:

Tabla 5.24. Estimadores de los parámetros del predictor lineal para MLG con distribución Gamma y función de enlace logarítmica.

$\hat{c}_0$	12.55954
$\hat{\alpha}_1$	0.31725
$\hat{\alpha}_2$	0.28342
$\hat{\alpha}_3$	0.16543
$\hat{\alpha}_4$	0.23059
$\hat{\alpha}_5$	0.27302
$\hat{\alpha}_6$	0.35231
$\hat{\alpha}_7$	0.46192
$\hat{\alpha}_8$	0.30715
$\hat{\alpha}_9$	0.18890

$\hat{\beta}_1$	0.90857
$\hat{\beta}_2$	0.93156
$\hat{\beta}_3$	0.99753
$\hat{\beta}_4$	0.41453
$\hat{\beta}_5$	0.11082
$\hat{\beta}_6$	-0.05421
$\hat{\beta}_7$	-0.44967
$\hat{\beta}_8$	-0.05944
$\hat{\beta}_9$	-1.43304

En la siguiente Tabla 5.25 se indican los errores de predicción usando la fórmula analítica para las provisiones por años de origen y para la provisión total y los coeficientes de variación correspondientes:

Tabla 5.25. Provisiones por año de origen y provisión total, errores de predicción y coeficientes de variación para MLG con distribución Gamma y función de enlace logarítmica usando fórmula analítica.

Año de origen	Provisión	Error de predicción	Coeficiente de variación
1	93316.3	45166.4	48.40%
2	446507.0	160557.2	35.96%
3	611147.2	177624.6	29.06%
4	992027.2	254470.9	25.65%
5	1453086.3	351334.3	24.18%
6	2186161.9	526287.9	24.07%
7	3665072.1	941322.3	25.68%
8	4122404.7	1175945.9	28.53%
9	4516082.0	1667392.4	36.92%

Total	18085805	2702710	14.94%

De manera análoga se calculan, en la Tabla 5.26, los errores de predicción para los pagos futuros por años de calendario y los coeficientes de variación:

Tabla 5.26. Pagos futuros por años de calendario, errores de predicción y coeficientes de variación para MLG con distribución Gamma y función de enlace logarítmica usando formula analítica.

Año de calendario	Pago	Error de predicción	Coeficiente de variación
10	5096855.3	847281.6	16.62 %
11	4050001.5	749549.8	18.51 %
12	3064407.7	628141.0	20.50 %
13	2078010.5	431885.8	20.78 %
14	1510392.7	345880.7	22.90 %
15	1095402.7	292255.7	26.68 %
16	692118.4	220057.8	31.79 %
17	416539.9	181226.5	43.51 %
18	82075.9	47918.1	58.38 %

Se generan $B = 1000$ muestras con la metodología *bootstrapping residuals* para obtener los errores de predicción para las provisiones por años de origen, para la provisión total y para los pagos futuros por años de calendario.

En la siguiente Tabla 5.27 se muestran los resultados para las provisiones por años de origen y para la provisión total:

Tabla 5.27. Provisiones medias por años de origen y provisión media total, errores de predicción y coeficientes de variación para MLG con distribución Gamma y función de enlace logarítmica usando *bootstrap* con 1000 muestras.

Año de origen	Provisión media	Error de predicción	Coeficiente de variación
1	93329.9	30298.7	32.47%
2	446523.7	119334.9	26.73%
3	611181.0	126847.6	20.76%
4	992028.9	174635.3	17.60%
5	1453085.0	227208.1	15.64%
6	2186166.0	316433.8	14.47%
7	3665091.0	532016.3	14.52%
8	4122432.0	556356.6	13.50%
9	45166073.0	566682.4	12.55%
Total	18085636	1061700	5.87%

Por último, se indican los resultados correspondientes a los pagos futuros por años de calendario en la siguiente Tabla 5.28:

Tabla 5.28. Pagos futuros medios por año de calendario, errores de predicción y coeficientes de variación para MLG con distribución Gamma y función de enlace logarítmica usando *bootstrap* con 1000 muestras.

Año de calendario	Pago medio	Error de predicción	Coeficiente de variación
10	5096897.2	652964.8	12.81%

11	4050047.9	545647.9	13.47 %
12	3064465.1	434825.7	14.19 %
13	2078021.4	297581.3	14.32 %
14	1510393.3	233914.7	15.49 %
15	1095419.1	194252.2	17.73 %
16	692130.4	142601.5	20.60 %
17	416548.9	109441.6	26.27 %
18	82080.9	26649.2	32.47 %

Los resultados obtenidos, tanto aplicando la fórmula analítica como aplicando *bootstrap*, en el cálculo de los errores de predicción se pueden consultar también en Boj y Costa (2015c).

Resultados con MLGBD con distribución Poisson sobre-dispersa y función de enlace logarítmica

Si se considera la distribución del error Poisson sobre-dispersa, la función de enlace logarítmica y la distancia ℓ^2 Euclídea en el MLGBD, los importes para la provisión total, las provisiones por años de origen y los pagos futuros por años de calendario coinciden con los importes obtenidos en el MLG clásico con distribución Poisson sobre-dispersa y función de enlace logarítmica y también con los importes obtenidos en el método Chain-ladder.

La instrucción para ajustar el MLGBD con distribución Poisson sobre-dispersa, con función de enlace logarítmica y con distancia ℓ^2 Euclídea es la siguiente:

```
R> library(dbstats)
R> dbglm(formula = cij ~ i + j, family = quasipoisson, method = "rel.gvar", metric =
"euclidean ", rel.gvar = 1)

Call:  dbglm(formula = cij ~ i + j, family = quasipoisson, method = "rel.gvar",
    metric = "euclidean ", rel.gvar = 1)
```

```
family: quasipoisson
metric: euclidean
Degrees of Freedom: 54 Total (i.e. Null);  36 Residual
Null Deviance:      10700000
Residual Deviance: 1903000

AIC:        NA
BIC:        NA
GCV:        76454.18
```

En cuanto a los errores de predicción, si se aplican las expresiones analíticas (5.26), (5.27) y (5.28) también coinciden con los resultados numéricos en el caso del MLG clásico.

Por tanto, la aplicación numérica en el caso del MLGBD se centra en los resultados obtenidos al aplicar la metodología *bootstrap*. Cabe recordar que en el MLG clásico se ha aplicado *bootstrapping residuals*, basado en los residuos de Pearson, mientras que en el MLGBD se aplica *bootstrapping pairs*. Se ha considerado un número $B = 1000$ de muestras para los cálculos.

En la siguiente Tabla 5.29 se presentan los resultados de los errores de predicción para las provisiones por año de origen (5.33) y para la provisión total (5.35), así como la media de las distribuciones predictivas en cada caso.

Tabla 5.29. Provisiones medias por año de origen y provisión media total, errores de predicción y coeficientes de variación para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea usando *bootstrap* con 1000 muestras.

Año de origen	Provisión media	Error de predicción	Coeficiente de variación
1	197097.3	155179.5	163.97%

2	567831.6	229653.5	48.91%
3	802433.7	292340.3	41.19%
4	1101788.1	317124.6	32.19%
5	1563680.6	391937.7	27.61%
6	2311566.8	489300.1	22.46%
7	3926323.4	835373.9	21.30%
8	4339191.5	660743.7	15.44%
9	4826806.1	677215.6	14.63%
Total	19554135.3	2231053.6	11.94%

En la siguiente Tabla 5.30 se recogen los resultados correspondientes a los pagos futuros por años de calendario:

Tabla 5.30. Pagos futuros medios por año de calendario, errores de predicción y coeficientes de variación para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea usando *bootstrap* con 1000 muestras.

Año de calendario	Pago medio	Error de predicción	Coeficiente de variación
10	5322047.9	651320.2	12.46%
11	4277227.1	595874.5	14.25%
12	3250816.7	522637.4	16.68%
13	2240502.3	436147.2	20.50%
14	1671934.2	374468.5	23.97%
15	1273321.1	327033.4	27.76%
16	861294.6	276203.3	37.10%
17	543109.0	217777.7	48.88%
18	174215.7	144268.3	166.67%

Resultados con MLGBD con distribución Gamma y función de enlace logarítmica

Si se considera la distribución de error Gamma, la función de enlace logarítmica y la distancia ℓ^2 Euclídea en el MLGBD, los importes para la provisión total, las provisiones por años de origen y los pagos futuros por años de calendario coinciden con los importes obtenidos en el MLG clásico con distribución del error Gamma y función de enlace logarítmica.

La instrucción para ajustar el MLGBD con distribución Gamma, con función de enlace logarítmica y con distancia ℓ^2 Euclídea es la siguiente:

```
R> dbglm(formula = cij ~ i + j, family = Gamma(link="log"), method = "rel.gvar",
metric = "euclidean", weights = rep(1, n), rel.gvar = 1)

Call:  dbglm(formula = cij ~ i + j, family = Gamma(link = "log"), method = "rel.gvar",
    metric = "euclidean", weights = rep(1, n), rel.gvar = 1)

family: Gamma
metric: euclidean

Degrees of Freedom: 54 Total (i.e. Null);  36 Residual
Null Deviance:      20.1
Residual Deviance: 4.023

AIC:        1500.771
BIC:        1538.91
GCV:         0.1616
```

En cuanto a los errores de predicción, si se aplican las expresiones analíticas (5.26), (5.27) y (5.28) también coinciden con los resultados numéricos en el caso del MLG clásico.

En la siguiente Tabla 5.31 se presentan los resultados de los errores de predicción para las provisiones por año de origen (5.33) y para la provisión total (5.34), así como la media de las

distribuciones predictivas en cada caso, cuando se aplica *bootstrapping pairs* con $B=1000$ muestras para los cálculos.

Tabla 5.31. Provisiones medias por año de origen y provisión media total, errores de predicción y coeficientes de variación para MLGBD con distribución Gamma, función de enlace logarítmica y distancia ℓ^2 Euclídea usando *bootstrap* con 1000 muestras.

Año de origen	Provisión media	Error de predicción	Coeficiente de variación
1	199807.2	149998.0	160.74%
2	555157.9	216572.3	48.50%
3	722770.1	238360.9	39.00%
4	1127075.6	300218.6	30.26%
5	1589023.2	368042.5	25.33%
6	2326038.3	475971.9	21.77%
7	3756355.5	869596.4	23.73%
8	4216383.7	725539.0	17.60%
9	4755761.6	733279.3	16.24%
Total	19248389.0	2270695.2	12.55%

Por último, en la Tabla 5.32 se muestran los resultados correspondientes a los pagos futuros por años de calendario:

Tabla 5.32. Pagos futuros medios por año de calendario, errores de predicción y coeficientes de variación para MLGBD con distribución Gamma, función de enlace logarítmica y distancia ℓ^2 Euclídea usando *bootstrap* con 1000 muestras.

Año de calendario	Pago medio	Error de predicción	Coeficiente de variación
10	5237837.3	760410.0	14.92%
11	4185514.9	659058.1	16.27%
12	3199644.1	547190.1	17.86%
13	2211179.4	407489.5	19.61%
14	1641124.5	338332.6	22.40%
15	1229712.7	290088.8	26.48%
16	829033.2	238753.1	34.50%
17	534531.8	190841.6	45.82%
18	179831.5	132739.0	161.73%

5.3.3 Cálculo de provisiones incluyendo márgenes de riesgo

El actual contexto de Solvencia II requiere una exigente gestión empresarial del riesgo de las entidades aseguradoras. En el problema de cálculo de provisiones en seguros no vida es de interés calcular el error de predicción cometido con la metodología utilizada para la estimación de los pagos futuros de la entidad. Además, la distribución predictiva de las estimaciones, que es descriptiva respecto del riesgo, permite obtener, por ejemplo, su valor en riesgo a un nivel de confianza fijado.

