Aplicación del análisis discriminante al estudio de la siniestralidad en el ramo del seguro de automóviles

Melgar Hiraldo, María del Carmen – mcmelhir@upo.es
Ordaz Sanz, José Antonio – jaordsan@upo.es

Departamento de Economía, Métodos Cuantitativos e Historia Económica
Universidad Pablo de Olavide, de Sevilla

RESUMEN

Conocer aquellas características del asegurado más asociadas a la ocurrencia de siniestros podría ser una cuestión de gran interés para el sector asegurador. Esto se puede abordar desde diversos puntos de vista, según se pretenda determinar las variables que influyen, y el modo en que lo hacen, en la probabilidad de sufrir siniestros, en el número de siniestros ocurridos o en ser proclive a pertenecer al colectivo de los que tienen siniestros frente al de los que no tienen, por ejemplo. De este modo, son múltiples las técnicas estadístico-econométricas que podrían usarse para esta tarea: modelos de elección discreta binaria de tipo logit o probit, modelos de recuento de datos o técnicas de análisis multivariante, entre otras.

En este trabajo nos centraremos más detenidamente en el último enfoque señalado, aplicando el análisis discriminante a los datos cedidos por una multinacional aseguradora que opera en el mercado español del ramo de automóviles. El objetivo con ello es determinar las características del asegurado que más contribuyen a la ocurrencia de siniestros y el sentido en que lo hacen en dicho sector.

***Palabras claves*:** Seguro de automóviles; siniestralidad; análisis multivariante.

***Área temática*:** Métodos Cuantitativos e Informáticos.

ABSTRACT

Knowing the policy-holders' characteristics that are more related to the occurrence of accidents could be a matter of great interest for the insurance industry. This issue can be approached from different points of view depending on the purpose of the research: to determine the variables that influence, and how they do in the likelihood of accidents, the number of incurred claims, or the probability to belong to the group of clients with accidents or with not, for example. So, there are many statistical and econometric techniques that could be used for this task: discrete choice models like logit or probit, count data models, or multivariate analysis techniques, among others.

In this paper we will focus more closely on the latter approach by applying discriminant analysis to the data provided by a multinational insurance company that works in the Spanish market for auto insurance. The aim is thus to determine the characteristics of the insured that most contribute to the occurrence of accidents and the way that they do in that sector.

***Keywords*:** Automobile insurance; Accidents; Multivariate analysis.

1. INTRODUCCIÓN

El seguro de automóviles conforma uno de los principales ramos del conjunto de la actividad aseguradora. Las últimas cifras en España de este sector del sistema financiero (DGSFP, 2012) ofrecen un volumen de primas imputadas brutas de 10.617 millones de €, representando con ello el 17,4% del total del sector asegurador; por su parte, el montante total de la siniestralidad en el ramo supone el 78,3% de la cuantía de las primas. Estas cifras ponen de relieve la importancia macroeconómica de esta industria y, dentro de la actividad, el fuerte peso que tiene el importe de la siniestralidad.

Para las entidades aseguradoras, conocer el "perfil" de sus clientes frente a la siniestralidad registrada que declaran es una cuestión fundamental, ya que influye directamente en sus cuentas de resultados. Un análisis apropiado de las variables relativas a las características del vehículo asegurado y del conductor así como de otras referidas a la póliza, se revela de este modo como un tema de sumo interés. Este análisis puede ser abordado desde distintas perspectivas. Así, se pueden encontrar trabajos como los de Ordaz y Melgar (2010) y Ordaz *et al*. (2011), que analizan la probabilidad de sufrir siniestros a través de modelos de elección discreta. O los de Shankar *et al*. (1997) y Richaudeau (1999), que usan modelos econométricos de tipo recuento o *count data* tradicionales para estimar el número de siniestros. O bien, las referencias de Lee *et al*. (2002), Melgar *et al*. (2005) y Melgar (2011), que consideran modelos inflados de ceros para la estimación del número de siniestros, como superación, a nivel teórico, de los modelos de recuento tradicionales.

El objetivo de la investigación que aquí se presenta se centra en tratar de especificar un modelo que pudiese permitir a las compañías de seguros clasificar, a priori, a sus clientes en 2 grupos: aquéllos más propensos a sufrir y declarar sus siniestros, frente a los que no. Para ello se recurrirá al análisis discriminante que, como herramienta estadístico-econométrica de la familia del análisis multivariante, se aplicará sobre el conjunto de variables que usualmente manejan las entidades para fijar las primas de sus asegurados.

La estructura del estudio llevado a cabo es como sigue. Después de este primer apartado de carácter introductorio, en el Apartado 2 describiremos los aspectos más

importantes de la información extraída de la base de datos utilizada. A continuación, el Apartado 3 se ocupará de describir la técnica multivariante utilizada en el análisis; se hará de forma muy breve, dado que es ampliamente conocida. En el Apartado 4 se ofrecerán los principales resultados del análisis empírico llevado a cabo y el Apartado 5 se dedicará finalmente a las conclusiones más relevantes de la investigación. Las Referencias bibliográficas (Apartado 6) y un Anexo con la relación de las variables empleadas cerrarán el trabajo.

2. DESCRIPCIÓN DE LA BASE DE DATOS

La base de datos utilizada en esta investigación ha sido cedida por una importante multinacional aseguradora privada que opera en todo el territorio español[1]. Comprende un conjunto de variables relativas a un total de 131.408 pólizas de vehículos asegurados, refiriéndose al periodo temporal comprendido entre el 16 de junio de 2002 y el 15 de junio de 2003. De este volumen, se ha extraído una muestra aleatoria de 15.000 pólizas[2].

