

DOCUMENTS DE TREBALL
DE LA DIVISIÓ DE CIÈNCIES JURÍDIQUES
ECONÒMIQUES I SOCIALS

Col·lecció d'Economia

HERRAMIENTAS ESTADÍSTICAS PARA
EL ESTUDIO DE PERFILES DE RIESGO *

Eva Boj del Val

M^a Mercè Claramunt Bielsa

Josep Fortiana Gregori

Adreça de correspondència:

Dep. Matemàtica Econòmica, Financera i Actuarial
Facultat de Ciències Econòmiques i empresarials
Universitat de Barcelona
Av. Diagonal 690,
08034_Barcelona, ESPANYA
e-mails: *boj@eco.ub.es claramun@eco.ub.es fortiana@mat.ub.es*

* Treball finançat en part per la beca “Riesgo y Seguro 2000” de la fundació MAPFRE
Estudis: “*Anàlisi multivariant aplicat a la selecció de factors de risc en la tarificació*”.

Resum

En aquest treball s'il·lustra, en sentit pràctic, la utilització de tres eines que permeten a l'actuari definir els grups de tarifa i estimar les primes de risc en un procés de tarificació *a priori* no vida. La primera és l'anàlisi de segmentació (*CHAID* i *XAID*) utilitzat inicialment per *UNESPA* al 1997 en la seva cartera comú d'automòbils. La segona és un procés de selecció de predictors pas a pas amb el model de regressió basada en distàncies. I la tercera és un procés amb els models lineals generalitzats que representen la tècnica més actual de la bibliografia actuarial. D'aquests últims, combinant diferents funcions link i distribucions de l'error, es desprenen els clàssics models additiu i multiplicatiu, dels quals s'interpreta el significat.

Abstract

In this paper it is illustrated, in a practical way, the use of three tools that permit the actuary to define tariff groups and to estimate risk premiums in the class-rating process for non-life insurance. The first is the segmentation analysis (*CHAID and XAID*) used firstly at 1997 by *UNESPA* in its common portfolio of automobiles. The second is a stepwise selection process with the distance-based regression model. And the third is a process with the well known generalized linear regression model, which represents the most modern technique in the actuarial bibliography. From the later, if we combine different link functions and error distributions, we can obtain the classical additive and multiplicative models.

Key words: Segmentation analysis, credibility models, distance-based models, generalized linear models, risk profiles, class-rating, non-life insurance.

JEL: C12, C51, C52, G22.

1. INTRODUCCIÓN

En el trabajo nos ocupamos de la **selección de variables de tarifa** de entre el conjunto de factores potenciales de riesgo. Inicialmente nos va a interesar estudiar la relación de dependencia entre la variable univariante experiencia de siniestralidad, $\mathbf{Y}_{(n \times 1)}$, y cada factor de riesgo univariante, $\mathbf{X}_{(n \times 1)}$, de forma individualizada. También estudiaremos las relaciones entre predictores. Para ello haremos uso de medidas de asociación, que tenderán a 0 si hay independencia y a 1 si existe algún tipo de relación funcional. Distinguiendo el tipo de variables, tenemos [Cuadras y Sánchez (1997)]:

- Continua-continua: partimos de dos variables, $\mathbf{X}$ e $\mathbf{Y}$, continuas $(n \times 1)$, y calculamos la ya conocida correlación al cuadrado, ρ^2 :

$$\rho = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{Cov(\mathbf{Y}, \mathbf{X})}{\sqrt{Var(\mathbf{Y})Var(\mathbf{X})}} \quad (1)$$

- Continua-categorica: partimos de una variable, $\mathbf{Y}$, continua $(n \times 1)$ y de una variable, $\mathbf{X}$, categorica con k clases, de manera que para cada valor x_i de $\mathbf{X}$ tenemos los valores $y_{i1}, y_{i2}, \dots, y_{in_i}$ de $\mathbf{Y}$ para $i = 1, 2, \dots, k$, entonces a partir del análisis de la variancia tenemos la medida

$$\eta^2 = 1 - \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2} = 1 - \frac{Q_d}{Q_t} = \frac{Q_e}{Q_t} \quad (2)$$

- Categorica-categorica: partimos de dos variables, $\mathbf{X}$ e $\mathbf{Y}$, categoricas con p y q clases respectivamente. Sea $N = (n_{ij})$ la tabla de contingencia $p \times q$ que las resume. Una buena medida de asociación entre filas y columnas de la tabla debe estar basada en el estadístico Chi-cuadrado que se utiliza usualmente para contrastar la independencia:

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \left(n_{ij} - \frac{n_{i.} \times n_{.j}}{n_{..}} \right)^2 \bigg/ \left(\frac{n_{i.} \times n_{.j}}{n_{..}} \right) \quad (3)$$

donde $n_{i.} = \sum_{j=1}^q n_{ij}$, $n_{.j} = \sum_{i=1}^p n_{ij}$, $n_{..} = \sum_{i=1}^p \sum_{j=1}^q n_{ij}$, y como medida de asociación se propone el siguiente coeficiente

$$C^2 = \frac{\chi^2}{r \times n_{..}} \quad (4)$$

donde $r = \min\{p, q\} - 1$. Si además se tiene interés en tener en cuenta la ordinalidad de al menos una de las dos variables categóricas implicadas en una tabla de contingencia, nos referimos a Agresti (1984) pp. 156-179.

Las medidas de asociación nos permitirán conocer la relación variable a variable con la siniestralidad, pero **el objetivo es la obtención de un conjunto “equilibrado” de variables de tarifa**. Si seleccionamos separadamente las variables que una a una están más asociadas con el riesgo, es posible que en el conjunto de variables seleccionadas resultante dispongamos de información redundante o bien que no tengamos incorporadas variables que de manera conjunta con otras resulten significativas. Así, se hace necesario realizar el estudio conjunto teniendo en cuenta a la vez todos los factores potenciales del riesgo e “idealmente” todas sus interacciones. Para ello haremos uso de técnicas de análisis estadístico multivariante. Algunas de las técnicas nos serán útiles además para seguir cubriendo las siguientes etapas hasta la formación de los grupos de tarifa exclusivos y exhaustivos, y/o hasta la posterior estimación de la prima pura para cada asegurado. Distinguiremos: técnicas de predicción mediante modelos de regresión, técnicas de agrupación de individuos mediante análisis cluster y técnicas de clasificación de individuos mediante análisis discriminante.

Las técnicas de regresión consisten en la estimación de la respuesta a partir de una serie de variables explicativas o predictores. Disponemos de una gran variedad de modelos a los que adaptar nuestro tipo de datos. Su aplicación a la

selección de variables de tarifa es la siguiente: dado un modelo concreto, por supuesto acorde con nuestros datos, buscaremos mediante un proceso de selección de predictores la “mejor” combinación de ellos para la estimación del riesgo, y éstos pasarán a ser el conjunto de variables de tarifa. Con estas técnicas podemos, si nos interesa, realizar ya una estimación apurada de las primas puras. Destacamos especialmente los modelos lineales generalizados y los modelos basados en análisis de distancias de los que haremos uso en la aplicación. Incluimos aquí los modelos de credibilidad, pues son técnicas de regresión que nos permiten realizar una estimación de la siniestralidad a partir de unos grupos homogéneos de riesgo. Aunque cabe notar que precisan ser combinados previamente con alguna metodología de selección de variables que nos indique cuáles serán los grupos de tarifa homogéneos iniciales.