Como se indica en Bermúdez (2009), el error de predicción se puede utilizar para calcular las provisiones de una manera más prudente, añadiendo un porcentaje del error de predicción al mejor estimador.

En la aplicación de los MLG a las cuantías de siniestros de un triángulo *run-off* y asumiendo para la distribución del error una familia paramétrica dependiente de un parámetro, junto con la función de enlace logarítmica, en este trabajo se han podido desarrollar las fórmulas del error de predicción de los pagos futuros por años de calendario para la familia paramétrica general, tanto para el caso de utilizar formulación analítica como para el caso de realizar estimación *bootstrap*. En la práctica, las formulaciones presentadas permiten realizar cálculos teniendo en cuenta un ambiente financiero, a partir del valor actual de los pagos futuros para siniestros pendientes, incluyendo márgenes de riesgo con significado estadístico.

En este apartado se describe cómo calcular el valor actual de los pagos futuros por años de calendario, es decir, teniendo en cuenta el valor temporal del dinero, siguiendo la Directiva Solvencia II (artículo 77.2). Para calcular este valor actual, sería preciso conocer la estructura temporal de los tipos de interés y aplicar tipos de interés libres de riesgo (Albarrán y Alonso, 2010).

En España, al igual que en otros países occidentales, la referencia básica de la estructura temporal de los tipos de interés son los valores del Estado. Así, para valorar los flujos es habitual utilizar los tipos de cupón cero, a partir de la cotización de los bonos del Estado en los mercados financieros a distintos plazos.

Se plantean tres maneras distintas de calcular el valor actual de los pagos futuros por años de calendario:

1. En primer lugar, calcular el valor actual de los importes de los pagos futuros por años de calendario, sin incluir ningún tipo de margen de riesgo:

$$IBNR_{actual} = \sum_{t=k+1}^{2k} PF_t \left(1+I_1^t\right)^{-(t-k)}, \tag{5.36}$$

donde I_1^t, para $t = k+1,...,2k$, son los tipos de interés anuales para cada uno de los futuros años de calendario que se tengan en cuenta en el triángulo *run-off*.

2. En segundo lugar, calcular el valor actual de los pagos futuros por años de calendario añadiendo un porcentaje, δ, del error de predicción. De esta manera se está incluyendo un margen de riesgo cada año de calendario igual al porcentaje fijado del correspondiente error de predicción.

   Cabe recordar que se definen dos maneras alternativas de obtener el error de predicción. En el caso de utilizar la formula analítica (5.27) se tiene que:

$$IBNR_{actual} = \sum_{t=k+1}^{2k} \left(PF_t + \delta \sqrt{MSE\left(PF_t\right)} \right) \left(1 + I_1^t\right)^{-(t-k)}. \tag{5.37}$$

   En cambio, si se aplica la metodología *bootstrap* se utiliza la expresión (5.34) y se deriva que:

$$IBNR_{actual} = \sum_{t=k+1}^{2k} \left(PF_t + \delta PE^{boot}\left(PF_t\right) \right) \left(1 + I_1^t\right)^{-(t-k)}. \tag{5.38}$$

3. En tercer lugar, se puede calcular el valor actual de los valores en riesgo, VaR, para un nivel de confianza fijado α de la distribución predictiva de los pagos futuros por años de calendario:

$$IBNR_{actual} = \sum_{t=k+1}^{2k} VaR_{\alpha}\left(\widehat{PF}_t^{boot} \right) \left(1 + I_1^t\right)^{-(t-k)}. \tag{5.39}$$

   En este caso se está añadiendo un margen de riesgo porque se reemplaza la media esperada de las distribuciones predictivas por el cuantil α de dichas distribuciones. Cabe notar que esta manera de incluir un margen de riesgo sólo es posible cuando se estima la distribución predictiva de los pagos futuros por años de calendario mediante *bootstrap.*

5.3.3.1 Aplicación práctica

A continuación se calcula numéricamente la provisión de siniestros pendientes utilizando los

datos de Taylor y Ashe (1983) de las tres maneras distintas que acaban de describirse.

Se supone que el tipo de interés que se aplica es un 1,5% anual fijo para todo el plazo y que en el caso de añadir un porcentaje del error de predicción se utiliza el valor $\delta = 0.995$. Además, el nivel de confianza fijado para obtener el valor en riesgo es del 99.5%, tal como se indica en la Directiva Solvencia II.

Se ilustra numéricamente el cálculo de las provisiones bajo estos supuestos en el caso del MLG clásico con distribución Gamma y con distribución Poisson sobre-dispersa, con función de enlace logarítmica en ambos casos. Además, para el MLGBD se tienen en cuenta las mismas hipótesis y distancia ℓ^2 Euclídea, de manera que los resultados solamente son distintos en la parte referente a la metodología *bootstrap* con respecto al MLG clásico. El estudio se centra en estas dos distribuciones para acotar las dimensiones de la investigación, pero cabe notar que sería posible aplicar otras distribuciones como podría ser la Inversa-Gaussiana.

En primer lugar, en las siguientes Tablas 5.33 y 5.34, se calcula el valor actual de los pagos futuros por años de calendario, sin incluir ningún margen de riesgo, y el valor actual de los mismos pagos pero añadiendo un 99.5% del error de predicción con fórmula analítica. Se muestran los resultados considerando las distribuciones Poisson sobre-dispersa y Gamma, con una función de enlace logarítmica, que coinciden para el MLG clásico y para el MLGBD si se asume la distancia ℓ^2 Euclídea.

Tabla 5.33. Valores actuales de los pagos futuros por años de calendario y de los pagos futuros por años de calendario más un 99.5% del error de predicción usando fórmula analítica con distribución Poisson sobre-dispersa y función de enlace logarítmica en el MLG y en el MLGBD con distancia ℓ^2 Euclídea, asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción
10	1	5226535.8	5970168.6

11	2	4179394.4	4885988.3
12	3	3131667.5	3772586.3
13	4	2127271.9	2604001.9
14	5	1561878.9	1964811.8
15	6	1177743.7	1540217.1
16	7	744287.4	1037239.9
17	8	445521.3	695253.2
18	9	86554.6	194282.1
Valor actual		17873967	21639961

Tabla 5.34. Valores actuales de los pagos futuros por años de calendario y de los pagos futuros por años de calendario más un 99.55% del error de predicción usando fórmula analítica con distribución Gamma y función de enlace logarítmica en el MLG y en el MLGBD con distancia ℓ^2 Euclídea, asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción
10	1	5096855.3	5939900.5
11	2	4050001.5	4795803.6
12	3	3064407.7	3689408.0
13	4	2078010.5	2507736.9
14	5	1510392.7	1854544.0
15	6	1095402.7	1386197.1
16	7	692118.4	911075.9
17	8	416539.9	596860.3
18	9	82075.9	129754.4
Valor actual		17310125	20851676

Si se aplica metodología *bootstrap,* los resultados respecto del error de predicción son distintos para MLG y MLGBD con distancia ℓ^2 Euclídea, asumiendo la misma distribución del error y la misma función de enlace. En las Tablas 5.35 y 5.36 se recogen los resultados correspondientes a la distribución Poisson sobre-dispersa y en las Tablas 5.37 y 5.38 se recogen los resultados correspondientes a la distribución Gamma. En cada tabla se calcula el valor actual de los pagos futuros por años de calendario, el valor actual de los pagos futuros por años de calendario incrementados en un 99.5% del error de predicción y el valor actual de los valores en riesgo de las distribuciones predictivas de los pagos futuros por años de calendario a un 99.5% de nivel de confianza.

Tabla 5.35. Valores actuales de los pagos futuros por años de calendario, de los pagos futuros por años de calendario más un 99.5% del error de predicción y del valor en riesgo con un nivel de confianza del 99.5% en el MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica usando metodología *bootstrap* con 1000 muestras, asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción	$VaR_{99.5}$
10	1	5226535.8	5979316.2	7417055.0
11	2	4179394.4	4896856.4	6364764.7
12	3	3131667.5	3778171.9	5207534.8
13	4	2127271.9	2612827.8	3682095.3
14	5	1561878.9	1970829.6	2735533.8
15	6	1177743.7	1541463.7	2209257.2
16	7	744287.4	1035796.9	1841310.7
17	8	445521.3	698707.2	1262432.7
18	9	86554.6	194003.5	473412.3
Valor actual		17873967	21681420	29688278

Tabla 5.36. Valores actuales de los pagos futuros por años de calendario, de los pagos futuros por años de calendario más un 99.5% del error de predicción y del valor en riesgo con un nivel de confianza del 99.5% en el MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea usando metodología *bootstrap* con 1000 muestras, asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción	$VaR_{99.5}$
10	1	5226535.8	5874599.5	7154048.2
11	2	4179394.4	4772289.5	6154359.3
12	3	3131667.5	3651691.7	4734648.6
13	4	2127271.9	2561238.4	3576892.6
14	5	1561878.9	1934475.1	2788398.2
15	6	1177743.7	1503141.9	2262121.6
16	7	744287.4	1019109.7	1683506.6
17	8	445521.3	662210.1	1209831.3
18	9	86554.6	230101.6	631216.3
Valor actual		17873967	21203120	28720128

Tabla 5.37. Valores actuales de los pagos futuros por años de calendario, de los pagos futuros por años de calendario más un 99.5% del error de predicción y del valor en riesgo con un nivel de confianza del 99.5% en el MLG con distribución Gamma y función de enlace logarítmica, usando metodología *bootstrap* con 1000 muestras y asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción	$VaR_{99.5}$
10	1	5096855.3	5746555.3	5099652.6
11	2	4050001.5	4592921.2	4052656.1

12	3	3064407.7	3497059.3	3067038.9
13	4	2078010.5	2374103.9	2079754.3
14	5	1510392.7	1743137.8	1511867.6
15	6	1095402.7	1288683.6	1096662.3
16	7	692118.4	834006.9	693145.23
17	8	416539.9	525434.3	417478.6
18	9	82075.9	108591.9	82457.2
Valor actual		17310125	19810455	17324230

Tabla 5.38. Valores actuales de los pagos futuros por años de calendario, de los pagos futuros por años de calendario más un 99.5% del error de predicción y del valor en riesgo con un nivel de confianza del 99.5% en el MLGBD con distribución Gamma, función de enlace logarítmica y distancia ℓ^2 Euclídea, usando metodología *bootstrap* con 1000 muestras y asumiendo un tipo de interés fijo anual del 1.5%.

Año de calendario	Diferimiento (en años)	Pago	Pago + 0.995 Error predicción	$VaR_{99.5}$
10	1	5096855.3	5853463.2	6323486.2
11	2	4050001.5	4705764.3	5350768.6
12	3	3064407.7	3608861.8	4230436.4
13	4	2078010.5	2483462.6	2963346.4
14	5	1510392.7	1847033.6	2337500.1
15	6	1095402.7	1384041.1	1815354.8
16	7	692118.4	929677.7	1280757.3
17	8	416539.9	606427.3	880765.3
18	9	82075.9	214151.2	425189.5
Valor actual		17310125	20669343	18085774

Tabla 5.39. Valores actuales de los pagos futuros con MLG con distribución Poisson sobre-dispersa y función de enlace logarítmica y con MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea, usando metodología *bootstrap* con 1000 muestras y asumiendo un tipo de interés fijo anual del 1.5%.