Las variables incluidas en cada póliza contienen información sobre determinadas características del tipo y uso del vehículo; la edad, sexo, antigüedad del permiso de conducción y comunidad autónoma de residencia del titular de la póliza; así como la prima anual y el grado de cobertura de la póliza. Estas variables, o las categóricas que se definen a partir de ellas, se tomarán, como más adelante veremos en nuestra descripción del análisis discriminante, como variables explicativas en el mismo. Asimismo, cada póliza recoge el registro o no de siniestros asociados a la misma; este hecho lo definiremos mediante una variable dicotómica[3] y será el objeto central de este estudio. Al final del trabajo, un Anexo ofrece la definición de todas las variables consideradas.

Seguidamente, en la Tabla 1, se muestran los principales resultados del análisis descriptivo de los datos utilizados. Se ofrece, para cada variable explicativa, el peso de

[1] Pese a formar parte hoy día de una multinacional suiza, el origen de esta entidad radica en Andalucía; de ahí el peso que tienen las pólizas de esta Comunidad Autónoma en el conjunto de su cartera.

[2] Este número obedece a motivos computacionales, resultando ampliamente representativo de la población total.

[3] En realidad se dispone del número de siniestros (0, 1, 2…), pero puesto que en la presente investigación nuestro interés reside únicamente en si hay siniestros registrados o no, recogeremos este hecho mediante la definición de una variable dicotómica que adoptará los valores: 0 si no hay; 1 si hay.

las distintas categorías que se han establecido en este estudio, así como el porcentaje de registros que presenta siniestralidad en cada una de ellas.

Tabla 1. Distribución de las categorías y de la siniestralidad, por variable.

Variable Categoría	% sobre el total de la muestra	% con siniestros, sobre el total de la categoría
Tipo del vehículo asegurado		
Turismo-furgoneta	80,5%	26,5%
Camión	1,3%	25,3%
Autocar	0,2%	52,2%
Ciclomotor-moto	7,7%	7,0%
Vehículo especial	10,3%	6,8%
Uso del vehículo asegurado		
Particular	79,8%	24,7%
Profesional	19,6%	16,3%
Otros	0,6%	12,0%
Edad del asegurado		
De 18 a 25 años	3,1%	23,4%
De 26 a 45 años	39,8%	24,2%
De 46 a 70 años	51,8%	22,7%
De 71 años y más	5,3%	15,9%
Antigüedad del permiso de conducción del asegurado		
Menos de 2 años	0,8%	35,5%
2 años y más	99,2%	22,9%
Sexo del asegurado		
Hombre	85,3%	22,3%
Mujer	14,7%	26,5%
Región de residencia del asegurado		
Sur	40,3%	24,7%
Este	18,5%	22,6%
Norte	18,6%	23,0%
Centro	18,2%	19,7%
Canarias	4,4%	21,3%
Prima anual de la póliza		
Hasta 300 €	32,2%	11,8%
De 301 a 400 €	26,8%	22,6%
De 401 a 600 €	23,2%	28,1%
Más de 600 €	17,8%	36,9%
Grado de cobertura de la póliza		
Grado bajo	54,3%	16,1%
Grado medio	37,8%	29,3%
Grado alto	7,9%	39,4%
Total	100%	22,9%

Fuente: Elaboración propia.

Para empezar, deben resaltarse las cifras de nuestra variable dependiente: de los 15.000 asegurados, 11.558 no han declarado siniestro alguno, esto es el 77,1% sobre el total; la cifra de asegurados con siniestros supone, lógicamente, el 22,9% restante.

Ya en las variables explicativas, respecto al tipo del vehículo asegurado se han considerado cinco grupos: "turismo o furgoneta", "camión", "autocar", "ciclomotor o moto" y "vehículo especial". Como muestra la Tabla 1, destaca el primero de los citados, con el 80,5% del total. En el otro extremo encontramos que camiones y autocares significan sólo un 1,5% de manera conjunta. Sobre la siniestralidad registrada en cada una de estas categorías, ésta se observa en el 26,5% de los turismos y furgonetas, cifra muy similar (25,3%) a la de los camiones. Autocares por un lado, y ciclomotores y motos y vehículos especiales por el otro, presentan sin embargo porcentajes muy dispares: 52,2% y en torno al 7%, respectivamente.

En cuanto al análisis de las cifras alusivas al uso principal del vehículo asegurado, según se aprecia en la Tabla 1, el 79,8% de los casos se destina a uso "particular"; el uso "profesional" (que engloba el servicio público, utilizaciones industriales, transporte de mercancías, transporte escolar, transporte general de viajeros y labores agrícolas) representa el 19,6%; y la categoría de "otros" (alquiler, escuela de conductores, compra-venta y retirada del permiso de conducción) supone únicamente el 0,6% sobre el total. En relación a la siniestralidad registrada, el uso profesional y, más aún, los otros usos, muestran porcentajes de siniestralidad inferiores, 16,3% y 12,0% respectivamente, que el que se da en el uso particular: el 24,7%.

Entre las características del asegurado, analizamos en primer lugar la edad. Ésta se ha categorizado en cuatro tramos[4]: "de 18 a 25 años", "de 26 a 45 años", "de 46 a 70 años" y "de 71 años y más". Los tramos mayoritarios resultan ser los dos intermedios, conteniendo conjuntamente el 91,6% del total de asegurados. En lo que se refiere a la siniestralidad, el porcentaje de asegurados con siniestros es bastante similar (en torno al 22-24%). La cifra de los mayores de 71 años, en cambio, es bastante inferior (15,9%).