Los métodos de análisis cluster sirven para la formación de grupos homogéneos de individuos. Los métodos pueden ser [Andenberg (1973); Hartigan (1975); Hawkins y Kass (1982); Sierra (1986)]: jerárquicos / no jerárquicos; aglomerativos / divisivos; monotéticos / politéticos. En nuestro estudio destacamos únicamente la aplicabilidad de los métodos jerárquicos aglomerativos y divisivos politéticos.

- *Jerárquicos:*

Aglomerativos: empiezan con las clases básicas (individuo a individuo) y las van fusionando para formar subclusters. El punto común de partida lo constituye una matriz de distancias entre individuos que se calcula a partir de la experiencia de siniestralidad objeto de estudio. Éstos son de utilidad exclusivamente para la formación de clases de tarifa sólo si cada individuo representa una clase de tarifa inicial, por ejemplo, si queremos agrupar por zonas, cada individuo deberá ser una provincia. Así su aplicabilidad es limitada y no es adecuado para la selección de variables. Dentro de los jerárquicos aglomerativos también se clasifica al método de Ward [Campbel (1986); Ward (1963)]. Éste se separa del resto porque no se basa en una

función de distancias, sino en la minimización de la varianza dentro de los grupos que va formando. Nosotros lo recomendamos exclusivamente para la discretización de variables continuas.

Divisivos: empiezan con el conjunto completo de individuos que forman un solo cluster y van particionando sucesivamente en clases más finas. Dentro de ellos englobamos principalmente las técnicas politéticas de segmentación que veremos con detalle. Hace unos años se empezaron a utilizar fundamentalmente en el seguro de automóvil como herramienta de toma de decisiones de las entidades [Calatayud y Martínez (1997); Pérez (2001)]. Una aplicación concreta de CHAID la tenemos en la segmentación de la cartera común de automóviles de UNESPA [UNESPA (1995)] elaborada por Sánchez (1997). Éstas son un caso especial de análisis cluster que también puede ser considerado como técnica de regresión, pues necesita de una variable respuesta y de un conjunto de predictores. El resultado final del análisis de segmentación son unos segmentos terminales que nos resumen los grupos de tarifa a partir de una clasificación no cruzada y que tiene en cuenta, aunque de un modo jerárquico, el efecto de interacción entre predictores. Éstas cubren todas las fases de la tarificación, aunque su predicción está limitada a las clases ya existentes de los factores categóricos seleccionados y la única opción para la estimación de la prima pura es alguna media de los grupos terminales.

- *No jerárquicos*: son métodos que optimizan un funcional objetivo fijado un número de clusters. Los que como objetivo tienen alguna variedad de minimización de la varianza dentro de los grupos a formar, pueden ser utilizados unidimensionalmente, al igual que el método jerárquico de Ward, en la discretización de variables continuas.

El análisis discriminante clasifica a los individuos en dos o más poblaciones según los valores de siniestralidad y posteriormente con un proceso de selección de predictores escogemos aquellos que “mejor” discriminan a las poblaciones

formadas *a priori*. Es usual distinguir dos poblaciones, la que no conlleva riesgo y la que conlleva riesgo (extrapolable al caso de más de dos poblaciones). Y lo usual es basarse en la experiencia del número de siniestros, así la población sin riesgo será la que no tenga siniestros y la de riesgo la que tenga al menos un siniestro. Al seleccionar las variables que mejor discriminan las poblaciones lo que cubrimos es la selección de variables de tarifa. Por lo tanto el análisis discriminante no es una técnica predictiva en el sentido de la tarificación, aunque sí nos permite clasificar a un nuevo individuo en una población concreta según los valores que tome en las variables de tarifa. En cualquier caso deberemos asignarle posteriormente de algún modo la prima pura.

En general, recomendamos la utilización de varios métodos para decidir finalmente un “buen” subconjunto de variables tarificadoras. Los métodos deberían coincidir aproximadamente en los resultados obtenidos, y es importante extraer de cada uno, no sólo el resultado final, sino la información que se desprende durante los procesos respecto a relaciones entre factores seleccionados y no seleccionados. Por su puesto, cada metodología incorpora sus hipótesis y con ellas ventajas e inconvenientes, así como un coste computacional mayor o menor.

A continuación damos detalle del *análisis de segmentación* y de los *modelos lineales generalizados* para pasar después a la aplicación. Respecto a los *modelos basados en distancias* nos referimos a Boj et al. (2000).

2. ANÁLISIS DE SEGMENTACIÓN (AS)

El Análisis de Segmentación es una técnica estadística de cluster jerárquico divisivo que trabaja sobre datos tipo regresión. Las variables independientes son categóricas, de tipo nominal u ordinal, y la variable dependiente puede ser cuantitativa o categórica. Se utiliza con fines exploratorios y descriptivos, con el objetivo básico de encontrar una clasificación de la población en grupos capaces de describir la variable dependiente de la mejor manera posible. El AS reduce la

complejidad de los problemas, rechazando tabulaciones cruzadas no significativas, detectando automáticamente los mejores predictores y creando subgrupos no cruzados potencialmente explicativos de la variable dependiente. Puede ser utilizado para la predicción a partir del árbol resultante. Adicionalmente es útil como paso previo en la aplicación de otras técnicas especializadas para datos cualitativos como los modelos logarítmico-lineales o el análisis de correspondencias. Nosotros lo utilizaremos para formar los grupos homogéneos de que parte un modelo de credibilidad en la estimación de primas, tal y como veremos en la aplicación.

En el trabajo se ha hecho uso del módulo CHAID (*CHi-squared Automatic Interaction Detector*) del programa SPSS [Magidson (1993)], y del CHAID y XAID (*eXtended Automatic Interaction Detector*) que encontramos programados en Fortran 77 en <http://www.stat.umn.edu/users/FIRM/index.html> bajo el nombre de FIRM (*Formal Inference-based Recursive Modeling*) [Hawkins (1997)]. Existen bastantes alternativas en programación de algoritmos de segmentación, pero lo usual es que los paquetes estadísticos no lo lleven implementado por defecto, si lo incorporan, son módulos separados, como por ejemplo el SPSS CHAID.

Distinguiremos el tipo de algoritmo dependiendo de dos factores: la naturaleza de la variable dependiente y la manera de calcular los *p-valores* de la fase de agrupación de categorías y de selección del mejor del predictor:

a) **SPSS CHAID nominal**: Realiza un CHAID para respuesta categórica nominal.

b) **SPSS CHAID ordinal**: Realiza un CHAID para respuesta categórica ordinal. En este caso se supone que las categorías de la variable dependiente están ordenadas y lo tiene en cuenta. Para su tratamiento, utiliza los modelos logarítmico-lineales *linear by linear* [Agresti (1984) pp. 79] que contemplan el hecho de que las dos variables analizadas (fila y columna) puedan estar ordenadas [Goodman (1979)]. Concretamente hace uso del modelo de efectos

fila (que supone que la variable ordinal es la columna –la respuesta– y que la fila es nominal –el factor–). Respecto a los *p-valores*, utiliza el contraste *Y-association* [Magidson (1992)] con el correspondiente *Likelihood Ratio*.

c) **FIRM nominal (CATFIRM)**: Realiza un CHAID para respuesta categórica nominal. Respecto al *p-valor* a diferencia del SPSS CHAID nominal, podemos predeterminar lo siguiente:

- Presta la posibilidad de una modificación para paliar el efecto de los ceros de las tablas dispersas: pregunta por el valor de la constante *A* a añadir al denominador del estadístico χ^2 . Si ponemos 0, nos dará la χ^2 de Pearson estándar. Si le damos un valor diferente de 0 calculará:

$$\frac{(\text{observadas} - \text{esperadas})^2}{\text{esperadas} + A} \quad (5)$$

de este modo se reducirá la significación e implicará estar menos inclinado a realizar particiones con un número pequeño de casos. En la práctica para tablas dispersas desalienta la formación de agrupaciones en las cuales las frecuencias esperadas sean pequeñas. El valor de *A* ha de ser pequeño (por ejemplo entre 0.5 y 1), sino podemos tener serias distorsiones en la significación estadística de los valores χ^2 .