Valor actual	MLG	MLGBD
Sin margen de riesgo (5.36)	17873967	17873967
Incluyendo un 99.5% del error de predicción estimado con expresión analítica (5.37)	21639961	21639961
Incluyendo un 99.5% del error de predicción estimado con *bootstrap* (5.38)	21681420	21203120
Considerando los VaR al 99.5% de nivel de confianza (5.39)	29688278	28720128

Tabla 5.40. Valores actuales de los pagos futuros con MLG con distribución Gamma y función de enlace logarítmica y con MLGBD con distribución Gamma, función de enlace logarítmica y distancia ℓ^2 Euclídea, usando metodología *bootstrap* con 1000 muestras y asumiendo un tipo de interés fijo anual del 1.5%.

Valor actual	MLG	MLGBD
Sin margen de riesgo (5.36)	17310125	17310125
Incluyendo un 99.5% del error de predicción estimado con expresión analítica (5.37)	20851676	20851676
Incluyendo un 99.5% del error de predicción estimado con *bootstrap* (5.38)	19810455	20669343
Considerando los VaR al 99.5% de nivel de confianza (5.39)	17524230	24399235

A modo de resumen, se incluyen en la Tabla 5.39 los valores obtenidos para (5.36), (5.37), (5.38) y (5.39) con los MLG y MLGBD aplicados, en los que se asume distribución de Poisson sobre-dispersa con función de enlace logarítmica, y distancia ℓ^2 Euclídea para la versión en distancias.

De forma análoga se presentan en la Tabla 5.40 los resultados en el caso de distribución Gamma, realizando los mismos supuestos que en la Tabla 5.39.

En la Tabla 5.39, al comparar los dos modelos aplicados, se observa lo siguiente: en el MLGBD se obtienen unos errores de predicción estimados con *bootstrap* menores que en el MLG y, a su vez, menores que los errores de predicción estimados mediante fórmula analítica. El valor actual con MLGBD considerando un margen de riesgo del 99.5% del error de predicción estimado con *bootstrap* supone un incremento del 18.63% con respecto al valor actual sin margen de riesgo. En el caso de considerar los valores en riesgo de las distribuciones predictivas a un nivel de confianza del 99.5%, en el caso del MLGBD el valor actual obtenido es un 60.68% mayor que el valor actual sin margen de riesgo mientras que en el caso del MLG la diferencia es del 66.10%. Es decir, tanto para el MLG como para el MLGBD los valores actuales incluyendo este margen de riesgo son mayores que al incluir un porcentaje del error de predicción.

En caso de asumir distribución Gamma, tal como se refleja en la Tabla 5.40, tanto el MLG como el MLGBD proporcionan unos errores de predicción con *bootstrap* inferiores al caso de expresión analítica, y el valor actual más pequeño se obtiene para el MLG, con un resultado que es un 14.44% mayor que el valor actual sin margen de riesgo. En caso de incluir un margen de riesgo a partir de los valores en riesgo, para MLG se incrementa el valor actual sin margen de riesgo en un 1.24% mientras que para MLGBD se incrementa en un 40.95%. Por ello, para MLG cuando se incluye este margen de riesgo se obtiene un valor actual menor que en caso de incluir un porcentaje del error de predicción, mientras que para MLGBD se observa la situación contraria, es decir, el valor actual teniendo en cuenta los valores en riesgo es mayor que cuando se incluye un porcentaje del error de predicción

Así, para los datos en estudio, en caso de incluir un margen de riesgo teniendo en cuenta el error de predicción, se observa que la metodología *bootstrap* proporciona unos resultados menores en casi todos los modelos en comparación con los casos en que se utiliza expresión analítica. Por tanto, los errores de predicción estimados con *bootstrap* son, en general, menores que los errores de predicción obtenidos a partir de la expresión analítica. Si el experto elige como criterio incluir los errores de predicción para el cálculo de provisiones de una manera prudente, la metodología *bootstrap* resulta más adecuada para estos datos. Por último, señalar que los menores errores de predicción se obtienen para el MLG que asume una distribución Gamma cuando se estiman con *bootstrap.*

Si se realiza un análisis de los resultados obtenidos al calcular el valor actual de los valores en riesgo de las distribuciones predictivas de los pagos futuros por año de calendario a un nivel de confianza del 99.5% cuando se asume una distribución Poisson sobre-dispersa, los incrementos con respecto al valor actual sin margen de riesgo son más elevados que en el caso de asumir una distribución Gamma.

Adicionalmente se pueden ilustrar gráficamente las distribuciones predictivas de la provisión total y de los pagos futuros por años de calendario al aplicar *bootstrap*. A modo de ejemplo, en los Gráficos del 5.1 al 5.39, se representan los histogramas de las distribuciones predictivas de la provisión total y de los pagos futuros por años de calendario en dos casos: para el MLGBD asumiendo distribución de Poisson sobre-dispersa y utilizando distancia ℓ^2, que generaliza el método Chain-ladder clásico; y para el MLG asumiendo distribución Gamma, que proporciona un mejor ajuste de resultados que la distribución de Poisson, ya que no presenta colas tan pesadas en caso de *bootstrap* y tiene, en general, unos menores errores de predicción.

Gráfico 5.1. Distribución predictiva de la provisión total para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

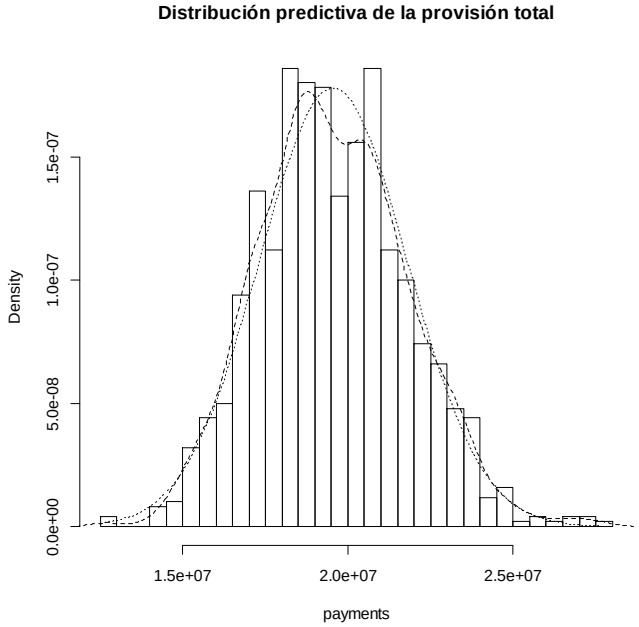


Gráfico 5.2. Distribución predictiva de la provisión total para MLG con distribución Gamma y función de enlace logarítmica.

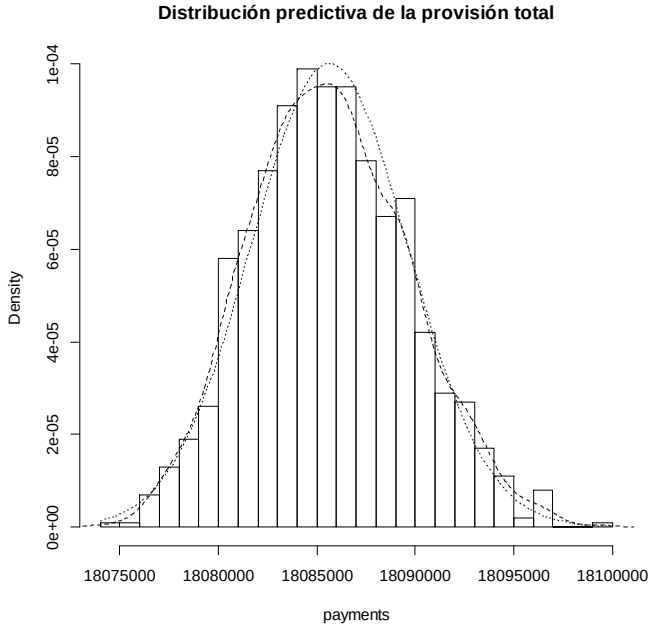


Gráfico 5.3. Distribución predictiva de los pagos futuros del año de calendario 10 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

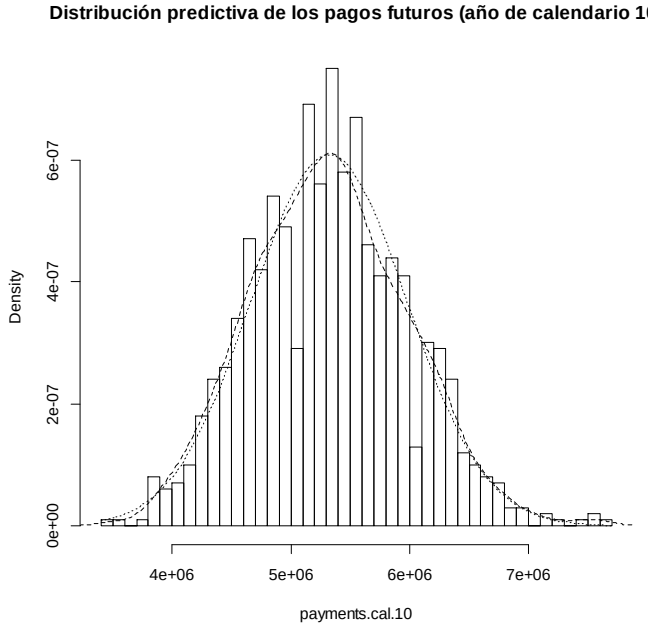


Gráfico 5.4. Distribución predictiva de los pagos futuros del año de calendario 10 para MLG con distribución Gamma y función de enlace logarítmica.

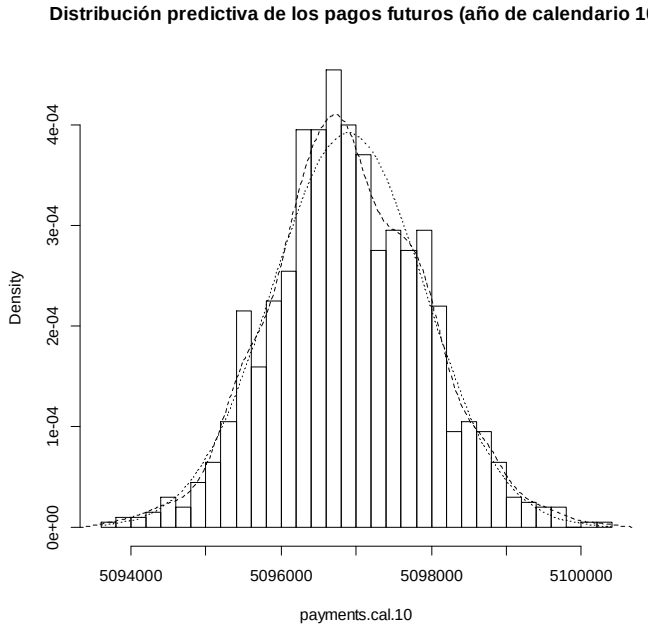


Gráfico 5.5. Distribución predictiva de los pagos futuros del año de calendario 11 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

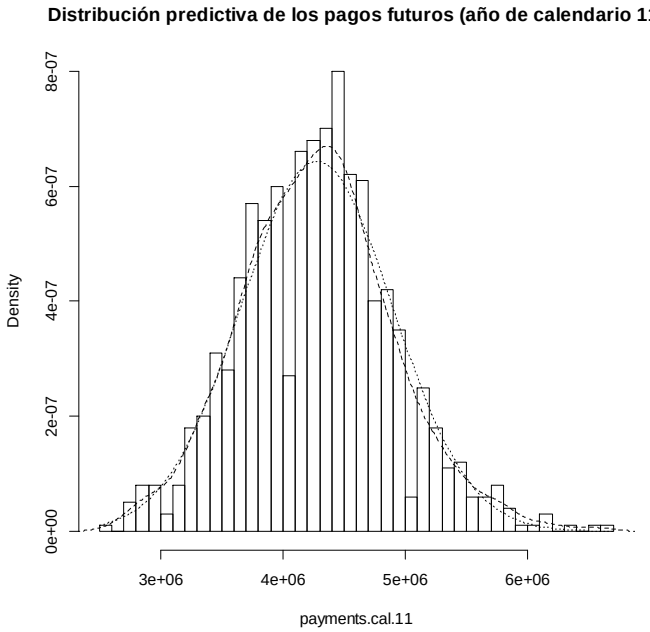


Gráfico 5.6. Distribución predictiva de los pagos futuros del año de calendario 11 para MLG con distribución Gamma y función de enlace logarítmica.