La antigüedad del permiso de conducción del tomador del seguro es otro de los aspectos relevantes que se han considerado. En la Tabla 1 se muestra la distinción que se hace entre menos de dos años y dos años o más (según el criterio empleado por la

[4] Esta variable se ha dividido ahora en tramos con fines descriptivos. Sin embargo, posteriormente en la aplicación del análisis discriminante se considerará directamente como variable cuantitativa.

entidad aseguradora estudiada). Del total de asegurados de la muestra, sólo el 0,8% tiene una antigüedad en su permiso inferior a 2 años. Su siniestralidad (35,5%), sin embargo, es muy superior a la de los asegurados con mayor experiencia (22,9%).

El sexo de los asegurados es también otra de las variables tenidas en cuenta. El 85,3% del total de la muestra son hombres, presentando siniestros el 22,3%. Esta cifra resulta algo superior en las mujeres, que es del 26,5%, según muestra la Tabla 1.

El lugar de residencia del asegurado constituye otra importante variable característica del asegurado. Constituye una "*proxy*" del lugar habitual de circulación de éste. Se han considerado cinco grandes regiones o zonas geográficas, como resultado de la agregación de distintas comunidades autónomas (véase el Anexo): "Sur", "Este", "Norte", "Centro" y "Canarias". La región "Sur" es la que concentra un mayor número de registros: el 40,3% del total de asegurados. En el otro extremo, encontramos a "Canarias", representada únicamente por un 4,4% de los asegurados. Las tres regiones restantes tienen cada una de ellas un peso en torno al 18,5%. Respecto a la siniestralidad que se encuentra en cada una de estas regiones, la Tabla 1 evidencia que los residentes en la zona mayoritaria, la región "Sur", son quienes presentan un mayor porcentaje de siniestros; en particular, en el 24,7% de los casos. En el extremo opuesto se sitúa la región "Centro", donde el 19,7% de los asegurados registran siniestros.

El último bloque de variables estudiadas hace referencia a características directamente relacionadas con las pólizas. En concreto, se han analizado el importe anual pagado en concepto de prima y el grado de cobertura aseguradora.

Respecto a la primera de las variables indicadas, el importe de la prima, ésta se ha dividido, a efectos descriptivos, en cuatro intervalos[5]: "hasta 300 €", "de 301 a 400 €", "de 401 a 600 €" y "más de 600 €". Como puede observarse en la Tabla 1, el tramo mayoritario es el más bajo, con el 32,2% de los asegurados. Los dos centrales ofrecen cifras similares (26,8% y 23,2%, respectivamente). La categoría de primas más caras (más de 600 €), supone sólo el 17,8% del total. En cuanto a la distribución de la siniestralidad registrada en cada uno de los intervalos considerados, ésta resulta creciente de manera muy apreciable, yendo del 11,8% para las pólizas inferiores a 300 € hasta el 36,9% de las pólizas con primas superiores a 600 €.

[5] En la aplicación del análisis discriminante que se llevará a cabo más adelante, esta variable se introduce como variable cuantitativa de modo directo.

Finalmente, en lo tocante al grado de cobertura de la póliza, éste se ha dividido en tres niveles distintos en función de las garantías contratadas en el seguro. El "grado bajo" incluye exclusivamente las garantías obligatorias según la ley; las pólizas con este nivel de cobertura representan el 54,3% del total. Los que se decantan por alguna garantía adicional opcional, como puede ser la referida a la rotura de lunas, incendio y/o robo del vehículo, disfrutan del que hemos denominado "grado medio" de cobertura; éste es el tipo elegido por el 37,8% de los asegurados considerados. Por último, el "grado alto" cubre, además, los daños propios del vehículo; aquí se encuentra el 7,9% del total de registros. El análisis de la siniestralidad para cada uno de los grados de cobertura se muestra, una vez más, también en la Tabla 1, donde puede comprobarse cómo el porcentaje de siniestralidad declarada va creciendo conforme lo hace el grado de cobertura. Así, éste es del 16,1% para el nivel bajo; del 29,3% para el medio; y del 39,4% para el alto. Estos resultados resultan enormemente interesantes por cuanto pueden encerrar comportamientos tan inherentes a la actividad aseguradora como son los de riesgo moral y/o de selección adversa (Boyer y Dionne (1989), Dionne *et al*. (1999), Chiappori y Salanié (2000), Abbring *et al*. (2003), Cohen (2005), entre otros).

3. METODOLOGÍA: EL ANÁLISIS DISCRIMINANTE

El análisis discriminante es una técnica bien conocida de análisis multivariante (Sharma, 1998; Hair *et al*., 1999) que pertenece al grupo de los llamados métodos de dependencia. Partiendo de un conjunto de elementos o individuos que pertenecen a diferentes grupos previamente establecidos, se trata de analizar la información relativa a una serie de variables independientes con un doble fin: explicativo y predictivo. La pertenencia de los elementos o individuos objeto de estudio a un grupo u otro se introduce en el análisis a través de una variable cualitativa que toma tantos valores como grupos existentes y juega el papel de variable dependiente. Las variables independientes reciben el nombre de variables discriminantes o clasificadoras. La información disponible se sintetiza en las denominadas funciones discriminantes, que no son más que combinaciones lineales de las variables discriminantes.