- Además, se puede elegir entre una aproximación asintótica a la distribución chi-cuadrado para el cálculo del p-valor o una distribución exacta [Mielke y Berry (1985)]. Al utilizar la distribución exacta, que es recomendable, se consiguen *p-valores* más seguros y formales para tablas con frecuencias pequeñas, puesto que la distribución asintótica de estadístico χ^2 de Pearson, se deteriora cuando las frecuencias en la tabla de contingencia son pequeñas. Los procedimientos de segmentación no pueden utilizarse en conjuntos pequeños de datos, pues a medida que se va segmentando no sólo hay una pérdida de potencia del test, sino que la distribución del estadístico, que es válida sólo asintóticamente, se hace cada vez más imprecisa.

d) **FIRM continua (CONFIRM)**: Realiza un XAID para variable respuesta cuantitativa, esta opción **no** es la misma que la del SPSSCHAID ordinal con el test *Y-association*. A diferencia del CHAID, el XAID se basa en las F de *Fisher* y en las t de *Student* procedentes del análisis de la varianza a la hora de calcular los *p-valores*. En la F para la selección del mejor predictor y en la t para la agrupación de categorías de un predictor.

3. MODELOS LINEALES GENERALIZADOS (MLG)

Nos referimos a los libros Dobson (2001), McCullagh y Nelder (1989), y en castellano a López y López de la Manzanara (1986, pp. 125-145) para una descripción detallada del MLG. Y redireccionamos a: <http://www.statsci.org/glm/bibliog.html> donde encontramos bibliografía selecta.

Supongamos la variable aleatoria $\mathbf{Y}_{(n \times 1)}$, con (y_i) para $i = 1, 2, \dots, n$ observaciones independientes, que recogen la siniestralidad y juegan el papel de variable respuesta en la regresión, y supongamos los predictores o factores potenciales de la estructura de riesgo $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$, vectores $(n \times 1)$: (x_{ij}) para $i = 1, 2, \dots, n$ y $j = 1, 2, \dots, p$. Recordemos 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, $\varepsilon_i \sim N(0, \sigma^2)$. La relación lineal de la respuesta con la estructura sistemática dada por los predictores es:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i \quad (6)$$

Así, tenemos observaciones independientes, $y_i \sim N(\mu_i, \sigma^2)$, con esperanza:

$$E[y_i] = \mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (7)$$

y varianza constante:

$$Var[y_i] = \sigma^2 \quad (8)$$

En los MLG seguimos teniendo (y_i) para $i=1,2,\dots,n$ observaciones independientes de la respuesta, unos errores centrados $E[\varepsilon_i]=0$, y un predictor lineal determinista al que llamaremos $\eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$. Y las dos extensiones respecto al modelo clásico son,

- La respuesta está ligada con el predictor lineal a través de una función F :

$$E[y_i] = \mu_i = F(\eta_i) = F\left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}\right) \quad (9)$$

- La distribución de $\mathbf{Y}$ no tiene porqué ser la Normal, puede ser cualquier distribución derivada de la familia exponencial [McCullagh y Nelder (1989)], con:

$$Var[y_i] = \frac{\phi \cdot V(\mu_i)}{w_i} \quad (10)$$

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 .

Como vemos para tener despejada la respuesta, deberemos hacer la función inversa de F , $g = F^{-1}$, a la que llamaremos función link, pues es la que nos linkará la respuesta con el predictor lineal. A la función link g le exigiremos que sea monótona y diferenciable:

$$g(E[y_i]) = g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (11)$$

Así, en el MLG, tenemos dos extensiones, una respecto a la distribución del error plasmada en $Var(\mathbf{Y})$, que podrá proceder de cualquiera de las de la familia exponencial y no tiene por qué ser la Normal; y otra respecto a la función link, $g(\boldsymbol{\mu})$, que debe ser una función monótona diferenciable y no tiene por qué ser la Identidad. Los coeficientes de la regresión se estiman por criterios máximo-verosímiles. Cabe notar que suponiendo una distribución del error Normal y la Identidad como función link, obtenemos como caso particular la solución por MCO del modelo clásico.

Podemos combinar diferentes links con una función de error (y al revés), pero existen unos links “naturales” asociados a algunas distribuciones los cuales se denomina links canónicos. Estos links proporcionan unas características simplificadoras en la formulación, lo que no implica que vayan a ser siempre los más adecuados para unos datos determinados.

Si nos fijamos en los valores que puede tomar la respuesta, Y , será más apropiado utilizar una distribución u otra según analicemos las cuantías de los siniestros o el número de siniestros por póliza. Por ejemplo [Brockman y Wright (1992); Coutts (1984); Haberman y Renshaw (1998); Hipp (2000)], si estamos analizando la cuantía de los siniestros será apropiado utilizar una distribución Gamma o una Inversa Gaussiana preferiblemente a una Normal, que no toman valores negativos, y si estamos analizando el número de siniestros será más adecuado utilizar una distribución de Poisson, una Binomial o una Binomial Negativa.

En los MLG la variabilidad no explicada (VNE) por un modelo M (fijada una función link y una distribución del error) se plasma en la desvianza escalada $D^*(\cdot; \hat{\beta}_{MV})$. Si $L(\cdot; \hat{\beta}_{MV})$ denota la función de verosimilitud para el modelo M y L_{sat} la función de verosimilitud del modelo saturado¹, entonces:

$$D^*(\cdot; \hat{\beta}_{MV}) = -2 \cdot \log \left(\frac{L(\cdot; \hat{\beta}_{MV})}{L_{sat}} \right) \quad (12)$$

ésta se hace menor a mayor número de predictores incluidos en el modelo, hasta llegar a explicar la variabilidad total (VT) de los datos.

Históricamente, tanto en el campo actuarial como en otros, se han utilizado dos casos particulares de los MLG: el modelo clásico aditivo de link la identidad y de error la Normal [Lemaire (1979)], y el multiplicativo de link el logarítmico combinado con una distribución de Poisson [Zehnwirth (1994)]. Cabe notar que el hecho de que suponer una tarifa multiplicativa implica de algún modo “la

¹ El modelo saturado es el que tiene tantos parámetros como individuos, y por lo tanto se cumple que $\hat{\mu}_i = y_i$ para $i = 1, 2, \dots, n$

hipótesis de independencia entre las variables utilizadas” o lo que es equivalente que “no existen términos de interacción” o que “no hay asociación entre las variables”, lo que nos es de utilidad a la hora de seleccionar sólo los efectos principales de los factores, pues se obtiene antes una estimación ajustada de la siniestralidad sin necesidad de interacciones, además hay evidencias empíricas de que las estimaciones obtenidas tienden a ser casi siempre positivas a diferencia de lo que ocurre con una aditiva [Brockman y Wright (1992)].

Usualmente se indica que el modelo más adecuado para unos datos determinados será aquel que nos ofrezca una menor desviación. Como se intuye, tenemos diferentes maneras de reducirla si variamos la función link, la distribución del error y/o los diferentes factores de riesgo [Millenhall (1999)]. Puesto que nosotros tenemos como objetivo la selección de variables de tarifa, lo usual será fijar un link y una distribución de la familia exponencial, y a partir de aquí realizar el proceso de selección: seleccionaremos predictores para un modelo dado.

