

VEGAS MONTANER, A. (1993) "Métodos estadísticos para el cálculo y comprobación de la provisión técnica para prestaciones". Previsión y Seguro nº 28

VILAR ZANÓN, J.L., VEGAS ASENSIO, J. (1998) "Estimación de la provisión de estabilización y del recargo técnico sobre primas a partir del ajuste de una distribución de poisson compuesta para el número de siniestros". Anales I.A.E. 3ª Época nº 4

## HERRAMIENTAS ESTADÍSTICAS PARA EL ESTUDIO DE PERFILES DE RIESGO

Eva Boj del Val<sup>1</sup>, M<sup>a</sup> Mercè Claramunt Bielsa<sup>2</sup> y Josep Fortiana Gregori<sup>3</sup>

### RESUMEN

En este trabajo se ilustran, a modo práctico, la utilización de tres herramientas que permiten al actuario definir los grupos de tarifa y estimar las primas en un proceso de tarificación *a priori* no vida. La primera es el análisis de segmentación (*CHAID* y *XAID*) utilizado inicialmente por *UNESPA* en 1997 en su cartera común de automóviles. La segunda es un proceso de selección de predictores paso a paso con el modelo de regresión basada en distancias. Y la tercera es un proceso con los modelos lineales generalizados que representan la técnica más actual de la bibliografía actuarial. De estos últimos, combinando diferentes funciones link y distribuciones del error, se desprenden los clásicos modelos aditivo y multiplicativo, de los cuales se interpreta el significado.

**PALABRAS CLAVE:** Análisis de segmentación, modelos de credibilidad, modelos basados en distancias, modelos lineales generalizados, perfiles de riesgo, tarificación *a priori*, seguros no vida.

<sup>1</sup> Profesora Ayudante del Departament de Matemàtica Econòmica, Financera i Actuarial de la Universitat de Barcelona

<sup>2</sup> Profesora Titular de Universidad del Departament de Matemàtica Econòmica, Financera i Actuarial de la Universitat de Barcelona

<sup>3</sup> Profesor Titular de Universidad del Departament d'Estadística de la Universitat de Barcelona

# 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,  $Y_{(n \times 1)}$ , y cada factor de riesgo univariante,  $X_{(n \times 1)}$ , 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 i Sánchez (1997)]:

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

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

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

$$\eta^2 = 1 - \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - y_i)^2}{\sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - y)^2} = 1 - \frac{Q_d}{Q_t} = \frac{Q_c}{Q_t}$$

- **Categorica-categorica:** partimos de dos variables,  $X$  e  $Y$ , 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:

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

y como medida de asociación se propone el siguiente coeficiente

$$C^2 = \frac{\chi^2}{r \times n_{..}}$$

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 categoricas 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**. No podemos limitarnos a escoger las variables que una a una están más asociadas con el riesgo, pues es posible que, dentro del conjunto elegido, alguna de ellas se vuelva no significativa. 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 and 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 (univariante o multivariante) 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 [Campbell (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 de manera positiva 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 si 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  $\chi^2$ . Si ponemos 0, nos dará la  $\chi^2$  de Pearson estándar. Si le damos un valor diferente de 0 calculará:

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

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  $\chi^2$ .

- Además, se puede elegir entre una aproximación asintótica a la distribución chi-cuadrado para el cálculo del *p*-valor o una distribución exacta [Mielke and 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  $\chi^2$  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), MacCullagh and 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  $\varepsilon_i$  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:

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

Así, tenemos observaciones independientes  $y_i \sim N(\mu_i, \sigma^2)$ : con esperanza  $E[y_i] = \mu_i$  y varianza constante  $Var[y_i] = \sigma^2$ . Tras la estimación por mínimos cuadrados ordinarios (MCO) de los coeficientes tenemos:

$$\hat{\mu}_i = E[y_i] = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$$

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

predicador lineal determinista al que llamaremos  $\eta_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij}$ . Y  
 as dos extensiones respecto al modelo clásico son,