El número de funciones discriminantes que se deben estimar es el mínimo entre J-1 y m, siendo J el número de grupos (mutuamente excluyentes) definidos mediante

una variable dependiente nominal y *m* el número de variables clasificadoras empleadas[6]. En el caso que nos ocupa, tenemos 2 grupos: el de aquellos individuos que declaran siniestros a sus aseguradoras y el de los que no. Así pues, determinaremos una única función discriminante, *D,* que dependerá de un conjunto de *m* variables clasificadoras, referidas a ciertas características personales del asegurado, del tipo y uso del vehículo y de la póliza contratada, que denotaremos por X_1, X_2, …, X_m, de tal forma que:

$$D = a_1 X_1 + a_2 X_2 + \ldots + a_m X_m . \qquad [1]$$

El objetivo que se persigue es que los valores de esta función se diferencien lo más posible de un grupo a otro y sean muy parecidos para los elementos de un mismo grupo. Habrá que encontrar entonces los valores de los coeficientes a_1, a_2, …, a_m para que esto se cumpla. El objetivo que persigue el análisis es no sólo ver qué variables resultan significativas en este proceso, sino también, una vez creada la función discriminante, calcular el valor de ésta para nuevos sujetos (*puntuación discriminante*) y clasificar a éstos en el grupo correspondiente según la puntuación obtenida.

Existen varios procedimientos para calcular las funciones discriminantes. Uno de los más usados es el de Fisher, consistente en maximizar la variación de la función *D* entre grupos, tratando al mismo tiempo, para evitar errores, que la variación dentro de cada grupo sea la menor posible. Esto es, lo que se persigue es maximizar la ratio:

$$\lambda = \frac{\text{variación inter - grupos}}{\text{variación intra - grupos}} . \qquad [2]$$

Las variaciones inter-grupos e intra-grupos se calculan a partir de las correspondientes sumas de los cuadrados de las desviaciones de las puntuaciones discriminantes con respecto a las de los valores de las variables independientes iguales a las medias de cada grupo (*centroides*); esto es:

$$\overline{D_1} = a_1 \overline{X_1^{(1)}} + a_2 \overline{X_2^{(1)}} + \ldots + a_m \overline{X_m^{(1)}} ; \qquad \overline{D_2} = a_1 \overline{X_1^{(2)}} + a_2 \overline{X_2^{(2)}} + \ldots + a_m \overline{X_m^{(2)}} . \quad [3]$$

De este modo, λ puede expresarse en función de los coeficientes desconocidos a_1, a_2, …, a_m; se tratará entonces de maximizar esa función de 2 variables para

[6] Como toda técnica estadística, el análisis discriminante debe cumplir una serie de requisitos previos a su aplicación para la obtención de resultados plenamente fiables. Los que se acaban de citar son sólo algunos de ellos, pero hay más: ausencia de multicolinealidad entre las variables discriminantes, distribución normal de éstas y homoscedasticidad intra-grupos. En ocasiones, sin embargo, puede que algunas de ellas no se cumplen, no suponiendo ello necesariamente la invalidación de los resultados (Hair *et al*., 1999).

obtener los coeficientes de la función discriminante. La solución resultante indica que $a_1,\ a_2,\ \ldots,\ a_m$ son las coordenadas de un autovector asociado al mayor autovalor λ de cierta matriz cuyos elementos dependen únicamente de los valores observados de las variables independientes $X_1,\ X_2,\ \ldots,\ X_m$.

Tras definir la función discriminante, se debe fijar un criterio para clasificar a los nuevos elementos. Entre ellos, figura el *punto de corte discriminante* (*PCD*), que no es más que la media de las puntuaciones discriminantes medias de cada grupo:

$$PCD = \frac{\overline{D^{(1)}} + \overline{D^{(2)}}}{2} \qquad [4]$$

y finalmente aplicar el siguiente criterio para clasificar un elemento *i*:

- Si $D_i < PCD$, se clasifica al elemento en el grupo 1.
- Si $D_i > PCD$, se clasifica al elemento en el grupo 2.

Como último paso, se debe evaluar la capacidad predictiva del modelo. El indicador más inmediato que suele utilizarse es la ratio de aciertos obtenidos al clasificar los distintos casos en los grupos considerados. Este valor puede luego compararse con el resultante de la aplicación del *criterio de aleatoriedad proporcional* (Hair *et al*., 1999), siendo tanto mejor el modelo cuanto más se supere este valor:

$$C_{PRO} = p^2 + (1-p)^2, \qquad [5]$$

donde *p* y (1-*p*) hacen referencia a la proporción de cada uno de los 2 grupos considerados sobre el total de la muestra.

Otro contraste de la capacidad discriminatoria del modelo es el que se realiza a través del estadístico *Q* de PRESS (Allen, 1974), que sigue una distribución χ^2 con 1 grado de libertad, y viene dado por la expresión:

$$Q = \frac{(N - nK)^2}{N(K-1)}, \qquad [6]$$

donde *N* es el tamaño muestral, *n* el número de observaciones correctamente clasificadas y *K* el número de grupos considerado en el análisis. El rechazo de la hipótesis nula implica que el resultado obtenido por la clasificación de nuestro modelo resulta aceptable en términos estadísticos.

A continuación, en el capítulo dedicado a mostrar los resultados de la aplicación empírica de nuestro trabajo, se mostrarán algunos resultados adicionales a los aquí expuestos, cuyo resultado iremos explicando convenientemente.