Supongamos dos modelos anidados $M_s \subset M_r$:

M_r : con $r + 1$ parámetros $(\beta_0, \beta_1, \dots, \beta_s, \beta_{s+1}, \dots, \beta_r)$

M_s : con $s + 1$ parámetros $(\beta_0, \beta_1, \dots, \beta_s)$

queremos testar si $r - s$ de los $r + 1$ parámetros son cero, i.e., si las variables explicativas $\mathbf{X}_{s+1}, \mathbf{X}_{s+2}, \dots, \mathbf{X}_r$ tienen una influencia significativa en la experiencia de siniestralidad esperada. En otras palabras, testamos si el modelo más pequeño, M_s , describe los datos de manera más adecuada que el modelo mayor, M_r . Sin pérdida de generalidad, testamos para los últimos $r - s$ de los $r + 1$ parámetros:

$$H_0 : \beta_{s+1} = \beta_{s+2} = \dots = \beta_r = 0$$

versus

$$H_1 : \text{no todos los } \beta_i \text{ (} i = s+1, s+2, \dots, r \text{) son } 0$$

Utilizaremos una distribución asintótica de la familia exponencial para el cálculo del p-valor asociado al contraste [Albrecht (1983)]: si $\hat{\beta}_s$ denota la estimación máximo verosímil y L_s la función de verosimilitud respecto al modelo M_s , y $\hat{\beta}_r$, L_r de igual modo para el modelo M_r , entonces el estadístico razón de verosimilitud, R , para este problema de hipótesis es $R = \frac{L_s(\cdot; \hat{\beta}_s)}{L_r(\cdot; \hat{\beta}_r)}$, y

$-2 \cdot \log(R)$ tiene una distribución asintótica Chi-cuadrado: χ_{r-s}^2 . Que reescrito en términos de desviaciones escaladas de los correspondientes modelos,

$$D_0^* = -2 \cdot \log\left(\frac{L_s(\cdot; \hat{\beta}_s)}{L_{sat}}\right) \sim \chi_{n-s-1}^2 \quad \text{y} \quad D_1^* = -2 \cdot \log\left(\frac{L_r(\cdot; \hat{\beta}_r)}{L_{sat}}\right) \sim \chi_{n-r-1}^2, \text{ tenemos}$$

$$\text{que, } D_0^* - D_1^* = -2 \cdot \log\left(\frac{L_s(\cdot; \hat{\beta}_s)}{L_r(\cdot; \hat{\beta}_r)}\right) \rightarrow \chi_{r-s}^2, \text{ y para la realización del contraste}$$

utilizaremos la distribución asintótica *F-de Fisher Snedecor* (división de chi-cuadrados) siguiente:

$$\frac{(D_0^* - D_1^*)/(r-s)}{D_1^*/(n-r-1)} \rightarrow F_{(r-s, n-r-1)} \quad (13)$$

Una vez sabemos cómo contrastar si un conjunto de coeficientes es significativo en la bondad del ajuste de la regresión, se trata de organizar un proceso de selección. Puesto que lo usual es disponer de un conjunto elevado de predictores, resulta casi imposible estudiar el ajuste de todos los modelos posibles, por lo que lo usual es optar por un proceso de introducción progresiva, por uno de eliminación progresiva o por uno paso a paso que combine los dos anteriores.

No existe ninguna manera óptima de validar el modelo resultante del proceso de selección, pero sí alguna más adecuada que otra. Una manera bastante adecuada es el siguiente contraste estadístico, consistente en la comprobación de si la variabilidad explicada por los predictores seleccionados es suficiente como para que la regresión en cuestión tenga poder predictivo. Para una regresión con p parámetros:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

y mediante el estadístico F , que no es más que el cociente entre la variabilidad total explicada sobre la total a explicar, tenemos en términos de desvianzas escaladas:

$$\frac{(D_0^* - D_1^*)/p}{D_1^*/(n - p - 1)} \rightarrow F_{(p, n-p-1)} \quad (14)$$

calculamos el p -valor asociado: si fijado α^* , $p\text{-valor} \geq \alpha^*$ aceptaremos la hipótesis nula, y significará que el conjunto de p variables no tiene poder predictivo en la regresión, y si $p\text{-valor} < \alpha^*$ el conjunto sí tendrá poder predictivo. No sólo podremos ver si aceptamos o rechazamos la hipótesis, sino también en qué medida. Además, puede ser una manera correcta de comparar dos regresiones con los mismos predictores pero con diferentes funciones link y/o distribuciones del error, nos quedaríamos con aquella que nos proporcionara un menor p -valor en el contraste.

4. APLICACIÓN EMPÍRICA

La aplicación ha sido realizada con los datos de Bermúdez y Pons (1997), que presentamos en la siguiente tabla:

		A: Antigüedad laboral											
		A1: Menos de 2 años				A2:Entre 2 y 10 años				A3:Más de 10 años			
E: Estado civil	E1: Aparejado	363.10	540.88	15.10	24.92	173.69	12.82	403.24	350.82	48.56	386.37	394.34	685.96
		267.58	523.66	371.92	94.36	63.47	142.28	221.20	497.90	57.06	103.11	145.99	162.40
		66.82	103.35	166.47	261.02	414.06	608.41	196.30	12.11	518.60	331.60	95.04	602.69
		93.05	140.82	312.18	109.34	205.19	311.53	400.06	122.62	287.28	105.88	110.22	203.00
		295.78	196.93	103.37	81.67	295.89	496.28	405.70	457.48	480.38	643.38	463.98	591.94
		59.36	298.89	452.58	150.11	536.37	268.45	203.34	168.02	423.87	724.22	129.01	239.66
		313.61	96.28	293.79	87.44	48.29	111.55	398.01	309.32	711.86	648.69	469.67	221.54
		395.56	187.40	286.09	137.64	74.51	168.30	90.92	381.98	259.00	783.54	516.97	182.72
		118.71	136.81	96.27	113.57	100.01	33.02	609.78	748.68	505.18	814.95	203.34	297.93
		51.60	332.67	403.13		20.97	84.40	366.05		177.51	251.56	509.60	556.78
										469.48	133.16	274.23	208.54
	E2: Separado Divorciado	54.92	109.34	268.77	47.18	513.14	97.04	316.37	483.78	215.06	414.72	238.92	157.63
		164.22	154.87	209.69	87.88	40.35	150.13	133.37	134.27	440.24	529.82	313.01	89.62
		396.50	184.61	80.80	321.16	529.67	332.02	532.93	147.06	48.16	471.13	388.53	36.85
		554.12	55.57	166.02	86.70	260.59	184.14	145.99	192.16	10.29	272.92	139.42	99.73
		163.77	173.10	148.62	425.24	425.27	102.29	42.67	149.56	382.19	520.18	77.37	290.97
		52.67	7.06	71.84	300.52	163.77	398.88	157.20	309.78	374.19	258.75	314.01	78.43
		77.99	486.38	84.18	69.13	640.47	77.34	202.68	98.96	563.36	232.75	60.60	
		65.40	176.24	254.47	141.89	53.00	47.94	331.85	111.04	271.07	94.52	134.11	
		102.35	62.66	97.79	134.91	272.63	319.02	27.32	166.00	37.10	421.51	163.87	
		107.72	67.74	80.97	118.57	81.67	97.72	563.61	56.23	36.43	487.45	290.97	
	E3: Soltero	75.30	156.06	242.69	265.84	158.72	101.68	165.78	47.99	293.78	155.79	295.07	157.63
		7.72	43.80	113.68	231.10	616.01	358.21	477.92	638.10	185.62	34.23	447.74	145.27
		133.02	484.22	196.30	10.29	405.37	56.18	107.48	97.31	488.89	170.73	242.96	314.85
		202.63	90.07	380.65	325.23	53.58	513.20	253.04	133.48	314.31	252.84	298.26	400.06
		223.68	27.06	388.53	68.63	123.04	82.79	406.20	113.57	46.47	84.31	222.98	56.56
		199.95	122.02	326.63	87.61	34.98	151.52	58.31	177.84	49.86	358.32	219.79	210.21
		49.62	80.97	91.41	84.79	316.44	342.59	77.93	406.86	483.03	287.12	539.43	251.92
		63.97	61.24	326.63	27.48	323.07	402.08	103.37	225.10	73.09	411.58	22.68	561.44
		293.02	133.68	106.38	314.31	364.56	413.70	400.06	234.76	527.52	434.72	420.87	
218.24		445.67	266.48	248.60	221.55	257.69	550.82	406.86	214.25	502.44	145.92		
				147.45	56.09	25.05			37.12	247.09	117.82		