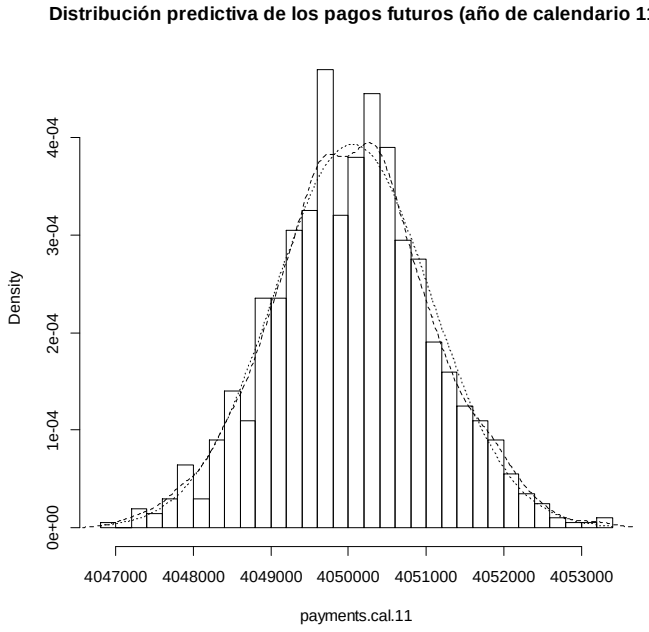


Gráfico 5.7. Distribución predictiva de los pagos futuros del año de calendario 12 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

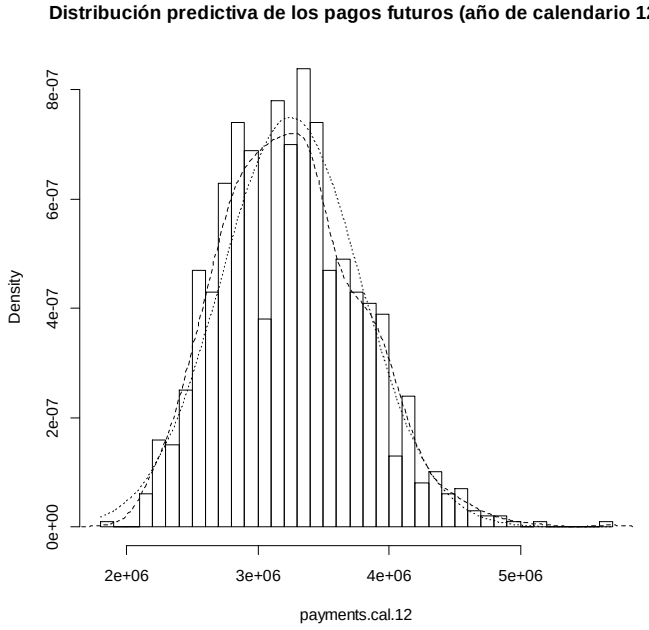


Gráfico 5.8. Distribución predictiva de los pagos futuros del año de calendario 12 para MLG con distribución Gamma y función de enlace logarítmica.

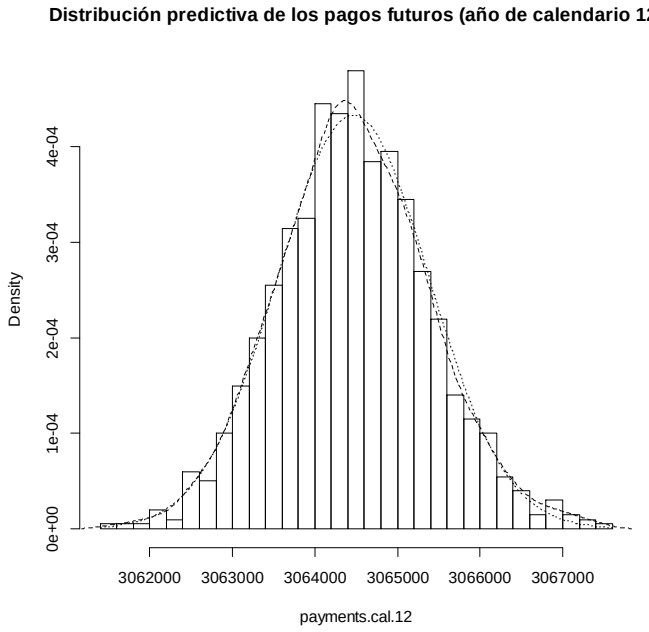


Gráfico 5.9. Distribución predictiva de los pagos futuros del año de calendario 13 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

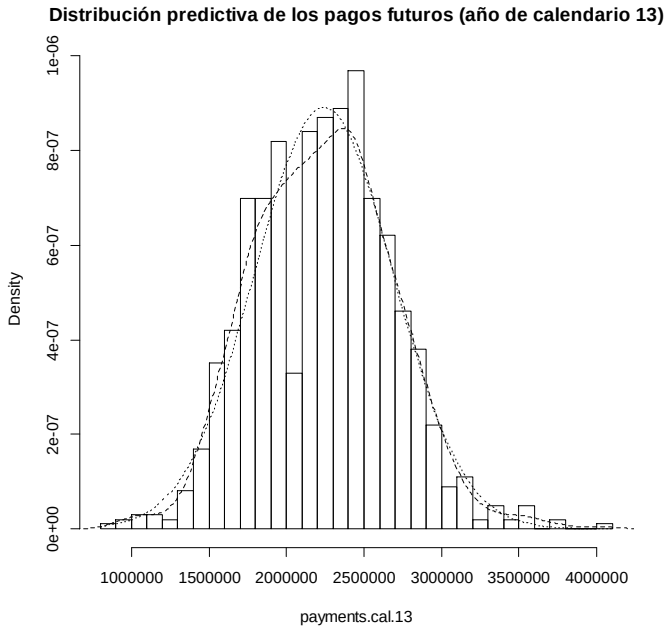


Gráfico 5.10. Distribución predictiva de los pagos futuros del año de calendario 13 para MLG con distribución Gamma y función de enlace logarítmica.

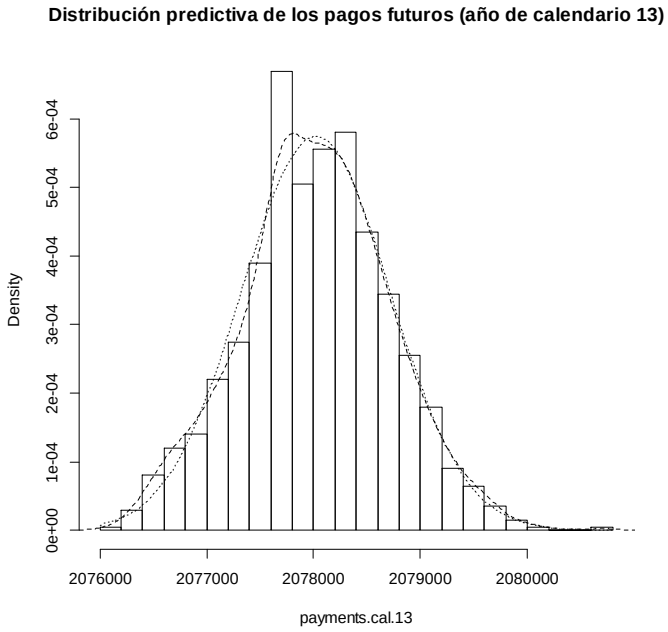


Gráfico 5.11. Distribución predictiva de los pagos futuros del año de calendario 14 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

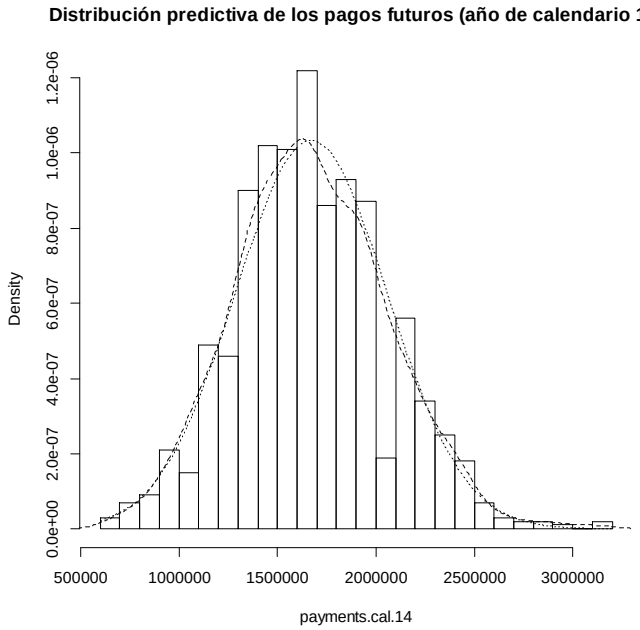


Gráfico 5.12. Distribución predictiva de los pagos futuros del año de calendario 14 para MLG con distribución Gamma y función de enlace logarítmica.