4. RESULTADOS DEL ANÁLISIS EMPÍRICO

El objetivo central del presente estudio consiste en determinar cuáles de las variables que habitualmente utilizan las compañías aseguradoras para fijar las primas de sus clientes podrían ayudar a explicar el registro de siniestros por parte de éstos y en qué sentido. El propósito final es construir un modelo que pudiese ayudar a clasificar, a priori, a un individuo en el grupo de los propensos a la declaración de siniestros o en el de los que no. Como se ha expuesto previamente, para este fin recurrimos al análisis discriminante aplicado sobre la base de datos descrita en el Apartado 2, haciendo uso para ello del paquete estadístico *IBM SPSS Statistics v.20.0.0*.

Se consideran 2 grupos excluyentes en los que se clasifican los asegurados: uno donde están los que no han registrado ningún siniestro (para quienes la variable SINIESTR = 0), y otro que está formado por los que sí han declarado algún siniestro en el periodo considerado (SINIESTR = 1). Respecto a las variables discriminantes, como ya se ha indicado en el análisis descriptivo, figuran tanto variables cuantitativas como variables categóricas, que se han introducido en el estudio a través de variables ficticias excluyendo la correspondiente a la categoría base (véase el Anexo).

Como primer paso de la aplicación del análisis discriminante, resulta habitual llevar a cabo los contrastes de igualdad de medias de las distintas variables independientes consideradas en los 2 grupos establecidos, con objeto de conformar una idea inicial de las variables que pueden ser más relevantes en el modelo. Los resultados (Figura 1) señalan que en los casos de las variables CAMION, R_ESTE, R_NORTE y R_CANARIAS se aceptaría la hipótesis nula de igualdad de medias en ambos grupos, por lo que, en principio, estas variables serían las únicas que no deberían ser discriminantes. No obstante, el hecho de que una variable individualmente no dé muestras de tener un comportamiento diferenciado en cada grupo, no implica necesariamente que no lo tenga al interactuar con otras variables consideradas en el análisis (Huberty, 1994).

Figura 1. Pruebas de igualdad de las medias de los grupos.

	Lambda de Wilks	F	gl1	gl2	Sig.
CAMION	1,000	,594	1	14998	,441
AUTOCAR	,999	11,136	1	14998	,001
CICL_MOT	,988	181,209	1	14998	,000
VEH_ESP	,983	261,775	1	14998	,000
USO_PROF	,994	91,199	1	14998	,000
OTR_USOS	1,000	6,325	1	14998	,012
EDAD	,997	39,868	1	14998	,000
ANTIG_2A	,999	9,631	1	14998	,002
MUJER	,999	18,759	1	14998	,000
R_ESTE	1,000	,254	1	14998	,614
R_NORTE	1,000	,005	1	14998	,944
R_CENTRO	,999	20,052	1	14998	,000
R_CANARIAS	1,000	1,042	1	14998	,307
PRIMA_T	,970	462,227	1	14998	,000
GR_MEDIO	,986	209,414	1	14998	,000
GR_ALTO	,987	201,619	1	14998	,000

Tras esto, se ha procedido a estimar la función discriminante, incluyendo en el modelo todas las variables simultáneamente. La Figura 2 señala que se ha obtenido una única función discriminante, como era de esperar al existir únicamente 2 grupos, y que, además, ésta resulta ser plenamente significativa como muestra el *p*-valor asociado a la Lambda de Wilks (0,000).

Figura 2. Construcción de la función discriminante.

Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	,063[a]	100,0	100,0	,244

Se ha empleado 1 primera función discriminante canónica en el análisis.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	Gl	Sig.
1	,940	922,799	16	,000

Los coeficientes estandarizados de las funciones discriminantes canónicas se han usado tradicionalmente para determinar la importancia relativa de cada variable en la función discriminante, de modo que cuanto más elevado sea el valor absoluto de su coeficiente, mayor será la aportación de la variable en cuestión. Así, según se aprecia en la Figura 3, las variables "PRIMA_T", "CICL_MOT", "GR_MEDIO", "VEH_ESP" y "GR_ALTO" serían las que presentan mayor relevancia a la hora de clasificar a los asegurados en un grupo u otro. Por su parte, las variables menos relevantes serían "USO_PROF", "AUTOCAR", "R_CANARIAS", "R_ESTE" y el sexo del asegurado ("MUJER").

Figura 3. Coeficientes estandarizados de las funciones discriminantes canónicas.

	Función
	1
CAMION	,207
AUTOCAR	-,033
CICL_MOT	,370
VEH_ESP	,324
USO_PROF	,000
OTR_USOS	,090
EDAD	,113
ANTIG_2A	-,114
MUJER	-,052
R_ESTE	,047
R_NORTE	,150
R_CENTRO	,093
R_CANARIAS	,040
PRIMA_T	-,469
GR_MEDIO	-,364
GR_ALTO	-,236

Sin embargo, de acuerdo con Hair *et al*. (1999), las cargas discriminantes son más válidas que los coeficientes estandarizados de las funciones discriminantes canónicas para llevar a cabo la tarea anterior. Dichas cargas las proporciona *IBM SPSS Statistics* ordenadas de mayor a menor bajo el título *Matriz de estructura* (Figura 4). Según este enfoque, las variables se consideran relevantes si la carga, en valor absoluto, es mayor o igual que 0,30. Así, las variables que más discriminan serían "PRIMA_T",

"VEH_ESP", "GR_MEDIO", "GR_ALTO", "CICL_MOT" y "USO_PROF". Excepto por esta última, se obtiene el mismo resultado que con los coeficientes estandarizados, si bien el orden de importancia difiere.

Figura 4. Matriz de estructura.