Se trata de una cartera de clientes de una entidad financiera que no pudieron hacer frente a la devolución del préstamo contraído. El tamaño total de la muestra es de 401 individuos. Sobre cada individuo se ha observado: la experiencia de siniestralidad plasmada en las pérdidas monetarias ocasionadas por los clientes a la entidad, de naturaleza continua, y los predictores antigüedad en el puesto laboral que ya encontramos discretizada en tres categorías: A1 (< 2 años), A2 (entre 2 y 10 años), A3 (> 10 años), y por lo tanto de naturaleza

categorica ordinal y estado civil del cliente de naturaleza categorica nominal con tres clases: E1 (aparejados), E2 (separados o divorciados), E3 (solteros). Esta información la podemos resumir de manera agregada mediante una tabla cruzada de $3 \times 3 = 9$ combinaciones. Una primera estimación posible del riesgo es la media aritmética de todos los individuos, o bien realizar la media de la tabla cruzada de cada combinación. En la siguiente tabla se plasman las medias y el número de casos correspondientes a cada combinación y a toda la muestra:

	A1	A2	A3	
E1	208.82 39	269.57 39	366.61 44	
E2	172.05 54	232.67 53	253.22 48	
E3	180.38 40	246.71 43	261.58 41	
				242.17 401

Antes de pasar a aplicar las tres metodologías realizaremos un estudio de asociación de variables. Nos interesa ver la relación de las cuantías con cada uno de los dos predictores y la relación entre ambos. Respecto a cuantías y predictores, utilizando (2), obtenemos:

- Cuantías y Estado: $\eta^2 = 0.028294$

- Cuantías y Antigüedad: $\eta^2 = 0.066174$

Observamos que la relación no es alta, pero que la antigüedad está mayormente relacionada. Y respecto a predictor con predictor, utilizando (4), obtenemos

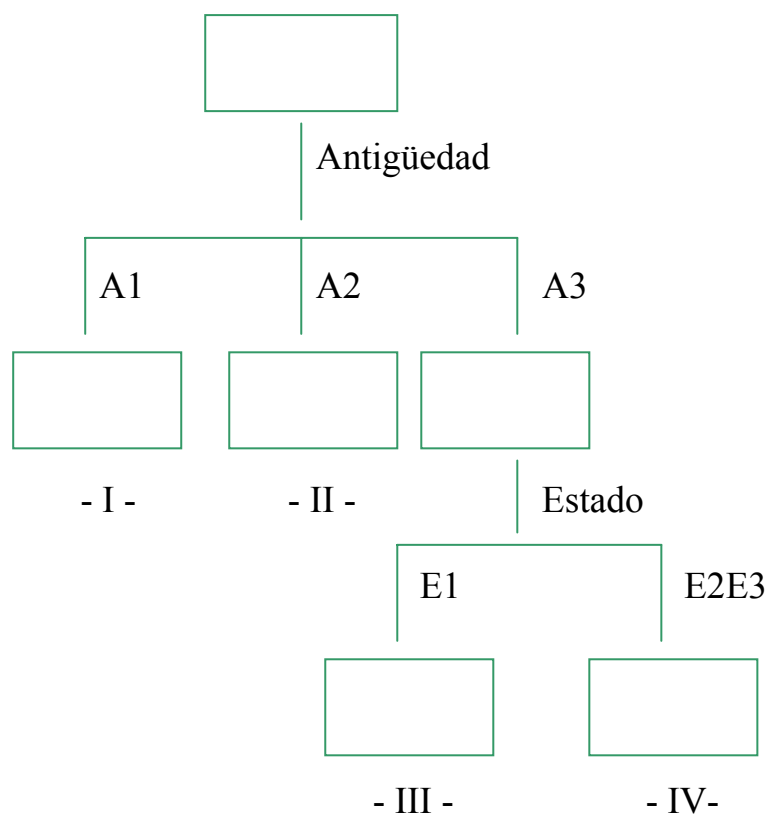
- Estado y Antigüedad: $C^2 = 0.905 / (2 \times 401) = 0.001284$

Vemos cómo los predictores no están relacionados entre si. Esto será un punto positivo para poder aplicar correctamente el análisis de segmentación, pues si los predictores estuvieran relacionados correríamos el peligro de topar con la “*paradoja de Simpson*”.

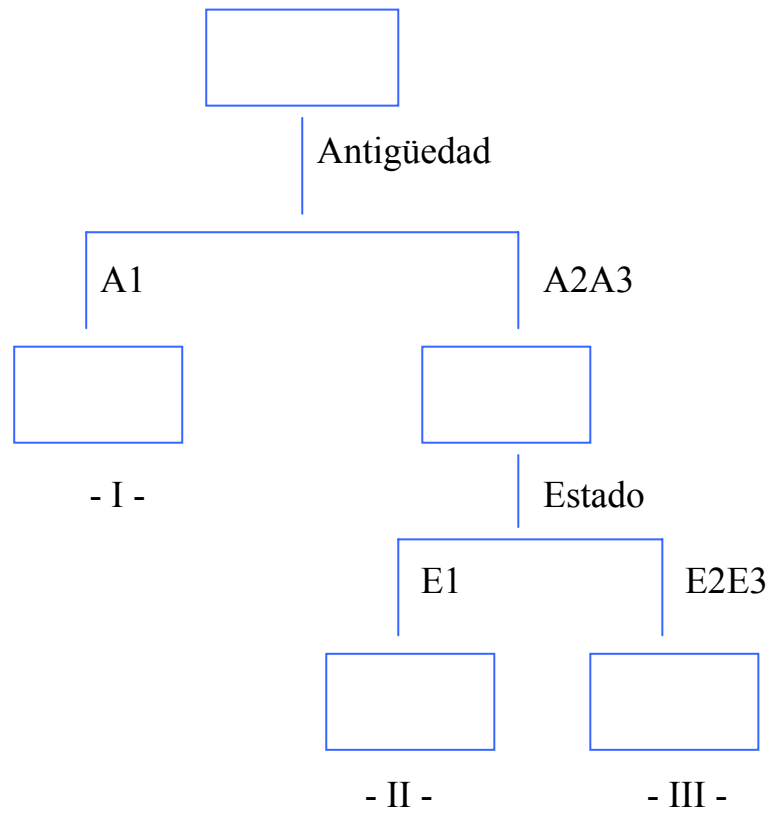
4.1. Análisis de segmentación:

Puesto que en nuestro caso la respuesta se corresponde con las cuantías de los siniestros y es continua, para la aplicación de CHAID hemos procedido a categorizarla haciendo uso del método de Ward, discretizando para 4, 10 y 31 categorías. La antigüedad en todos los casos ha sido definida como monótona y el estado como libre. El nivel de significación permitido para fase de agrupación de categorías y de selección del mejor predictor ha sido del 5%. El tamaño mínimo para analizar un nodo ha sido de 50 individuos. Obteniendo los siguientes árboles de segmentación:

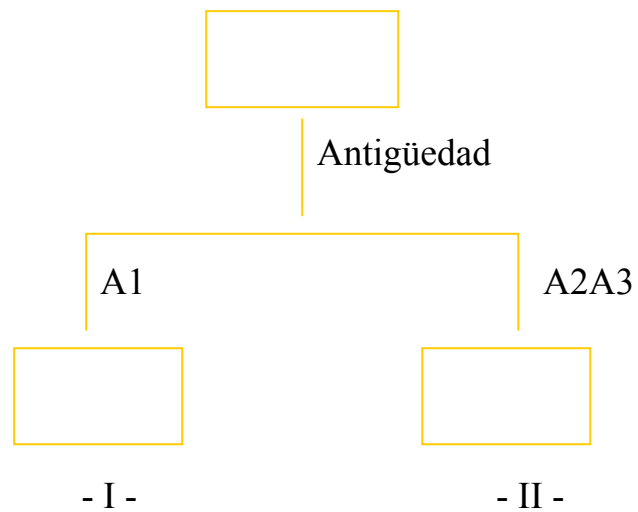
Árbol 1



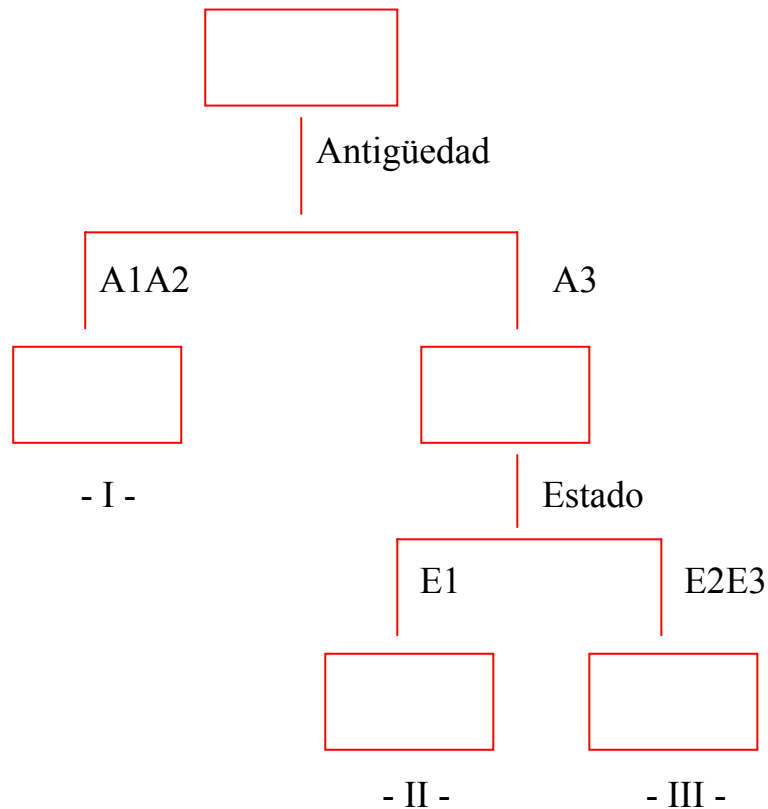
Árbol 2



Árbol 3



Árbol 4



El árbol 1 se corresponde con

- CONFIRM para variable respuesta continua
- CHAID ordinal con Ward 10 y 31 categorías

El árbol 2 se corresponde con

- CHAID ordinal con Ward 4 categorías
- CHAID nominal con Ward 4 y 10 categorías
- CATFIRM con Ward 4 categorías y $A = 0$ y 0.5
- CATFIRM con Ward 10 categorías y $A = 0$

El árbol 3 se corresponde con

- CATFIRM con Ward 10 categorías y $A = 0.5$

El árbol 4 se corresponde con

- CHAID nominal con Ward 31 categorías

Todo este repertorio nos sirve para realizar un estudio comparativo más amplio del efecto de discretizar a la respuesta en un número diferente de clases, de dar diferentes valores a los parámetros implicados en los algoritmos y de ver hasta qué punto es correcto considerar a la respuesta continua como categórica entre otras cosas. Pero nuestro interés es el de seleccionar el árbol más correcto, que como es de intuir es el que se corresponde al tratamiento de respuesta continua (CONFIRM) y como mucho al de categórica ordinal (CHAID ordinal) con el número máximo de categorías que permite el programa. Así, para este estudio nos quedaremos con el árbol 1. Sin embargo el resto nos sirve a modo informativo.

4.1.1. CHAID y XAID:

Una vez tenemos los grupos exhaustivos y homogéneos respecto al riesgo, la primera posibilidad en la estimación de la siniestralidad, sería la media aritmética de cada grupo:

	A1	A2	A3
E1	I 185.33 133	II 247.80 135	III 366.61 44
E2			IV 257.07 89
E3			
242.17 401			

4.1.2. Modelo de credibilidad:

Como ya hemos comentado, el AS nos sirve como paso previo para otras técnicas. Nosotros lo utilizaremos para configurar los grupos de partida de un modelo de credibilidad en la estimación del riesgo. Concretamente para el modelo de credibilidad de *Bühlmann-Straub*. El modelo aplica un promedio

ponderado de la media de la experiencia individual y de la experiencia de todo el colectivo:

$$z_j \cdot \bar{x}_j + (1 - z_j) \cdot \bar{x} \text{ para } j = 1, 2, 3, 4$$

donde z_j es el factor de credibilidad, que es una variable que recoge la ponderación de la media de cada grupo. Éste depende del tamaño del grupo, de la varianza interna del grupo y de la varianza entre grupos: $0 \leq z_j \leq 1$. Valores altos de z_j implican una significación alta del grupo por separado. Un buen resumen de las características de este modelo de credibilidad, junto con abundantes citas, puede encontrarse en Pons (1995). Los resultantes de la aplicación para cada segmento son:

Segmento	Factor de credibilidad
I:	$z_1 = 0.9478470046$
II:	$z_2 = 0.9485799125$
III:	$z_3 = 0.8573988926$
IV:	$z_4 = 0.9240224682$

Observamos que todos los factores de credibilidad son muy altos, cercanos 1, esto indica que los 4 grupos de tarifa son homogéneos internamente y diferentes entre si. Las diferencias entre los factores de credibilidad de cada grupo son debidas a los distintos tamaños. Empíricamente hemos comprobado que aplicando el modelo de *Bühlmann-Straub* a cualquier otra agrupación posible de las 9 combinaciones iniciales da en su conjunto factores de credibilidad inferiores a los aquí obtenidos. Empíricamente, la aplicación de otros modelos de credibilidad, como el *two-way* o el *jerárquico de Jewell* [Bermúdez y Pons (1997)], proporcionan una tabla cruzada cuyos factores de credibilidad son en su

conjunto también inferiores a los aquí obtenidos. La estimación con este modelo para cada grupo de tarifa es:

	A1	A2	A3	
E1	I 185.30 133	II 247.51 135	III 348.86 44	
E2			IV 255.93 89	
E3				
				242.17 401

4.2. Modelos de regresión basada en distancias:

Nos referimos a Boj et al. (2000) para el detalle del proceso de selección con el Modelo de Regresión basada en Distancias (RBD).