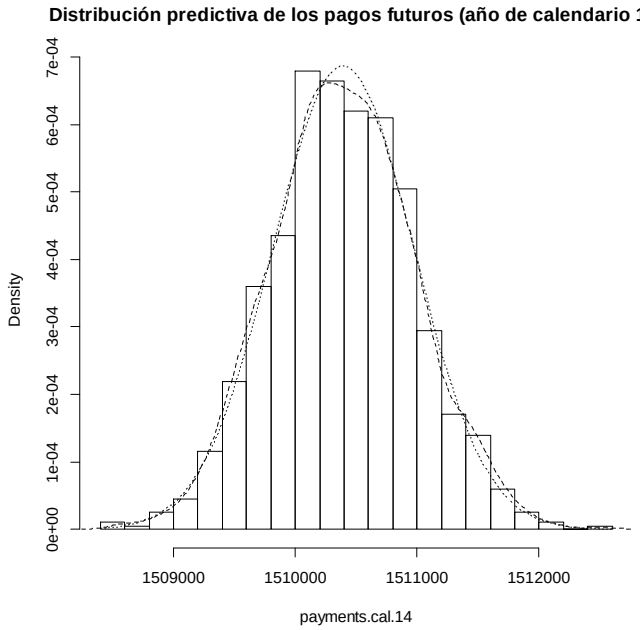


Gráfico 5.13. Distribución predictiva de los pagos futuros del año de calendario 15 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

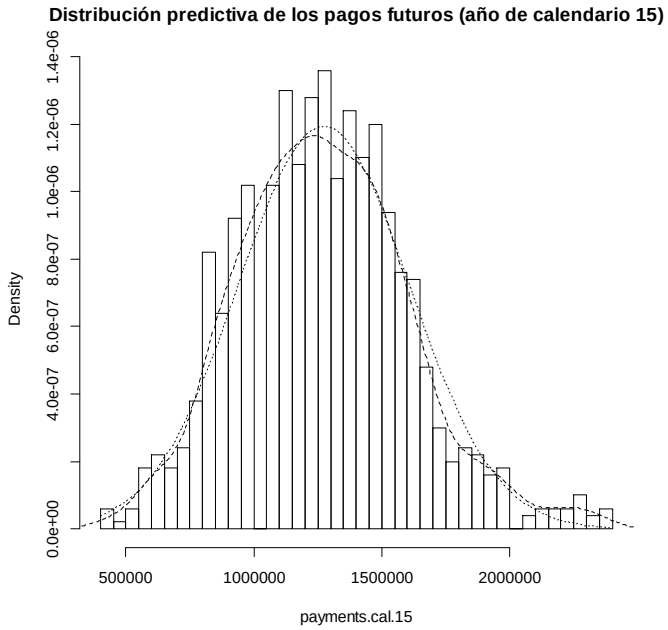


Gráfico 5.14. Distribución predictiva de los pagos futuros del año de calendario 15 para MLG con distribución Gamma y función de enlace logarítmica.

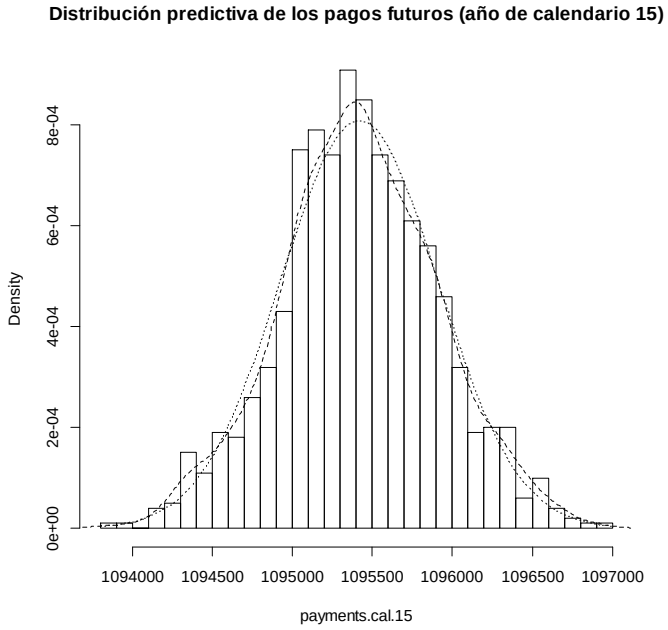


Gráfico 5.15. Distribución predictiva de los pagos futuros del año de calendario 16 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

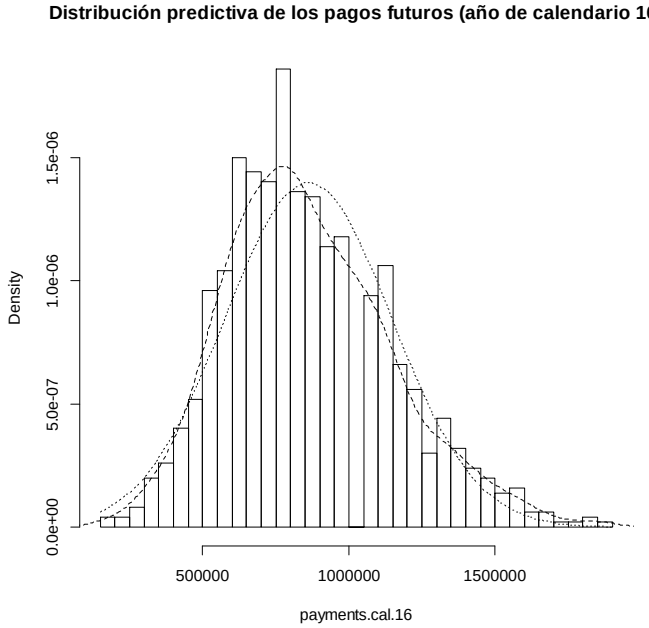


Gráfico 5.16. Distribución predictiva de los pagos futuros del año de calendario 16 para MLG con distribución Gamma y función de enlace logarítmica.

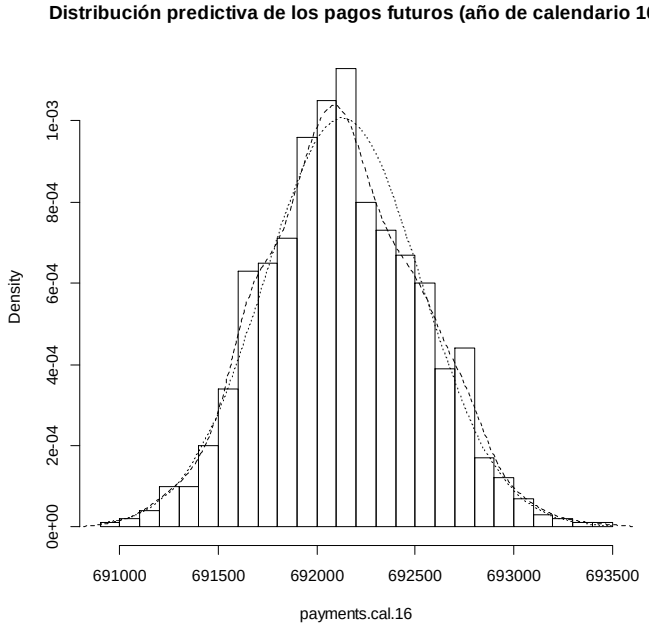


Gráfico 5.17. Distribución predictiva de los pagos futuros del año de calendario 17 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

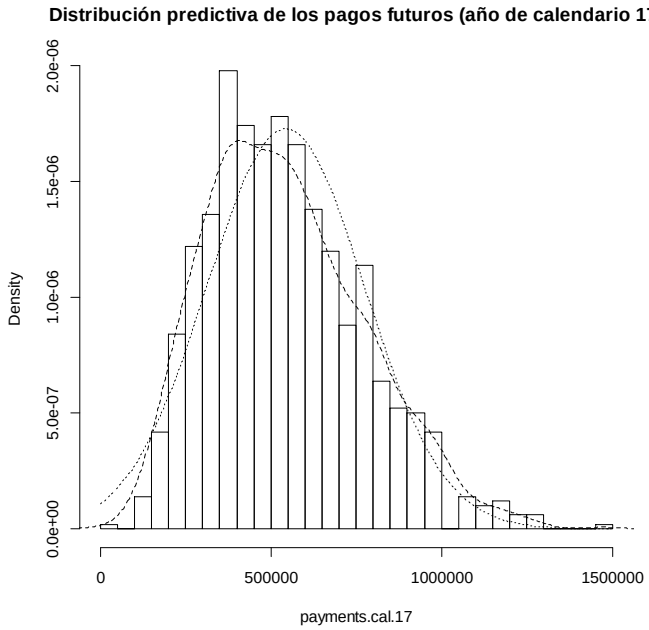


Gráfico 5.18. Distribución predictiva de los pagos futuros del año de calendario 17 para MLG con distribución Gamma y función de enlace logarítmica.

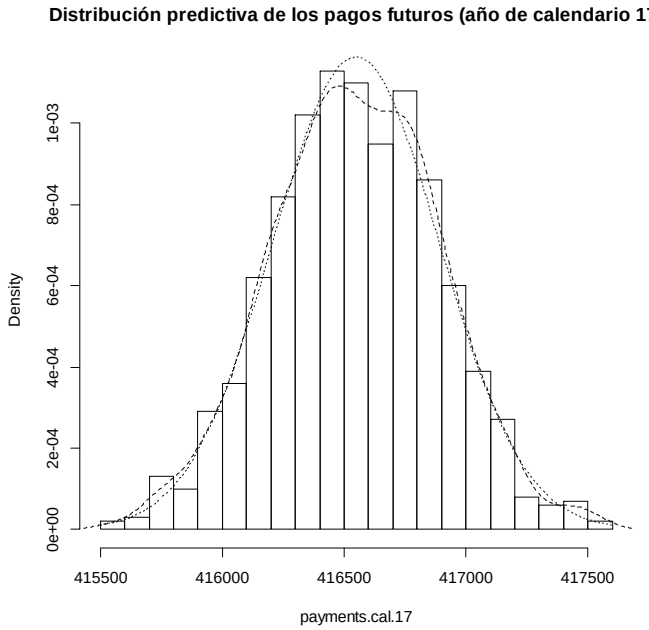


Gráfico 5.19. Distribución predictiva de los pagos futuros del año de calendario 18 para MLGBD con distribución Poisson sobre-dispersa, función de enlace logarítmica y distancia ℓ^2 Euclídea.

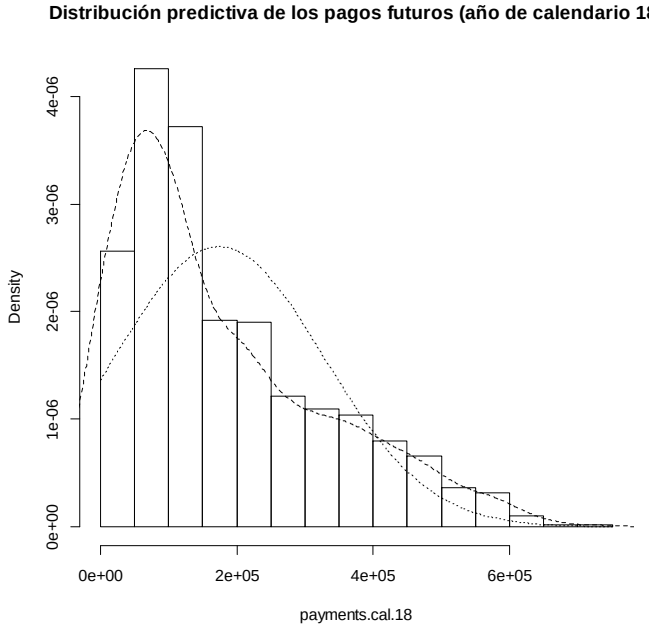
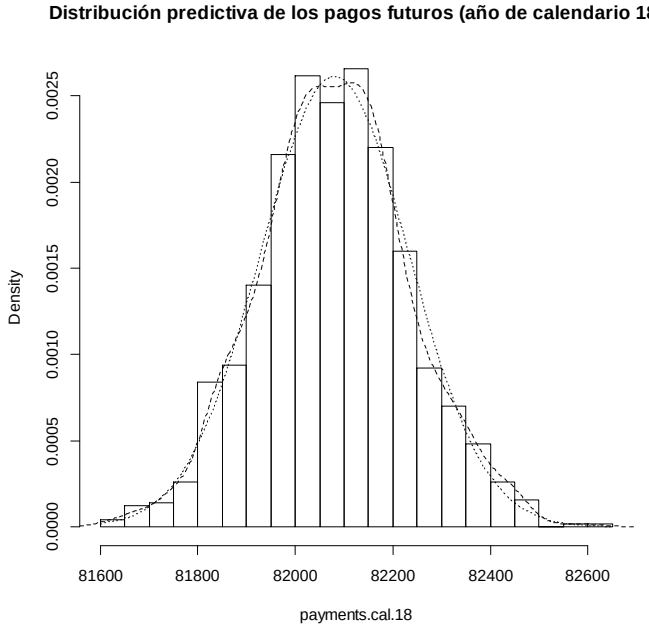


Gráfico 5.20. Distribución predictiva de los pagos futuros del año de calendario 18 para MLG con distribución Gamma y función de enlace logarítmica.