	Función
	1
PRIMA_T	-,697
VEH_ESP	,524
GR_MEDIO	-,469
GR_ALTO	-,460
CICL_MOT	,436
USO_PROF	,309
EDAD	,205
R_CENTRO	,145
MUJER	-,140
AUTOCAR	-,108
ANTIG_2A	-,101
OTR_USOS	,081
R_CANARIAS	,033
CAMION	-,025
R_ESTE	,016
R_NORTE	-,002

Las puntuaciones discriminantes que permiten clasificar a los asegurados en uno u otro grupo se calculan a partir de los coeficientes no tipificados de las funciones canónicas discriminantes, que muestra la Figura 5. La puntuación media obtenida en el grupo de los que no tienen siniestros es 0,138, por lo que este grupo se encuentra localizado, en promedio, en las puntuaciones positivas de la función discriminante no tipificada; por el contrario, el grupo de los que tienen siniestros se localiza en promedio en las puntuaciones negativas, dado que la puntuación media es en este caso de -0,462 (Figura 6).

El signo de los coeficientes estandarizados de las variables discriminantes (Figura 3), junto con el signo de las funciones discriminantes canónicas no tipificadas en los centroides de los grupos (Figura 6), permiten determinar el sentido del efecto de cada variable sobre la clasificación de los asegurados.

Figura 5. Coeficientes de las funciones canónicas discriminantes.

	Función
	1
CAMION	1,833
AUTOCAR	-,832
CICL_MOT	1,397
VEH_ESP	1,074
USO_PROF	,000
OTR_USOS	1,155
EDAD	,009
ANTIG_2A	-1,357
MUJER	-,144
R_ESTE	,120
R_NORTE	,385
R_CENTRO	,241
R_CANARIAS	,193
PRIMA_T	-,001
GR_MEDIO	-,757
GR_ALTO	-,877
(Constante)	,265

Figura 6. Funciones en los centroides de los grupos.

SINIESTR	Función
	1
0	,138
1	-,462

En el caso de las variables cuantitativas, si toman un valor superior a la media, el asegurado se clasificará en el grupo en el que el signo de la puntuación del centroide coincida con el signo del coeficiente de la variable. Así, en el caso de la edad, si ésta es superior a la media (48 años), el individuo se clasificará en el grupo de los que no tienen siniestros. Por el contrario, una prima superior a la media (721,19 €) contribuirá a que el asegurado se clasifique en el grupo de los que tienen siniestros.

Para las variables ficticias, un coeficiente positivo indicará que el asegurado tendrá una probabilidad mayor de pertenecer al grupo de los que no tienen siniestros que los que están en la categoría base. Así ocurre por ejemplo con los camiones,

ciclomotores y motos y vehículos especiales, frente a turismos y furgonetas; o también con el uso profesional y otros usos frente al uso particular; asimismo, los asegurados de las regiones Este, Norte, Centro y Canarias también son más propensos a no tener siniestros que los de las demás regiones (Sur). En cambio, un coeficiente negativo para una variable ficticia es señal de una mayor probabilidad de pertenecer al grupo de los que tienen siniestros. Es el caso de autocares frente a turismos y furgonetas; o de los que tienen una antigüedad en el carné de conducir inferior a 2 años frente a los que tienen una mayor antigüedad; o de las mujeres frente a los hombres; o de los residentes en la región Sur frente al resto; o de los que disfrutan de grados de cobertura media y de cobertura alta frente a los que tienen una cobertura baja.

En la práctica, para clasificar nuevos elementos en los 2 grupos considerados es habitual recurrir a las llamadas *funciones de clasificación*, en lugar de trabajar con las funciones canónicas discriminantes no tipificadas. La Figura 7 ofrece los coeficientes de dichas funciones, también conocidas como *funciones discriminantes lineales de Fisher*.

Figura 7. Coeficientes de la función de clasificación.

	SINIESTR	
	0	1
CAMION	-9,404	-10,503
AUTOCAR	-9,649	-9,150
CICL_MOT	6,774	5,937
VEH_ESP	2,915	2,272
USO_PROF	1,354	1,354
OTR_USOS	4,971	4,279
EDAD	,327	,322
ANTIG_2A	7,360	8,173
MUJER	3,192	3,279
R_ESTE	2,402	2,330
R_NORTE	1,528	1,298
R_CENTRO	2,701	2,557
R_CANARIAS	3,746	3,630
PRIMA_T	,005	,006
GR_MEDIO	3,323	3,776
GR_ALTO	-1,195	-,670
(Constante)	-12,110	-13,577

Funciones discriminantes lineales de Fisher.

A partir de ellos, se procedería calculando el valor de las 2 funciones (una por grupo) utilizando los valores conocidos de las variables independientes para un nuevo individuo, y éste se clasificaría en el grupo para el que se obtuviese una mayor puntuación.

Por último, la evaluación de la capacidad predictiva del modelo estimado resulta sumamente importante en nuestra tarea. Ésta suele medirse en primera instancia con la ratio de aciertos al clasificar los distintos casos en los grupos fijados en el análisis. La *matriz de clasificación* dada en la Figura 8 muestra el número de casos y porcentaje de clasificados correcta e incorrectamente con la aplicación del modelo en cada grupo. El porcentaje global de aciertos alcanza el 76,8% del total de casos.

Figura 8. Resultados de la clasificación.

		SINIESTR	Grupo de pertenencia pronosticado		Total
			0	1	
Original	Recuento	0	11407	151	11558
		1	3336	106	3442
	%	0	98,7	1,3	100,0
		1	96,9	3,1	100,0

Clasificados correctamente el 76,8% de los casos agrupados originales.