A groso modo, el modelo de RBD consiste en proyectar la respuesta sobre una configuración Euclídea obtenida a partir de una matriz de distancias entre individuos calculada en el espacio predictor. Para estimar la distribución de los estadísticos implicados en el proceso de selección de predictores, se requiere de la metodología *bootstrap*. En la referencia citada encontramos ya realizada la selección para los datos de la aplicación. Cabe notar que como función de distancias se ha utilizado el índice de similitud de Sokal-Michener en su extensión para variables cualitativas [Andenberg (1973) p.123, tabla 5.5].

Como resultado del proceso de selección paso a paso se ha obtenido que ambos factores son relevantes, entrando primeramente la antigüedad. Aquí plasmamos la tabla de estimaciones con ambos efectos principales:

	A1	A2	A3
E1	227.09	289.65	332.61
E2	163.84	226.40	269.36
E3	173.64	236.21	279.16
			242.17

4.3. Modelos Lineales Generalizados:

Nos referimos a Boj et al. (2002) para el detalle del Modelo Lineal Generalizado (MLG) y su utilización en tarificación *no vida*. Detallamos en la siguiente tabla las diferentes combinaciones de distribuciones del error y funciones link que hemos utilizado con estos datos referentes a cuantías de siniestros:

Error:	Link:	Canónico:
Gaussiana, $V(\mu_i)=1$	Identidad, $g(\mu_i) = \mu_i$	Sí
Gaussiana, $V(\mu_i)=1$	Logarítmico, $g(\mu_i) = \log(\mu_i)$	No
Gamma, $V(\mu_i) = \mu_i^2$	Identidad, $g(\mu_i) = \mu_i$	No
Gamma, $V(\mu_i) = \mu_i^2$	Logarítmico, $g(\mu_i) = \log(\mu_i)$	Sí
Inversa-Gaussiana, $V(\mu_i) = \mu_i^3$	Identidad, $g(\mu_i) = \mu_i$	No
Inversa-Gaussiana, $V(\mu_i) = \mu_i^3$	Logarítmico, $g(\mu_i) = \log(\mu_i)$	No
Inversa-Gaussiana, $V(\mu_i) = \mu_i^3$	$g(\mu_i) = 1/\mu_i^2$	Sí

Como veremos con posterioridad a la selección de variables, el modelo que nos ofrecerá mayor poder predictivo para estos datos haciendo uso de los efectos principales de los predictores, será el de la combinación Normal | Logarítmico. Puesto que en todos los casos las variables seleccionadas finalmente serán las mismas, tan solo a modo de ejemplo detallaremos el proceso para tal modelo.

Proceso de selección de predictores teniendo en cuenta los efectos principales y las correspondientes interacciones para el modelo Normal | Logarítmico:

- para el Estado, utilizamos dos variables binarias

$$X_{E1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in E1 \end{cases} \quad X_{E2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in E2 \end{cases}$$

- para la Antigüedad, utilizamos dos variables binarias

$$X_{A1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A1 \end{cases} \quad X_{A2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A2 \end{cases}$$

- y para su interacción,

$$X_{A1E1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A1 \cap E1 \end{cases} \quad X_{A1E2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A1 \cap E2 \end{cases}$$

$$X_{A2E1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A2 \cap E1 \end{cases} \quad X_{A2E2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A2 \cap E2 \end{cases}$$

Proceso de introducción progresiva:

<u>Variable</u>	<u>p-valor</u>	<u>p-valor X(1)</u>	<u>p-valor X(1)X(2)</u>
E	0.003307	0.001841	-----
A	0.000001	-----	-----
A:E	0.004719	0.689690	0.615348
	X(1) = A	X(2) = E	X(3) = A:E

Proceso de eliminación progresiva:

Variable	<i>p</i> -valor	<i>p</i> -valor / X[1]	<i>p</i> -valor / X[1]X[2]
E	0.001587	0.001841	-----
A	0.062327	0.000001	0.000001
A:E	0.615348	-----	-----
	X[1] = A:E	X[2] = E	X[3] = A

Si fijamos por ejemplo $\alpha^* = 0.05$ como nivel de significación permitido, nos quedaremos con los efectos principales, $X_{A1}, X_{A2}, X_{E1}, X_{E2}$. Veamos el poder predictivo de todas las combinaciones dados los efectos principales:

Modelo	Link:	Canónico:
Gaussiana, Identidad**	p-valor = 0	Sí
Gaussiana, Logarítmico*	p-valor = 0	No
Gamma, Identidad	p-valor = 0.000009	No
Gamma, Logarítmico	p-valor = 0.000549	Sí
Inversa-Gaussiana, Identidad	p-valor = 0.008463	No
Inversa-Gaussiana, Logarítmico	p-valor = 0.545475	No
Inversa-Gaussiana, $g(\mu_i) = 1/\mu_i^2$	p-valor = 0.508749	Sí

* $F_{4,396} = 10.4415$, ** $F_{4,396} = 9.9838$.

Como vemos el “mejor” será el Normal | Logarítmico. De la tabla podemos sacar alguna otra conclusión: Estamos en el caso de cuantías por siniestro, por lo que en general es recomendada la utilización de distribuciones con el rango de Y positivo, como es el caso de la Gamma y de la Inversa-Gaussiana, a ser posible con un efecto multiplicativo, link logarítmico, que ayuda a no obtener

estimaciones negativas para cualquier otra distribución. En este caso la Normal no nos ha proporcionado estimaciones negativas y ha sido la de mayor poder predictivo, por lo que en general no nos limitaremos a aceptar la utilización de la Gamma o la Inversa-Gaussiana. Comprobamos como para una distribución del error y unas variables dadas, no siempre el link canónico es el que mejor ajusta los datos, por ejemplo para la Normal el mejor es el Logarítmico, y para la Gamma y la Inversa-Gaussiana el mejor es la Identidad, y en ningún caso es el canónico. Éstos links proporcionan propiedades simplificadoras en la formulación genérica de la familia exponencial, cosa que no implica que sean la combinación más adecuada para unos datos determinados.

En las siguientes tablas detallamos las estimaciones con los efectos principales y datos desagregados para las diferentes opciones:

Gaussiana | Identidad:

	A1	A2	A3	
E1	227.09	289.65	332.61	
E2	163.84	226.40	269.36	
E3	173.64	236.21	279.16	
				242.17

Gaussiana | Logarítmico:

	A1	A2	A3	
E1	218.17	291.04	345.22	
E2	167.52	223.47	265.07	
E3	174.93	233.37	276.81	
				242.17

Gamma | Identidad:

	A1	A2	A3
E1	222.77	285.00	322.82
E2	168.09	230.32	268.15
E3	177.66	239.90	277.72
			242.17

Gamma | Logarítmico:

	A1	A2	A3
E1	216.17	289.78	335.74
E2	169.78	227.58	263.68
E3	177.69	238.20	275.97
			242.17

Inversa-Gaussiana | Identidad:

	A1	A2	A3
E1	220.33	282.56	318.63
E2	169.29	231.52	267.59
E3	178.71	240.94	277.01
			242.17

Inversa-Gaussiana | Logarítmico:

	A1	A2	A3	
E1	214.96	288.68	331.39	
E2	170.48	228.94	262.81	
E3	178.57	239.81	275.29	
				242.17

Inversa-Gaussiana | $g(\mu_i) = 1/\mu_i^2$:

	A1	A2	A3	
E1	201.59	288.56	356.19	
E2	176.51	226.46	255.06	
E3	181.43	237.13	270.60	
				242.17

Para cualquiera de las combinaciones el modelo con los efectos principales y las interacciones proporciona como estimación las medias de cada clase, al igual que para el basado en distancias.