ANEXO 5.1. Anexo informático

Chainladder

Introducción de datos:

```
R> cij <- c(357848, 766940, 610542, 482940, 527326, 574398, 146342,
139950, 227229, 67948, 352118, 884021, 933894, 1183289, 445745, 320996,
527804, 266172)
R> cij <- c(cij, 425046, 290507, 1001799, 926219, 1016654, 750816, 146923,
495992, 280405, 310608, 1108250, 776189, 1562400, 272482, 352053, 206286,
443160, 693190)
R> cij <- c(cij, 991983, 769488, 504851, 470639, 396132, 937085, 847498,
805037, 705960, 440832, 847631, 1131398, 1063269, 359480, 1061648,
1443370, 376686, 986608, 344014)
```

Construcción del triángulo run-off con cuantías acumuladas:

```
R> n <- length(cij)
R> k <- trunc(sqrt(2*n))
R> ii <- rep(1:k, k:1)
R> jj <- sequence(k:1)
R> Cij.mat <- cij.mat <- matrix(0, nrow = k, ncol = k)
R> for(i in 1:n) { cij.mat[ii[i], jj[i]] <- cij[i] }
R> for(i in 1:n) { Cij.mat[i,] <- cumsum(cij.mat[i,])}
R> Cij.mat
```

Cálculo de los factores de desarrollo:

```
R> m <- numeric(k – 1)
R> for (i in 1:(k-1))
{ m[k – i] <- sum(Cij.mat[1:i, k + 1 – i])/sum(Cij.mat[1:i, k – i]) }
R> m
```

Construcción del cuadrado con las cuantías acumuladas:

```
R> for (i in 1:(k – 1))
{ Cij.mat[k:(k – i + 1), I + 1] <- Cij.mat[k:(k – i + 1), i]*m[i] }
R> Cij.mat
```

Construcción del cuadrado con las cuantías desacumuladas:

```
R> cij.mat[,1] <- Cij.mat[,1]
R> for (i in 1:k) { cij.mat[i, 2:k] <- diff(c(Cij.mat[i, 1],Cij.mat[i, 2:k])) }
R> cij.mat
```

Cálculo de la provisión total:

```
R> prov <- sum(Cij.mat[, k]) – sum(cij)
R> prov
```

Cálculo de las provisiones por año de origen:

```
R> orig.prov <- numeric(k – 1)
R> for (orig in 1:(k – 1))
{ orig.prov[orig] <- sum(cij.mat[orig + 1,(k – (orig – 1)):k]) }
R> orig.prov
```

Cálculo de los pagos futuros por año de calendario:

```
R> pagofut <- numeric(k – 1)
R> for (fut in 1:k – 1) { future <- row(cij.mat) + col(cij.mat) – 1 == k + fut
pagofut[fut] <- sum(cij.mat[future]) }
R> pagofut
```

Mínimos cuadrados de De Vylder

Cálculo de los parámetros xi y pj:

```
R> pj <- numeric(k)
R> xi <- numeric(k)
R> pit <- numeric(k); xit <- numeric(k)
R> for (i in 1:k) { pj[i] <- cij.mat[1, i]/Cij.mat[1, k] }
R> pj2 <- pj^2
R> for (i in k:1) { xi[i] <- sum(cij.mat[i, 1:(k – i + 1)]*pj[1:(k – i +
1)])/sum(pj2[1:(k – i + 1)]) }
R> xi2 <- xi^2
R> it <- 1
R> for (i in k:2) { pit[i] <- sum(cij.mat[1:(k – i + 1), i]*xi[1:(k – i +
1)])/sum(xi2[1:(k – i + 1)])
pit[1] <- 1 – sum(pit[2:k]) }
R> pit2 <- pit^2
R> for (i in k:1) {  xit[i] <- sum(cij.mat[i, 1:(k – i + 1)]*pit[1:(k – i +
1)])/sum(pit2[1:(k – i + 1)]) }
R> xit2 <- xit^2
R> while(it<50)
{ for(i in k:2)
{ pit[i] <- sum(cij.mat[1:( k – i + 1),i]*xit[1:(k- i+1)])/sum(xit2[1:( k – i + 1)])
pit[1] <- 1-sum(pit[2:k]) }
pit2 <- pit^2
for (i in k:1)
{ xit[i] <- sum(cij.mat[i, 1:( k – i + 1)]*pit[1:( k – i + 1)])/sum(pit2[1:( k – i +
1)]) }
xit2 <- xit^2
it <- it + 1 }
R> xit
R> pit
```

Construcción del cuadrado con las cuantías no acumuladas:

```
R> xi.mat <- matrix(xit, k, k)
R> pj.mat <- matrix(pit, k, k, byrow=T)
R> cij.est.mat <- xi.mat*pj.mat
R> cij.est.mat
```

Cálculo de la provisión total:

```
R> prov <- sum(cij.est.mat[row(cij.est.mat) + col(cij.est.mat) -1> k])
R> prov
```

Cálculo de las provisiones por año de origen:

```
R> orig.prov <- numeric(k – 1)
R> for (orig in 1:(k – 1)) { orig.prov[orig] <-s um(cij.est.mat[orig + 1, (k – (orig
– 1)):k]) }
R> orig.prov
```

Cálculo de los pagos futuros por año de calendario:

```
R> pagofut<-numeric(k – 1)
R> for (fut in 1:k – 1)
{ future <- row(cij.est.mat) + col(cij.est.mat) – 1 == k + fut
pagofut[fut] <- sum(cij.est.mat[future]) }
R> pagofut
```

Método de separación aritmética

Introducción de datos:

```
R> ni <- c(606, 721, 697, 321, 600, 552, 543, 503, 435, 420)
```

Construcción del triángulo run-off con cuantías medias no acumuladas:

```
R> for(i in 1:n) { cij.mat[ii[i], jj[i]] <- cij[i] }
R> ni.mat <- matrix(rep(ni,each = k), nrow = k, ncol = k, byrow = T)
R> sij.mat <- cij.mat/ni.mat
R> sij.mat
```

Cálculo de los parámetros rj y lambda:

```
R> sum.col <- numeric(k)
R> for (j in 1:k ) { sum.col[j] <- sum(sij.mat[, j])}
R> sum.diag <- numeric(k)
R> for (j in 1:k) { diag.sij <- row(sij.mat) + col(sij.mat) – 1 == j
sum.diag[j] <- sum(sij.mat[diag.sij]) }
R> lambda <- numeric(k)
R> r <- numeric(k)
R> lambda[k] <- sum.diag[k]
R> r[k] <- sum.col[k]/lambda[k]
R> for (i in 1:(k – 1))
{ lambda[k – i] <- sum.diag[k – i]/(1-sum(r[(k – i + 1):k]))
r[k – i] <- sum.col[k – i]/sum(lambda[(k – i):k]) }
R> r
R> lambda
```

Introducción de la inflación futura constante para la estimación de las lambdas futuras, en tanto por uno:

```
R> inf <- 0.015
R> lambdafut <- lambda[k]*(1 + inf)^(1:(k – 1))
R> lambdatot <- c(lambda, lambdafut)
R> matriz.lambda<-matrix(0,k,k)
R> for (i in 1:k)  { matriz.lambda[, i] <- lambdatot[i:(k + i – 1)] }
```

Cálculo del rectángulo de cuantías no acumuladas:

```
R> r.vector <- rep(r, each = k)
R> r.mat <- matrix(r.vector, k, k)
R> cijest.mat <- ni*r.mat*matriz.lambda
R> cijest.mat
```

Cálculo de la provisión total:

```
R>  prov <- sum(cijest.mat[row(cijest.mat) + col(cijest.mat) –1 > k])
R> prov
```

Cálculo de las provisiones por año de origen:

```
R> orig.prov <- numeric(k – 1)
R> for (orig in 1:(k – 1))
R> { orig.prov[orig] <- sum(cijest.mat[orig + 1, (k – (orig – 1)):k]) }
R> orig.prov
```

Cálculo de los pagos futuros por año de calendario:

```
R> pagofut <- numeric(k – 1)
R> for (fut in 1:k – 1)
{ future <- row(cijest.mat) + col(cijest.mat) – 1 == k + fut
pagofut[fut] <- sum(cijest.mat[future]) }
R> pagofut
```

Capítulo 6

Conclusiones: Sinopsis y aportaciones

A continuación se destacan las principales ideas de cada capítulo:

Capítulo 2

En este capítulo se describe la metodología que se aplica para la gestión del riesgo en los problemas de tarificación *a priori*, de *credit scoring* y de cálculo de provisiones en los seguros no vida.

En Boj *et al.* (2004) se aplica el MLBD en el proceso de tarificación *a priori* en los seguros no vida. Posteriormente, en Boj *et al.* (2015b) se define el MLGBD como una generalización del MLBD en el mismo sentido que se generaliza en el caso clásico:

- Las desviaciones aleatorias respecto a la media pueden tener una distribución distinta de la Normal. En función de la variable de siniestralidad que se esté estudiando se puede usar una distribución del error distinta y se considera que puede ser cualquier distribución derivada de la familia exponencial de McCullagh y Nelder (1989). Esta familia paramétrica incluye, además de la Normal, las siguientes distribuciones: Poisson, Poisson sobre-dispersa, Binomial, Gamma e Inversa Gaussiana

- La función que relaciona el predictor lineal latente con la respuesta viene dada por una función de enlace, que puede ser cualquier función monótona y diferenciable

Esta generalización permite que el MLGBD sea competitivo para abordar problemas del ámbito actuarial en los que las hipótesis asumidas en un MLBD no se cumplen, en el mismo sentido que el MLG clásico con respecto al modelo de regresión lineal.

La regresión basada en distancias está implementada en el software *R* en la librería *dbstats* de Boj *et al.* (2014a) y ello permite realizar las aplicaciones prácticas del MLGBD con datos numéricos de seguros de no vida.

Capítulo 3

Las predicciones con el MLGBD se basan en variables latentes calculadas a partir de una matriz de distancias. Esta característica tiene el inconveniente de la no linealidad entre los predictores originales observados y la respuesta, que impide la interpretación de los coeficientes del predictor lineal latente como medidas de influencia de los predictores originales. En aplicaciones actuariales como la tarificación *a priori* no se puede renunciar a esta capacidad, crucial para evaluar la influencia relativa de los factores de riesgo. Con el objetivo de recuperar esta funcionalidad, en este capítulo se definen y estudian coeficientes de influencia, que miden la importancia relativa de los predictores observados.

Se construye un individuo de referencia o virtual que se utiliza como referencia u origen, y se estudian los valores que toman en él los predictores, de manera que los coeficientes de influencia que se obtienen dependerán del individuo elegido. Se puede considerar como referencia, por ejemplo, la media o la mediana para las coordenadas numéricas y la moda para las coordenadas cualitativas o binarias. Se definen coeficientes de influencia para los predictores categóricos o binarios y coeficientes de influencia para los predictores cuantitativos (ver Boj *et al.*, 2015a). Los valores de los coeficientes de influencia tendrán una validez local, en un entorno de un punto dado del espacio predictor, ya que dependen del individuo virtual.