Si comparamos este valor con el obtenido según el *criterio de aleatoriedad proporcional* indicado en la Ecuación 5 en el Apartado 3, para una proporción *p* de asegurados que no registran accidentes de 0,771, éste resulta ser del 64,7%. De acuerdo con ello pues, podemos concluir que la precisión clasificatoria de nuestro modelo es sustancialmente más elevada, por lo que el poder discriminante de las variables independientes consideradas en nuestro estudio resulta aceptable.

Para ratificar esta conclusión, llevamos también a cabo el contraste de la capacidad discriminatoria del modelo mediante el estadístico *Q* de PRESS (Ecuación 6, Apartado 3). Sabiendo que en este caso: $N = 15.000$, $n = 11.407+106 = 11.513$ y $K = 2$, se obtiene entonces que: $Q = 4.294,45$. Este valor resulta claramente superior al valor crítico de la χ_1^2 que es de 6,63 para un nivel de significación del 1% (el más exigente). Por tanto, las predicciones obtenidas con nuestro modelo son sustancialmente mejores que las obtenidas aleatoriamente.

En definitiva, los resultados de nuestro análisis pueden considerarse sumamente satisfactorios.

5. CONCLUSIONES

En este trabajo, se ha tratado de especificar un modelo que pudiese resultar útil para clasificar a los asegurados del sector del automóvil en el grupo de los que declaran siniestros o en el de los que no lo hacen, en función de determinadas características que manejan comúnmente las compañías de seguros en relación al vehículo asegurado, al conductor y a la póliza. Para ello se ha recurrido a una herramienta de la familia del análisis multivariante: el análisis discriminante, que se ha aplicado a una base de datos cedida por una multinacional aseguradora privada que opera a nivel nacional, formada por las pólizas de los vehículos asegurados durante un año concreto.

Dado que se pretende clasificar a los individuos en 2 grupos, el análisis discriminante proporciona una única función discriminante, que en nuestro estudio ha resultado ser significativa. Entre los resultados obtenidos en esta investigación, destacamos por una parte cuáles son las variables que más discriminan y, por otra, en qué sentido afectan las variables consideradas a la clasificación de los asegurados.

Respecto al primer punto, la prima pagada, el grado de cobertura y algunos tipos de vehículos, son las variables con mayor correlación con la función discriminante y, por tanto, las que más contribuyen a la correcta clasificación de los asegurados.

En cuanto al efecto de las variables consideradas, un asegurado con una edad superior a la media, o que conduzca un camión, un ciclomotor o moto o un vehículo especial, así como el que destine su vehículo a un uso distinto del particular, o bien circule por las regiones Este, Norte, Centro o Canarias, será más propenso a pertenecer al grupo de los que no declaran siniestros. Por el contrario, es mayor la probabilidad de pertenecer al grupo de los que declaran siniestros si se trata de un asegurado que paga una prima superior a la media, o tiene un grado de cobertura no básico, o conduce un autocar, o tiene menos de 2 años de antigüedad en el permiso de conducción, o es mujer, o vive en la región Sur.

Para finalizar, queremos señalar que el modelo ha alcanzado un porcentaje de aciertos en la clasificación de casi el 77%. Esta cifra es superior a la obtenida a través

del criterio de aleatoriedad proporcional. Asimismo el contraste *Q* de PRESS ha puesto de manifiesto que nuestro modelo clasifica significativamente mejor que si la clasificación hubiese sido aleatoria. Por ello, en definitiva consideramos que la capacidad predictiva del modelo propuesto resulta plenamente aceptable.

6. REFERENCIAS BIBLIOGRÁFICAS

- ABBRING, J.H.; CHIAPPORI, P.A.; HECKMAN J.J. y PINQUET, J. (2003). "Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish?". Journal of the European Economic Association, 1 (Papers and Proceedings), pp. 512-521.
- ALLEN, D.M. (1974). "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction". Technometrics, 16, pp. 125-127.
- BOYER, M. y DIONNE, G. (1989). "An Empirical Analysis of Moral Hazard and Experience Rating". Review of Economics and Statistics, 71, pp. 128-134.
- CHIAPPORI, P.A. y SALANIÉ, B. (2000). "Testing for Asymmetric Information in Insurance Markets". Journal of Political Economy, 108, 1, pp. 56-78.
- COHEN, A. (2005). Asymmetric Information and Learning: Evidence from the Automobile Insurance Market. Review of Economics and Statistics, 87, 2, pp. 197-207.
- DGSFP (2012). Seguros y Fondos de Pensiones: Informe 2011. Dirección General de Seguros y Fondos de Pensiones, Ministerio de Economía y Competitividad, Madrid.
- DIONNE, G.; GOURIÉROUX, C. y VANASSE, C. (1999). "Evidence of Adverse Selection in Automobile Insurance Markets". En Dionne, G. y Laberge-Nadeau, C. (eds.): Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation, Kluwer Academic Publishers, pp. 13-46, Montréal.
- HAIR, J.F. Jr.; ANDERSON, R.E.; TATHAM, R.L. y BLACK, W.C. (1999). "Análisis Multivariante", 5ª ed. Prentice Hall Iberia, Madrid.