La combinación de la Normal con el link Logarítmico, nos proporciona un modelo de efecto multiplicativo, para el modelo con los efectos principales tenemos las siguientes estimaciones de los coeficientes:

Coefficiente:	t-valor:
$\hat{\beta}_0 = 5.6233$	77.8052
$\hat{\beta}_{A1} = -0.4589$	-5.1128
$\hat{\beta}_{A2} = -0.1707$	-2.2999
$\hat{\beta}_{E1} = 0.2208$	2.7257
$\hat{\beta}_{E2} = -0.0433$	-0.4999

La interpretación, al tratarse de un modelo multiplicativo es la siguiente:

$$\log(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_{A1} \cdot X_{A1} + \hat{\beta}_{A2} \cdot X_{A2} + \hat{\beta}_{E1} \cdot X_{E1} + \hat{\beta}_{E2} \cdot X_{E2}$$

$$\log(\hat{\mu}_i) = 5.6233 - 0.4589 \cdot X_{A1} - 0.1707 \cdot X_{A2} + 0.2208 \cdot X_{E1} - 0.0433 \cdot X_{E2}$$

$$\hat{\mu}_i = e^{(5.6233 - 0.4589 \cdot X_{A1} - 0.1707 \cdot X_{A2} + 0.2208 \cdot X_{E1} - 0.0433 \cdot X_{E2})}$$

$$\hat{\mu}_i = e^{5.6233} \cdot e^{-0.4589 \cdot X_{A1}} \cdot e^{-0.1707 \cdot X_{A2}} \cdot e^{0.2208 \cdot X_{E1}} \cdot e^{-0.0433 \cdot X_{E2}}$$

$$\hat{\mu}_i = 276.81 \times 0.6320^{X_{A1}} \times 0.8431^{X_{A2}} \times 1.2471^{X_{E1}} \times 0.9576^{X_{E2}}$$

Si nos fijamos en los t -valores vemos cómo el de menor importancia es el de la clase 2 del estado, y que la variable que marca la pauta es la antigüedad.

REFERENCIAS

- Agresti, A. (1984), “*Analysis of ordinal categorical data*”, New York: Wiley.
- Albrecht, P. (1983), “Parametric multiple regression risk models: Connections with tariffication, especially in motor insurance”, *Insurance: Mathematics and Economics* 2, pp. 113-117.
- Andenberg, M. R. (1973), “*Cluster analysis for applications*”, New York: Academic Press, pp. 31-51.

- Bermúdez, Ll. y M. A. Pons (1997), “Determinación del riesgo de impago en una cartera de préstamos según el tipo de cliente”, *Matemática de las Operaciones Financieras 97*, Publicaciones de la Universidad de Barcelona, pp. 291-308.
- Boj, E., Claramunt, M. M. y J. Fortiana (2000), “Una alternativa en la selección de los factores de riesgo a utilizar en el cálculo de primas”, *Anales del Instituto de Actuarios Españoles*, Tercera Época, nº 6, pp. 11-35.
- Boj, E., Claramunt, M. M. y J. Fortiana (2002), “Selección exacta de variables de tarifa con MLG: Estudio bootstrap”, *VI Congreso de Matemática Financiera y Actuarial y V Italian-Spanish Conference of Financial Mathematics*, Valencia 20-22 de junio de 2002.
- Brockman, M. J. y T. S. Wright (1992), “Statistical Motor Rating: Making Effective Use of your Data”, *Journal of the Institute of Actuaries* 119:3, pp. 457-543.
- Calatayud, J. y I. Martínez (1997), “Pricing No Vida”, en: *Manual de banca, finanzas y seguros*, Ediciones gestión 2000 S.A., Colección Universitaria Eserp.
- Campbel, M. (1986), “An integrated system for estimating the risk premium of individual car models in motor insurance”, *ASTIN Bulletin* 16:2, pp. 165-184.
- Coutts, S. M. (1984), “Motor Insurance Rating - an Actuarial Approach”. *Journal of the Institute of Actuaries* 111, pp. 87-148.
- Cuadras, C. M. y P. Sánchez (1997), “*Relacions multivariants entre dos conjunts de variables*”. Curs de Doctorat del Departament d’Estadística de la Universitat de Barcelona.
- Dobson, A. J. (2001), “*An Introduction to Generalized Linear Models*”, Second Edition, London: Chapman & Hall.

- Goodman, L. A. (1979), "Simple models for the analysis of association in cross-classifications having ordered categories", *Journal of the American Statistical Association* 74, pp. 537-552.
- Haberman, S. y A. E. Renshaw (1998), "Actuarial applications of Generalized Linear Models", en: *Statistics in Finance*, Hand, D. J. and S. D. Jacka (eds), Arnold Applications of Statistics, London-Sydney-Auckland, pp. 42-65.
- Hartigan, J. A. (1975), "*Clustering Algorithms*", New York: Wiley.
- Hawkins, D. M. y G. V. Kass (1982), "*Topics in Applied Multivariate Analysis*", Ed. D M Hawkins, Cambridge University Press, pp. 269-302.
- Hawkins, D. M. (1997), "*FIRM: Formal Inference-based Recursive Modeling*", Technical Report Number 546, School of Statistics, University of Minnesota, <http://www.stat.umn.edu/users/FIRM/index.html>
- Hipp, Ch. (2000), "*Least Squares, Generalized Linear Models and Credibility Theory with applications to Tariffication*", Working Paper, University of Karlsruhe.
- López, M. y J. López de la Manzanara (1996), "*Estadística para actuarios*", Editorial MAPFRE, S.A. Madrid.
- Lemaire, J. (1979), "Selection procedures of regression analysis applied to automobile insurance. Part II: Sample Inquiry and Underwriting Applications", *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker* 79:1, pp. 65-72.
- McCullagh, P. y J. A. Nelder (1989), "*Generalized Linear Models*", Second Edition, London: Chapman & Hall.
- Magidson, J. (1993), "*SPSS for Windows: Chi-square Automatic Interaction Detection CHAID. Release 6.0*", SPSS Inc. Chicago.
- Magidson, J. (1992), "Chi-squared analysis of a scalable dependent variable", in: *Proceedings of the 1992 Annual Meeting of the American Statistical Association, Educational Statistics Section*.

- Mielke, P. W. y K. J. Berry (1985), “Non-asymptotic inferences based on the chi-square statistic for r by c contingency tables”, *Journal of Statistical Planning and Inference* 12, pp. 41-45.
- Millenhall, S. J. (1999), “A systematic relationship between minimum bias and generalized linear models”, *Proceedings of the Casualty Actuarial Society* 86, pp. 393-487.
- Pérez Torres, J. L. (2001), “*Conociendo el seguro*”, Editorial UMESER, S.A. Barcelona.
- Pons, M. A. (1995), “*Introducción a la teoría de la credibilidad*”, Colección de publicaciones del Departamento de Matemática Económica, Financiera y Actuarial de la Universidad de Barcelona, nº 30.
- Sánchez, M. (1997), “Segmentación de carteras. Aplicación al seguro de responsabilidad civil del automóvil”, *Actuarios*, diciembre96/enero-febrero97, pp. 62-65.
- Sierra, M. A. (1986), “*Análisis multivariante: teoría y aplicaciones en economía*”, Ediser, DL. Barcelona.
- UNESPA (1995), “*Estudio de la siniestralidad en la cartera de turismos. Modalidad Responsabilidad Civil. Segmentación año 1994*”, Dirección de Estudios.
- Ward, J. H. (1963), “Hierarchical grouping to optimise an objective function”, *Journal of the American Statistical Association*, pp. 236-244.
- Zehnwirth, B. (1994), “Ratemaking: from Bailey and Simon (1960) to Generalised Linear Regression Models”, *Casualty Actuarial Society Winter Forum*, pp. 615-659.