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

$$y_i = F(\eta_i) + \varepsilon_i = F\left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij}\right) + \varepsilon_i$$

La distribución de  $\mathbf{Y}$  no tiene porqué ser la Normal, puede ser  
 cualquier distribución derivada de la familia exponencial de  
 MacCullagh and Nelder, con:

$$\hat{\mu}_i = E[y_i] = F\left(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}\right)$$

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

donde  $\phi$  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:

$$g(\hat{\mu}_i) = g(E[y_i]) = \hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x_{ij}$$

así, en el MLG, tenemos dos extensiones, una respecto a la  
 distribución del error plasmada en  $\text{Var}(\mathbf{Y})$ , 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<sup>4</sup>. 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<sup>5</sup>,  
 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,  $\mathbf{Y}$ , 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 and Wright (1992); Coutts (1984); Haberman and  
 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 u  
 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  $D^*(\mathbf{Y}, \hat{\boldsymbol{\mu}}_{MV})$ . Si  $L(\mathbf{Y}, \hat{\boldsymbol{\mu}}_{MV})$  denota la función de

<sup>4</sup> Si suponemos una Ley de probabilidad para cada  $y_i$  perteneciente a una familia  
 exponencial de MacCullagh-Nelder, la función de verosimilitud en general será de  
 la forma:  $L(y; \theta, \phi) = \exp\left\{\frac{\theta \cdot y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$ , donde  $a$ ,  $b$  y  $c$  son funciones,  $\theta$  es el  
 parámetro canónico y  $\phi$  el de dispersión. Es fácil ver que en las formulación más  
 general:  $\mu = E[y] = b'(\theta)$  y  $\text{Var}[y] = a(\phi) \cdot b''(\theta)$ .

<sup>5</sup>  $g = \theta$ , y por lo tanto  $\theta(\mu) = \eta$ .

verosimilitud para el modelo M y  $L_{sat}$  la función de verosimilitud del modelo saturado<sup>6</sup>, entonces:

$$D^*(Y, \hat{\mu}_{MV}) = -2 \cdot \log \left( \frac{L(Y, \hat{\mu}_{MV})}{L_{sat}} \right)$$

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 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 and 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 un error y a partir de aquí realizar el proceso de selección: elegiremos predictores para un modelo dado.

<sup>6</sup> El modelo saturado es el que tiene tantos parámetros como individuos, y cumple  $\hat{\mu}_i = y_i$  para  $i = 1, 2, \dots, n$

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  $X_{s+1}, X_{s+2}, \dots, X_r$  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
- $H_1$ : no todos los  $\beta_i$  ( $i = s + 1, s + 2, \dots, r$ ) 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:  $\chi^2_{r-s}$ . Que reescrito en términos de desviaciones escaladas de los correspondientes modelos,

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

tenemos que,

$$D_0^* - D_1^* = -2 \cdot \log \left( \frac{L_s(\cdot; \hat{\beta}_s)}{L_r(\cdot; \hat{\beta}_r)} \right) \rightarrow \chi^2_{r-s}$$

para la realización del contraste utilizaremos la distribución *F*-de Fisher Snedecor (división de chi-cuadrados) siguiente:

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

La 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 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:

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

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 desviaciones escaladas:

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

Calculamos el *p*-valor asociado: si fijado  $\alpha^*$ , *p*-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*-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 aquélla 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			
E1: Aparejado		363,1	540,9	15,1	24,9	173,7	12,8	403,2	350,8	48,6	386,4	394,3	686,0
		267,6	523,7	371,9	94,4	63,5	142,3	221,2	497,9	57,1	103,1	146,0	162,4
		66,8	103,4	166,5	261,0	414,1	608,4	196,3	12,1	518,6	331,6	95,0	602,7
		93,1	140,8	312,2	109,3	205,2	311,5	400,1	122,6	287,3	105,9	110,2	203,0
		295,8	196,9	103,4	81,7	295,9	496,3	405,7	457,5	480,4	643,4	464,0	591,9
		59,4	298,9	452,6	150,1	536,4	268,5	203,3	168,0	423,9	724,2	129,0	239,7
		313,6	96,3	293,8	87,4	48,3	111,6	398,0	309,3	711,9	648,7	469,7	221,5
		395,6	187,4	286,1	137,6	74,5	168,3	90,9	382,0	259,0	783,5	517,0	182,7
		118,7	136,8	96,3	113,6	100,0	33,0	609,8	748,7	505,2	815,0	203,3	297,9
		51,6	332,7	403,1		21,0	84,4	366,1		177,5	251,6	509,6	556,8
										469,5	133,2	274,2	208,5
	E2: Separado Divorciado		54,9	109,3	268,8	47,2	513,1	97,0	316,4	483,8	215,1	414,7	238,9
		164,2	154,9	209,7	87,9	40,4	150,1	133,4	134,3	440,2	529,8	313,0	89,6
		396,5	184,6	80,8	321,2	529,7	332,0	532,9	147,1	48,2	471,1	388,5	36,9
		554,1	55,6	166,0	86,7	260,6	184,1	146,0	192,2	10,3	272,9	139,4	99,7
		163,8	173,1	148,6	425,2	425,3	102,3	42,7	149,6	382,2	520,2	77,4	291,0
		52,7	7,1	71,8	300,5	163,8	398,9	157,2	309,8	374,2	258,8	314,0	78,4
		78,0	486,4	84,2	69,1	640,5	77,3	202,7	99,0	563,4	232,8	60,6	
		65,4	176,2	254,5	141,9	53,0	47,9	331,9	111,0	271,1	94,5	134,1	
		102,4	62,7	97,8	134,9	272,6	319,0	27,3	166,0	37,1	421,5	163,9	
		107,7	67,7	81,0	118,6	81,7	97,7	563,6	56,2	36,4	487,5	291,0	
		79,2	26,8	665,6	180,4	378,2	416,7	273,7	134,3	324,4	380,7	209,3	
		208,8	305,3	73,2	106,4	494,4	456,6	228,2		133,2	416,8	209,5	
	106,4	213,1	321,2		264,0	108,0	194,8		392,6	277,1	143,7		
	94,0	5,1	491,7		95,0	150,9	46,7		135,3	214,4	361,6		
E3: Soltero		75,3	156,1	242,7	265,8	158,7	101,7	165,8	48,0	293,8	155,8	295,1	157,6
		7,7	43,8	113,7	231,1	616,0	358,2	477,9	638,1	185,6	34,2	447,7	145,3
		133,0	484,2	196,3	10,3	405,4	56,2	107,5	97,3	488,9	170,7	243,0	314,9
		202,6	90,1	380,7	325,2	53,6	513,2	253,0	133,5	314,3	252,8	298,3	400,1
		223,7	27,1	388,5	68,6	123,0	82,8	406,2	113,6	46,5	84,3	223,0	56,6
		200,0	122,0	326,6	87,6	35,0	151,5	58,3	177,8	49,9	358,3	219,8	210,2
		49,6	81,0	91,4	84,8	316,4	342,6	77,9	406,9	483,0	287,1	539,4	251,9
		64,0	61,2	326,6	27,5	323,1	402,1	103,4	225,1	73,1	411,6	22,7	561,4
		293,0	133,7	106,4	314,3	364,6	413,7	400,1	234,8	527,5	434,7	420,9	
		218,2	445,7	266,5	248,6	221,6	257,7	550,8	406,9	214,3	502,4	145,9	
						147,5	56,1	25,1		37,1	247,1	117,8	



trata de una cartera de clientes de una entidad financiera que no dieron frente a la devolución del préstamo contraído en un terminado periodo de tiempo. El tamaño total de la muestra es de 1 individuos. Sobre cada individuo se ha observado: la experiencia siniestralidad cuantía de los impagos, 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 categórica ordinal y estado civil del cliente de naturaleza categórica nominal con tres estados: E1 (aparejados), E2 (separados o divorciados), E3 (solteros). Esta información la podemos resumir de manera agregada mediante una tabla cruzada de 3x3=9 combinaciones. Una primera estimación del nivel de riesgo es la media aritmética de toda la población, o bien utilizar la media de la tabla cruzada de cada combinación:

	A1	A2	A3
E1	208.816   39	269.565   39	366.609   44
E2	172.045   54	232.667   53	253.215   48
E3	180.380   40	246.705   43	261.575   41
			242.171   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:

- 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 fuertemente relacionada.