Adicionalmente, se propone aplicar la metodología de *bootstrapping pairs* para obtener muestras de pseudo-datos a partir de los valores originales de los coeficientes. Esto permite calcular los errores estándar y construir intervalos de confianza para los coeficientes de influencia.

Los intervalos de confianza se pueden construir de dos maneras:

- Intervalos de confianza simples, basados en la distribución Normal estandarizada.

- Intervalos de confianza del percentil de *t*, basados en la distribución del estadístico *bootstrap* del test de Wald (ver Boj y Costa, 2015a).

En este capítulo se incluye una aplicación práctica con datos numéricos sobre el seguro de automóviles a terceros descritos en Hallin e Ingenbleek (1983), que ya se habían usado en Boj *et al.* (2012), para estudiar la frecuencia de siniestralidad. Se aplica un MLGBD con distribución de error Poisson y función de enlace logarítmica. En este modelo hay tres factores potenciales de riesgo, de los cuales dos son numéricos y uno es categórico nominal. En total, se estiman once coeficientes de influencia y aplicando *bootstrapping pairs* se calculan los errores estándar y se construyen los intervalos de confianza simples y los intervalos de confianza del percentil de *t*. Con un nivel de confianza del 95% se obtiene que todos los coeficientes son significativos para este conjunto de datos.

Capítulo 4

En este capítulo se estudia la viabilidad de un modelo de regresión logística BD para constituir una metodología alternativa en el cálculo de *scorings* en el problema del riesgo de crédito (ver Costa *et al.*, 2012). Se trata de un caso de MLGBD con distribución del error Binomial y función de enlace *logit.*

A partir de las probabilidades de insolvencia estimadas con el modelo se construye la matriz de confusión y se aplican dos criterios de selección del modelo, siguiendo a West (2000):

- En primer lugar, se estiman las probabilidades de mala clasificación de los individuos, tanto de la población de malos y buenos riesgos como la global. La probabilidad de equivocarse en conceder créditos a malos riesgos es realmente importante y tampoco es bueno clasificar mal a los buenos riesgos ya que, si no se conceden créditos a buenos clientes, en términos esperados no se podrán compensar las pérdidas de los siniestros. Por todo ello, es necesario elegir una técnica predictiva que mantenga un equilibrio entre las tres probabilidades.

- En segundo lugar, se calculan los costes del error en la clasificación de los individuos, bajo el supuesto que el coste de conceder un crédito a un solicitante que sea mal riesgo es significativamente mayor que el coste de denegar un crédito a un solicitante que sea un buen riesgo.

En el método de regresión logística BD se ha utilizado un punto de corte de 0.5 en la probabilidad de insolvencia estimada para decidir si un individuo va a ser un mal riesgo de crédito. Este criterio ya se ha aplicado en otras metodologías de *credit scoring* en West (2000) y en ADBD en Boj *et al.* (2009).

Para completar el análisis del modelo de regresión logística BD se considera que el punto de corte puede tomar valores entre 0 y 1, para ver de qué manera afecta a las probabilidades de mala clasificación de los individuos y a los costes del error en la clasificación. Con estos resultados se describen los criterios de calidad de ajuste con los que elegir un punto de corte "óptimo" para unos datos determinados.

En primer lugar, se puede calcular el índice de Kolmogorov-Smirnov, de manera que se identifica el punto de corte o *score* óptimo que maximiza este índice y es un buen indicador cuando el punto de corte esperado es cercano a este valor óptimo. Gráficamente, se puede identificar este punto de corte en la denominada curva ROC.

En segundo lugar, se puede calcular el índice de Gini para medir la capacidad del modelo para predecir con exactitud los buenos y malos riesgos, que puede utilizarse con fines comparativos.

En este capítulo se incluye una aplicación práctica con datos de riesgo de crédito de dos entidades financieras de Australia y Alemania, respectivamente. Estos datos ya se han utilizado en otros trabajos como West (2000) y Boj *et al.* (2009) y ello permite poder comparar los resultados obtenidos con la regresión logística BD con los resultados que proporcionan otras metodologías, algunas de las cuales también son basadas en distancias, como el ADBD y el método de los *k* vecinos más próximos.

En los datos australianos se puede comprobar como la regresión logística BD resulta ser la metodología de *credit scoring* que proporciona una menor probabilidad de mala clasificación de los malos riesgos y global y, además, los costes del error derivados de la mala clasificación que hace el modelo también son los más pequeños en comparación con los otros métodos. Se comprueba, además, que tanto el punto de corte óptimo, siguiendo el criterio de Kolmogorov-Smirnov, como el valor del *score* para el cual se minimiza tanto la probabilidad de mala clasificación global como los costes del error en la mala clasificación son valores de 0.5 o muy cercanos.

En los datos alemanes, la regresión logística BD es la segunda técnica en minimizar la probabilidad de mala clasificación de los individuos, pero obtiene mejores resultados que la regresión logística clásica y minimiza los costes del error derivados de la mala clasificación de los individuos. Se mantiene alrededor de 0.5 el valor del *score* con el que se maximiza el criterio de Kolmogorvo-Smirnov y se minimizan tanto las probabilidades de mala clasificación como los costes del error, aunque con un rango de valores que varían entre 0.47 y 0.55.

Capítulo 5

Este capítulo se centra en el estudio de las provisiones de siniestros pendientes en los seguros no vida, que forman parte de las provisiones técnicas que debe calcular una entidad aseguradora.

Se presenta el contexto legal, haciendo especial mención de la Directiva Solvencia II, a la que todas las entidades aseguradoras europeas deberán adaptarse antes del año 2016, debido a que introduce consideraciones a tener en cuenta en el cálculo de las provisiones técnicas.

Se describen algunos de los métodos estadísticos deterministas empleados tradicionalmente en el cálculo de provisiones de siniestros pendientes y se destaca el método de Chain-ladder, que ha sido ampliamente estudiado y generalizado en la literatura actuarial. Por otro lado, se introducen los métodos estocásticos de cálculo de provisiones, que son los que la entidad aseguradora deberá aplicar en el contexto de Solvencia II.

En concreto, el MLG clásico es una de las metodologías más estudiadas en su aplicación al cálculo de provisiones y además, cuando la distribución del error que se considera es Poisson sobre-dispersa y la función de enlace es la logarítmica, se obtienen las mismas estimaciones de provisiones que con el método de Chain-ladder determinista.

Por otro lado, se propone la aplicación del MLGBD para el cálculo de provisiones, que generaliza al MLG en el campo de las distancias y, además, contiene como caso particular el método de Chain-ladder clásico.

Con la aplicación de modelos estocásticos, además de obtener predicciones óptimas al generar el proceso subyacente en el triángulo *run-off*, se puede estimar el error de predicción cometido para la estimación de los pagos futuros de la entidad.

Los errores de predicción para las provisiones por años de origen y para la provisión total han sido estudiados en England y Verrall (1999) y England (2002) para el caso del MLG cuando se asume la distribución Poisson sobre-dispersa o la distribución Gamma junto con la función de enlace logarítmica. Dichos errores se pueden calcular o bien a partir de su expresión analítica, o bien, estimando una parte de su expresión a partir de la distribución predictiva de las provisiones obtenida mediante *bootstrap*.

En ese trabajo se deducen y obtienen las formulaciones relativas a los errores de predicción de los pagos futuros por años de calendario para el caso general de la familia paramétrica de distribuciones del error para el MLG. Su estimación se puede realizar mediante expresión analítica y también aplicando *bootstrap* a una parte de la formulación (ver Boj *et al.*, 2014b y Boj y Costa, 2015c).

En Solvencia II se indica que deberá tenerse en cuenta en valor temporal del dinero, es decir, calcular el valor actual de los pagos futuros que se realicen en un determinado año de calendario y, adicionalmente, se considerará un margen de riesgo.

En este sentido, se plantean distintas maneras de obtener el importe de las provisiones por siniestros pendientes:

- A partir del valor actual de los pagos futuros por años de calendario incrementados en un porcentaje del error de predicción (calculado con la fórmula analítica o con *bootstrap).*

- A partir del valor actual de los valores en riesgo de la distribución predictiva de los pagos futuros por años de calendario obtenida con metodología *bootstrap*.

En el caso del MLG clásico se aplica la metodología *bootstrapping residuals* y en el caso del MLGBD se aplica la metodología *bootstrapping pairs*.

Para ilustrar la aplicabilidad de los distintos métodos de cálculo de provisiones se utilizan los datos de Taylor y Ashe (1983), un triángulo con 55 cuantías que se han pagado en el año de origen del siniestro y los 9 años siguientes.

Se estiman los pagos futuros aplicando los distintos métodos estadísticos descritos, tanto deterministas como estocásticos. Además en el MLG y en el MLGBD se obtienen los errores de predicción y la distribución predictiva de los pagos futuros por años de calendario.

Finalmente, se ilustra cómo calcular las provisiones teniendo en cuenta el contexto de Solvencia II.

Aportaciones

Se indican, a modo de resumen, las principales aportaciones del trabajo:

- En el Capítulo 2 se describen a nivel teórico las metodologías de regresión, tanto clásicas como basadas en distancias, que pueden aplicarse a la gestión del riesgo en distintos ámbitos de los seguros no vida

- En el Capítulo 3 se definen coeficientes de influencia para el MLGBD, en el caso de predictores cuantitativos y en el caso de predictores cualitativos o binarios, con el fin de medir la importancia relativa de los factores de riesgo en la siniestralidad.

  Se construyen intervalos de confianza de los coeficientes de influencia de dos maneras distintas: basados en la distribución Normal estandarizada y haciendo uso de la distribución del estadístico *bootstrap* del test de Wald

  A partir de unos datos del seguro de automóviles se calculan los coeficientes de influencia y se construyen los intervalos de confianza del percentil de t para contrastar su significación.

- En el Capítulo 4 se estudia la viabilidad de un modelo de regresión logística BD para constituir una metodología alternativa en el cálculo de *scorings* en el problema del riesgo de crédito. Para ello se estiman con dicho modelo:

  - las probabilidades de mala clasificación de los malos riesgos y global, y

  - los costes del error derivados de la mala clasificación de los individuos.

  Se amplía el estudio para distintos puntos de corte entre 0 y 1 en las probabilidades de insolvencia estimadas para decidir si un individuo es un mal riesgo de crédito. Se incluyen medidas de bondad del ajuste del modelo como el índice Kolmogorov-Smirnov o el índice de Gini.

Se ilustra la idoneidad del modelo de regresión logística BD en comparación con otras metodologías con datos reales de riesgo de crédito de dos entidades financieras.

- En el Capítulo 5 se propone el MLGBD para estimar los pagos futuros que deben servir de base para calcular las provisiones de siniestros pendientes de una entidad aseguradora.

  Para el MLG clásico se deduce la expresión analítica del error de predicción para los pagos futuros por año de calendario en el caso general de la familia paramétrica de distribuciones del error considerando la función de enlace logarítmica.

  Mediante metodología *bootstrap* se obtiene la distribución predictiva de los pagos futuros por año de calendario y, a partir de ella, se estiman los errores de predicción, tanto para MLG como para MLGBD.

  Teniendo en cuenta el contexto de Solvencia II se plantean distintas maneras de calcular el mejor estimador de las provisiones y considerar márgenes de riesgo.

  Se realiza una comparativa de distintos métodos estadísticos de cálculo de provisiones a partir de unos datos sobre importes de siniestros que han sido ampliamente utilizados en la literatura actuarial.