- HUBERTY, C.J. (1994). "Applied Discriminant Analysis". Wiley-Interscience, Nueva York.
- LEE, A.H.; STEVENSON, M.R.; WANG, K. y YAU, K.K.W. (2002). "Modeling Young Driver Motor Vehicle Crashes: Data with Extra Zeros". Accident Analysis and Prevention, 34, 4, pp. 515-521.
- MELGAR, M.C. (2011). "Utilización de los modelos inflados de ceros en la estimación del número de siniestros en el seguro de automóviles". En Ayuso, M. (ed.): Métodos cuantitativos en economía del seguro del automóvil, pp. 35-51. Barcelona.
- MELGAR, M.C.; ORDAZ, J.A. y GUERRERO, F.M. (2005). "Diverses Alternatives pour Déterminer les Facteurs Significatifs de la Fréquence d'Accidents dans l'Assurance Automobile". Assurances et Gestion des Risques - Insurance and Risk Management, 73, 1, pp. 31-54.
- ORDAZ, J.A. y MELGAR, M.C. (2010). "Covariate-Based Pricing of Automobile Insurance". Insurance Markets and Companies: Analyses and Actuarial Computations, 1, 2, pp. 92-99.
- ORDAZ, J.A.; MELGAR, M.C. y KHAN, M.K. (2011). "An Analysis of Spanish Accidents in Automobile Insurance: The Use of the Probit Model and the Theoretical Potential of Other Econometric Tools". Equilibrium, 6, 3, pp. 117-134.
- RICHAUDEAU, D. (1999). "Automobile Insurance Contracts and Risk of Accident: An Empirical Test Using French Individual Data". Geneva Papers on Risk and Insurance Theory, 24, 1, pp. 97-114.
- SHANKAR, V.; MILTON, J. y MANNERING, F. (1997). "Modeling Accident Frequencies as Zero-Altered Probability Processes: an Empirical Inquiry". Accident Analysis and Prevention, 29, 6, pp. 829-837.
- SHARMA, S. (1998). "Applied Multivariate Techniques". John Wiley & Sons, Nueva York.

ANEXO: Definición de las variables utilizadas en el análisis

Tipo del vehículo asegurado
TUR_FUR = 1 si el vehículo asegurado es un turismo o una furgoneta; 0 en caso contrario (categoría base)
CAMION = 1 si el vehículo asegurado es un camión; 0 en caso contrario
AUTOCAR = 1 si el vehículo asegurado es un autocar; 0 en caso contrario
CICL_MOT = 1 si el vehículo asegurado es un ciclomotor o una moto; 0 en caso contrario
VEH_ESP = 1 si el vehículo asegurado es un vehículo especial; 0 en caso contrario
Uso del vehículo asegurado
USO_PART = 1 si el uso del vehículo asegurado es el uso particular; 0 en caso contrario (categoría base)
USO_PROF = 1 si el uso del vehículo asegurado es profesional; 0 en caso contrario
OTR_USOS = 1 si el vehículo asegurado se destina a otros usos; 0 en caso contrario
Edad del asegurado
EDAD: edad del asegurado (en años)
ED18_25 = 1 si el asegurado tiene entre 18 y 25 años; 0 en caso contrario
ED26_45 = 1 si el asegurado tiene entre 26 y 45 años; 0 en caso contrario
ED46_70 = 1 si el asegurado tiene entre 46 y 70 años; 0 en caso contrario
ED71_ = 1 si el asegurado tiene 71 años o más; 0 en caso contrario
Antigüedad del permiso de conducción del asegurado
ANTIG<2A = 1 si el asegurado obtuvo el permiso de conducción hace menos de 2 años; 0 en caso contrario
Sexo del asegurado
MUJER = 1 si el asegurado es mujer; 0 en caso contrario
Región de residencia del asegurado
SUR = 1 si el asegurado reside en la región Sur (Andalucía, Ceuta, Melilla); 0 en caso contrario (categoría base)
ESTE = 1 si el asegurado reside en la región Este (Cataluña, Comunidad Valenciana, Islas Baleares, Murcia); 0 en caso contrario
NORTE = 1 si el asegurado reside en la región Norte (Aragón, Navarra, País Vasco, La Rioja, Cantabria, Principado de Asturias, Galicia); 0 en caso contrario
CENTRO = 1 si el asegurado reside en la región Centro (Castilla y León, Comunidad Autónoma de Madrid, Castilla-La Mancha, Extremadura); 0 en caso contrario

CANARIAS = 1 si el asegurado reside en la región de Canarias (Islas Canarias); 0 en caso contrario
Prima anual de la póliza
PRIMA_T: importe de la prima total anual (en €)
P0_300 = 1 si la prima anual pagada por el asegurado no supera los 300 €; 0 en caso contrario
P301_400 = 1 si la prima anual pagada por el asegurado es mayor de 300 € y no supera los 400 €; 0 en caso contrario
P401_600 = 1 si la prima anual pagada por el asegurado es mayor de 400 € y no supera los 600 €; 0 en caso contrario
P601_ = 1 si la prima anual pagada por el asegurado es mayor de 600 €; 0 en caso contrario
Grado de cobertura de la póliza
GR_BAJO = 1 si el asegurado disfruta del grado de cobertura bajo; 0 en caso contrario (categoría base)
GR_MEDIO = 1 si el asegurado disfruta del grado de cobertura medio; 0 en caso contrario
GR_ALTO = 1 si el asegurado disfruta del grado de cobertura alto; 0 en caso contrario
Siniestralidad
SINIESTR = 1 si el asegurado ha declarado algún siniestro a la compañía de seguros; 0 en caso contrario

AGRADECIMIENTOS

Este trabajo ha recibido ayuda del Ministerio de Economía y Competitividad (Proyecto ECO2012-35584).