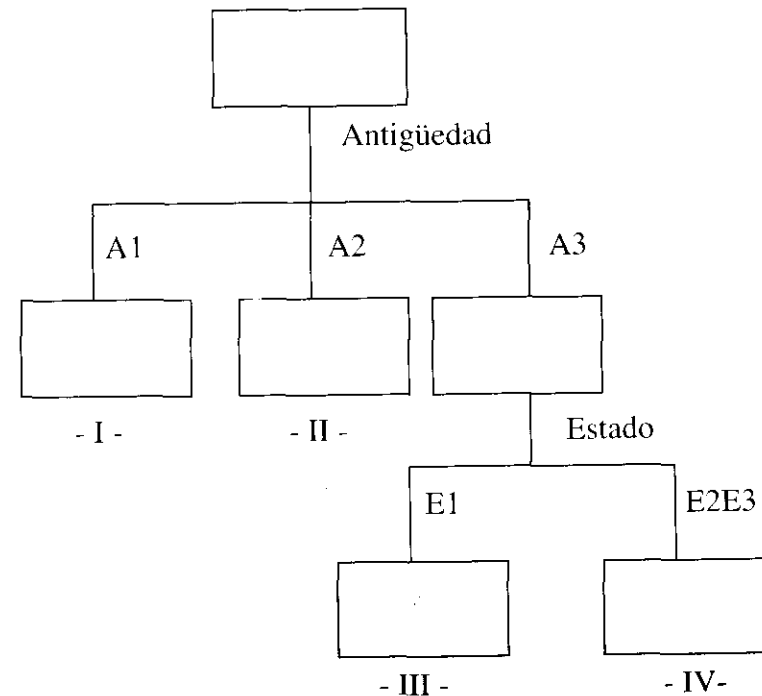
- Estado y Antigüedad:  $C^2 = 0.905_{2 \times 401} = 0.001284$

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".

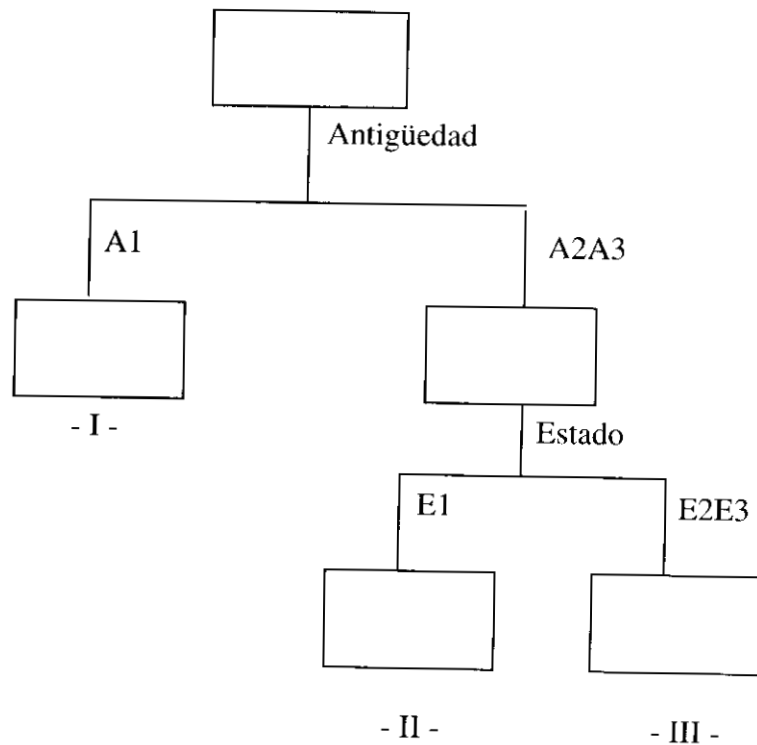
#### 4.1. Análisis de segmentación:

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:

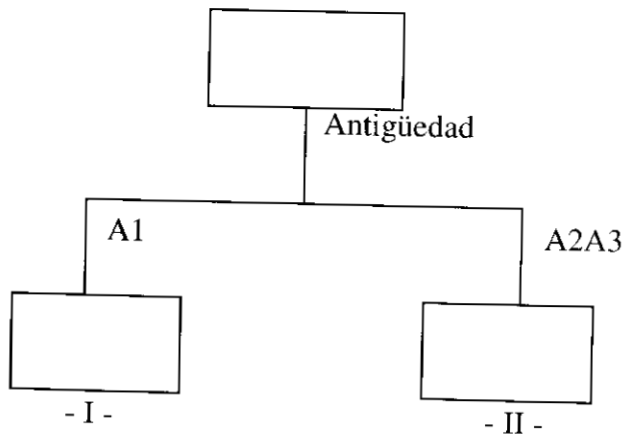
#### Árbol 1



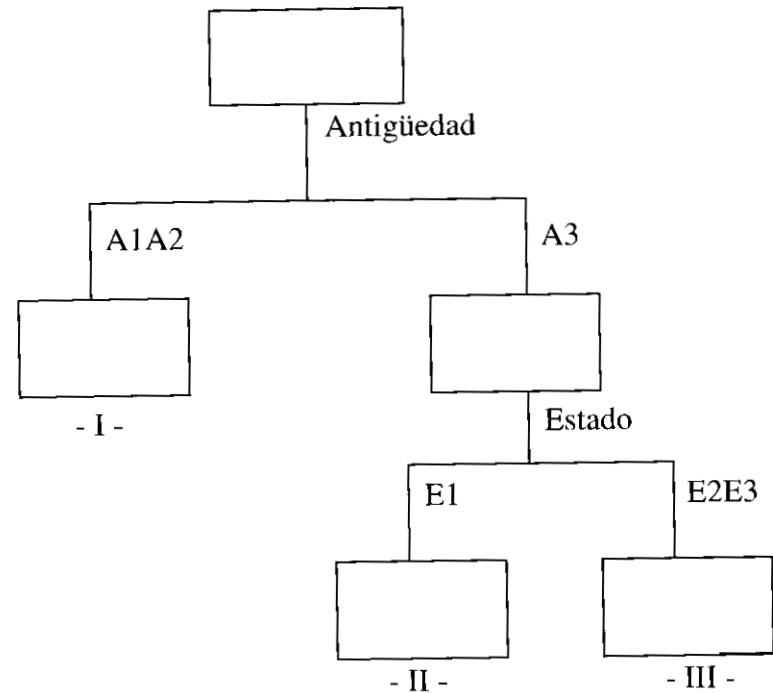
### Árbol 2



### 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.

### 1.1.1. 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	II	III 366.61   44
E2	185.33   133	247.8   135	IV
E3			257.07   89
	242.17   401		

### 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 tarifa de un modelo de credibilidad en la estimación del riesgo. Concretamente para el modelo de credibilidad de BÜLHMANN-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 sí. 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	II	III 348.86   44
E2	185.30   133	247.51   135	IV
E3			255.93   89
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 realizar una estimación por mínimos cuadrados ordinarios sobre una configuración Euclídea obtenida a partir de una matriz de distancias entre individuos calculada en el espacio predictor. En la referencia citada encontramos a realizado el proceso de selección de variables con estos datos. Aquí lasmamos la tabla de estimaciones con los 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			

**3. Modelos Lineales Generalizados:**

os referimos a Boj et al. (2002) para el detalle del Modelo Lineal generalizado (MLG) y su utilización en tarificación *no vida*. tallamos 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, $V(\mu_i)=\mu_i^3$	$g(\mu_i)=\frac{1}{\mu_i^2}$	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:

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

proceso de eliminación progresiva:

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

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

Modelo	Link:	Canónico:
Gaussiana, Identidad**	$R^2 = 0.0916$ , p-valor = 0	Sí
<b>Gaussiana, Logarítmico*</b>	$R^2 = 0.0954$ , <b>p-valor = 0</b>	<b>No</b>
Gamma, Identidad	$R^2 = 0.0697$ , p-valor = 0.000009	No
Gamma, Logarítmico	$R^2 = 0.0717$ , p-valor = 0.000549	Sí
Inversa-Gaussiana, Identidad	$R^2 = 0.0338$ , p-valor = 0.008463	No
Inversa-Gaussiana, Logarítmico	$R^2 = 0.0075$ , p-valor = 0.545475	No
Inversa-Gaussiana, $g(\mu_i) = 1/\mu_i^2$	$R^2 = 0.0080$ , 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:

*Inversa-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

*Inversa-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

Coefficiente:	t-valor:
$\hat{\beta}_0 = 5.5667$	56.6635
$\hat{\beta}_{A1} = -0.3716$	-2.1296
$\hat{\beta}_{A2} = -0.0585$	-0.4138
$\hat{\beta}_{E1} = 0.3375$	2.8299
$\hat{\beta}_{E2} = -0.0324$	-0.2391
$\hat{\beta}_{A1E1} = -0.1911$	-0.8469
$\hat{\beta}_{A2E1} = -0.2489$	-1.3476
$\hat{\beta}_{A1E2} = -0.0148$	-0.0625
$\hat{\beta}_{A2E2} = -0.0261$	-0.1335

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

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

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

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

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

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

nos fijamos en los t-valores vemos como el de menor importancia es el de la clase 2 del estado, y que la variable que marca la pauta es la igualdad. Para poder comparar los resultados del AS con un poco más de precisión, vamos a detallar cuáles serían los coeficientes si fuéramos acabado seleccionando también la interacción:

Observamos cómo los coeficientes nos confirman que los grupos no cruzados resultantes de la segmentación son lógicos.

## BIBLIOGRAFÍA

- AGRESTI, A. (1984). "Analysis of ordinal categorical data". John Wiley & Sons, Inc. New York.
- 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". Academic Press. New York. 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. AND 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, I.** (1997). "Pricing No Vida". In: Manual de banca, finanzas y seguros. Ediciones gestión 2000, S.A., Colección Universitaria Eserp.
- CAMPBELL, 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.
- GUADRAS, C. M. I P. SÁNCHEZ** (1997). "Relacions multivariants entre dos conjunts de variables". Curs de Doctorat del Departament Estadística de la Universitat de Barcelona.
- JOHNSON, A. J.** (2001). "An Introduction to Generalized Linear Models". Second Edition. Chapman & Hall. London.
- ABERMAN, S. AND A. E. RENSHAW** (1998). "Actuarial Applications of Generalized Linear Models". En: Statistics in Finance. and, D. J. and S. D. Jacka (eds). Arnold Applications of Statistics. London-Sydney-Auckland. pp. 42-65.
- ARTIGAN, J. A.** (1975). "Clustering Algorithms". John Wiley and Sons, Inc. New York.
- AWKINS, D. M., AND G. V. KASS** (1982). "Topics in Applied Multivariate Analysis". Ed. D M Hawkins. Cambridge University Press. pp. 269-302.
- AWKINS, D. M.,** (1997). "FIRM: Formal Inference-based Curvilinear Modeling". Technical Report Number 546, School of Statistics. University of Minnesota.
- BAUMANN, CH.** (2000). "Least Squares, Generalized Linear Models and Credibility Theory with applications to Tariffication". Working Paper. University of Karlsruhe.
- BLANCO, 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. AND J. A. NELDER** (1989). "Generalized Linear Models". Second Edition. Chapman & Hall. London.
- 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. AND 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.