Bibliografía

Albarrán, I. y P. Alonso (2010). *Métodos estocásticos de estimaciones de las provisiones técnicas en el marco de Solvencia II*. Cuadernos de la Fundación Mapfre **158**. Fundación MAPFRE Estudios, Madrid.

Andrews, D.F. y A.M. Herzberg (1985). *Data. A collection of problems from many fields for the student and research worker*. Springer, New York.

Balzarotti, V. y F. Castelpoggi (2009). Modelos de puntuación crediticia: la falta de información y el uso de datos de una central de riesgos. *Ensayos Económicos (Banco Central de la República Argentina)* **56**, 95–156.

Bermúdez, Ll. (2009). Métodos estocásticos para el cálculo de la provisión técnica de prestaciones pendientes en Solvencia II. *Cuadernos actuariales* **13**, 1–12.

Boj, E., Claramunt, M.M. y J. Fortiana (2004). *Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación*. Cuadernos de la Fundación MAPFRE **88**. Fundación MAPFRE Estudios, Madrid.

Boj, E., Claramunt, M.M. y J. Fortiana (2007). Selection of predictors in distance-based regression. *Communications in Statistics: Simulation and Computation* **36:1**, 87–98.

Boj, E., Delicado, P. y J. Fortiana (2008). Logistic and local logistic distance-based regression. *Proceedings of the International Seminar on Nonparametric Inference ISNI 2008*, 66–70.

Boj, E., Claramunt, M.M., Esteve, A. y J. Fortiana (2009). Criterios de selección de modelo en credit scoring, aplicación del análisis discriminante basado en distancias. *Anales del Instituto de Actuarios Españoles*, Tercera Época **15**, 209–230.

Boj, E., Delicado, P. y J. Fortiana (2010). Local linear functional regression based on weighted distance-based regression. *Computational Statistics and Data Analysis* **54**, 429–437.

Boj, E., Fortiana, J., Esteve, A., Claramunt, M.M. y T. Costa (2011). Aplicación de un modelo de regresión logística basado en distancias en el problema de credit scoring. En: Feria, J.M. *et al. Investigaciones en Seguros y Gestión de riesgos: RIESGO 2011, pp. 293–305*. Cuadernos de la Fundación MAPFRE **171**. Fundación MAPFRE Estudios, Madrid.

Boj, E., Delicado, P., Fortiana, J., Esteve, A. y A. Caballé (2012). Local distance-based generalized linear models using the dbstats package for R. *Documentos de Trabajo de la Xarxa de Referència en Economia Aplicada (XREAP), XREAP2012-11*.

Boj, E., Caballé, A., Delicado, P. y J. Fortiana (2014a). *dbstats: distance-based statistics (dbstats)*. R package version 1.4.
http://cran.r-project.org/web/packages/dbstats/index.html

Boj, E., Costa, T. y J. Espejo (2014b). Provisiones técnicas por años de calendario mediante modelo lineal generalizado. Una aplicación con R-Excel. *Anales del Instituto de Actuarios Españoles*, Tercera Epoca **20,** 83–116.

Boj, E. y T. Costa (2015a). Wald test and distance-based generalized linear models. Actuarial Application. *Global Journal of Pure and Applied Mathematics* **11:1**, 295–306.

Boj, E. y T. Costa (2015b). Claim reserving: calendar year reserves for the GLM. En: Guillén, M. *et al. Current Topics on Risk Analysis: ICRA6 and RISK2015 Conference*. Cuadernos de la Fundación MAPFRE **205**, 169–177. Fundación MAPFRE, Madrid.

Boj, E. y T. Costa (2015c). Provisions for claims outstanding, incurred but bot reported, with generalized linear models: prediction error formulation by calendar years. *Cuadernos de gestión* (to appear).

Boj, E., Costa, T., Fortiana, J. y A. Esteve (2015a). Assessing the Importance of Risk Factors in Distance-Based Generalized Linear Models. *Methodology and Computing in Applied Probability* (to appear).

Boj, E., Delicado, P., Fortiana, J., Esteve A. y A. Caballé. (2015b). Global and local distance-based generalized linear models. *TEST* (to appear).

Brockman, M.J. y T.S. Wright (1992). Statistical model rating: making effective use of your data. *Journal of the Institute of Actuaries* **119**, 457–543.

Claramunt, M.M. y T. Costa (2003). Matemática Actuarial No Vida. Un enfoque práctico. *Colección de Publicaciones del Departamento de Matemática Económica, Financiera y Actuarial* **63**.

Costa, T., Boj. E. y J. Fortiana (2012). Bondad de ajuste y elección del punto de corte en regresión logística basada en distancias. Aplicación al problema del credit scoring. *Anales del Instituto de Actuarios Españoles*, Tercera Época **18**, 19-40.

Cuadras, C.M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. En: Dodge, Y. *Statistical Data Analysis and Inference, pp. 459–473*. Elsevier Science Publishers B. V., North-Holland, Amsterdam.

Cuadras, C.M. y C. Arenas (1990). A distance-based regression model for prediction with mixed data. *Communications in Statistics: Theory and Methods* **19**, 2261–2279.

Cuadras, C.M., Arenas, C. y J. Fortiana (1996). Some computational aspects of a distance-based model for prediction. *Communications in Statistics: Simulation and Computation* **25:3**, 593–609.

Davidson, A.C. y D.V. Hinkley (1997). *Bootstrap methods and their application*. Cambridge University Press, New York.

De Vylder, F. (1978). Estimation of IBNR claims by least squares. *Mittelungen der Vereinigung Schewizerischer Versicherungsmathematike* **78**, 249–254.

Efron, B. (1979). Bootstrap methods: another look at the Jacknife. *Annals of Statistics* **7**, 1–26.

Efron, B. y R. Tibshirani (1998). *An introduction to the bootstrap*. Chapman & Hall, London.

England, P.D. y R.J. Verrall (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics* **25**, 281–293.

England, P. (2002). Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving". *Insurance: Mathematics and Economics* **31**, 461–466.

England, P.D. y R.J. Verrall (2002). Stochastic claims reserving in general insurance (with discussion). *British Actuarial Journal* **8**, 443–544.

England, P.D. y R. J. Verrall (2006). Predictive Distributions of Outstanding Liabilities in General Insurance. *Annals of Actuarial Science* **1:II**, 221–270.

Esteve, A. (2003). *Distancias Estadísticas y Relaciones de Dependencia entre Conjuntos de Variables*. Tesis Doctoral. Universidad de Barcelona.

Esteve, A., Boj, E. y J. Fortiana (2009). Interaction terms in distance-based regression. *Communications in Statistics: Theory and Methods* **38**, 3498-3509.

Flachaire, E. (1999). A better way to bootstrap pairs. *Economics Letters* **64**, 257–262.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics & Data Analysis* **49**, 361–376.

Freedman, D.A. (1981). Bootstrapping regression models. *Annals of Statistic* **9**, 1218-28.

Gesmann, M., Murphy, D., Zhang, Y., Carrato, A., Wuthrich, M. y F. Concina (2015). ChainLadder: Statistical methods for the calculation of outstanding claims reserves in general insurance. R package version 0.2.2.
http://cran.r-project.org/web/packages/ChainLadder/index.html

Gower J.C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–874.

Gower, J. C. y S. Harding (1988). Nonlinear biplots. *Biometrika* **75**, 445–455.

Haberman, S. y A.E. Renshaw (1996). Generalized Linear Models and Actuarial Science. *Journal of the Royal Statistical Society. Series D (The Statistician)* **45:4**, 407–436.

Haberman, S. y A.E. Renshaw (1998). Actuarial applications of generalized linear models. En: *Statistics in Finance*, D. J. Hand and S. D. Jacka (eds). Arnold, London.

Hallin, M. y J.F. Ingenbleek. (1983). The Swedish automobile portfolio in 1977. A statistical study. *Scandinavian Actuarial Journal* **83**, 49-64.

Hosmer, D.W. y S. Lemeshow (2000). Applied logistic regression (Second edition). John Wiley & Sons, Inc., New York.

Institute and Faculty of Actuaries (1997). *Claims Reserving Manual*. Institute of Actuaries, London.

Iñiguez, C.A. y M.G. Morales (2009). *Selección de perfiles de clientes mediante regresión logística para muestras desproporcionadas, validación, monitoreo y aplicación en la proyección de provisiones*. Proyecto de Fin de Carrera, Escuela Politécnica Nacional, Ecuador.

Kaas, R., Goovaerts, M., Dhaene, J. y M. Denuit (2008*). Modern Actuarial Risk Theory. Using R (Second edition)*. Springer-Verlag, Heidelberg.

Mack, T. (1993). Distribution Free Calculation of the Standard Error of Chain Ladder Reserve Estimates. *ASTIN Bulletin* **23**, 213–225.

Mack, T. y G. Venter (2000). A Comparison of Stochastic Models that Reproduce Chain Ladder Reserve Estimates. *Insurance: Mathematics and Economics* **26**, 101–107.

MacKinnon, J.G. (2002). Bootstrap inference in econometrics. *The Canadian Journal of Economics* **35:4**, 615–645.

MacKinnon, J.G. (2006). Bootstrap methods in econometrics. *The Economic Record* **82**, *special issue, september 2006*, s2–s18.

MacKinnon, J.G. (2007). Bootstrap hypothesis testing. *Queen's Economics Department Working Paper* **1127**.

McCullagh, P. y J.A. Nelder (1989). *Generalized Linear Models (Second edition)*. Chapman & Hall, London.

Millenhall, S.J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Society* **86**, 393-487.

Pérez, J. L. (2001). *Conociendo el seguro (Segunda Edición).* Editorial UMESER S.A., Barcelona.

Pinheiro, P.J.R., Andrade e Silva, J.M. y M.d.L. Centeno (2003). Bootstrap Methodology in Claim Reserving. *The Journal of Risk and Insurance* **4**, 701–714.

R Development Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Renshaw, A.E. (1989). Chain ladder and interactive modelling (claims reserving and GLIM). *Journal of the Institute of Actuaries* **116:III**, 559-587.

Renshaw, A. y R. Verrall (1998). A Stochastic Model Underlying the Chain-Ladder Technique. *British Actuarial Journal* **4:IV**, 903–923.

Reyes, J., Escobar, C., Duarte, J. y P. Ramírez (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Estudios Pedagógicos* **23:2**, 101–120.

Řezáč, M. y F. Řezáč (2011). How to Measure the Quality of Credit Scoring Models. *Journal of Economics and Finance* **61:5**, 486–507.

Taylor, G. y F.R. Ashe (1983). Second Moments of Estimates of Outstanding Claims. *Journal of Econometrics* **23**, 37–61.

Van Eeghen, J.; Greup, E.K. y J.A. Nijssen (1981). Loss Reserving Methods. *Surveys of Actuarial Studies* **1**, National Nederlanden.

Verbeek, H.G. (1972). An approach to the analysis of claims experience in motor liability excess of loss reinsurance. *ASTIN Bulletin* **6**, 195-202.

Verrall, R. (2000). An Investigation into Stochastic Claims Reserving Models and the Chain-ladder Technique. *Insurance: Mathematics and Economics* **26**, 91–99.

Verrall, R. y P. England (2000). Comments on: "A Comparison of Stochastic Models that reproduce Chain Ladder Reserve Estimates, by Mack and Venter". *Insurance: Mathematics and Economics* **26**, 109–111.

West, D. (2000). Neural network credit scoring models. *Computer & Operations Research* **27**, 1131–1152.

Wood, S.N. (2006). *Generalized additive models: an introduction with R*. Chapman & Hall, Boca Raton.