3. ARTÍCULOS DE APLICACIÓN

TARIFICACIÓN DEL SEGURO DEL AUTOMÓVIL: MÉTODOS DE ANÁLISIS MULTIVARIANTE

Eva Boj del Val
Departamento de Matemática Económica, Financiera y Actuarial
Universidad de Barcelona

1. Introducción

La tarificación de los seguros, desde el punto de vista técnico[1], tiene como objetivo el correcto cálculo de primas equitativas[2] y suficientes[3]. Distinguimos dos sistemas de tarificación: *a priori* y *a posteriori*. Para la realización de un proceso de tarificación el actuario hace uso de métodos estadísticos, principalmente de técnicas de análisis multivariante. En este trabajo nos centramos en las técnicas estadísticas que se aplican en la tarificación *a priori* del seguro del automóvil, en concreto en el Modelo Lineal Generalizado (MLG) y la Regresión Basada en Distancias (RBD).

Para una correcta tarificación es necesario tener en cuenta las peculiariedades que cada cobertura tiene actualmente en España y en qué modo se realiza la recogida de datos, tanto a nivel de entidad aseguradora individual como a nivel de bases de datos y plataformas del sector asegurador.

El trabajo se ha estructurado como sigue: en el apartado 2 realizamos una breve descripción de los procesos de tarificación de los seguros *no vida*; en el apartado 3 describimos el tipo de datos con el que realizar el cálculo de primas y los métodos estadísticos aplicables para su resolución, en especial el MLG y la RBD; en el apartado 4 discutimos las pecularidades del seguro obligatorio del automóvil en España y en qué modo deben ser tenidas en cuenta en el cálculo de sus tarifas; finalmente, en el apartado 5, comentamos las aportaciones recientes y perspectivas futuras de investigación respecto de la RBD en su aplicación en el ámbito actuarial.

2. Tarificación de los seguros *no vida*

El actuario debe elaborar unas *bases técnicas* que comprendan, entre otros aspectos: información genérica (explicación del riesgo asegurable, los factores de riesgo considerados en la tarifa y los sistemas de tarificación utilizados) e información estadística sobre el riesgo (la estadística utilizada indicando el tamaño de la muestra, las fuentes, el método de obtención y el período a que se refiera). Desde el punto de vista legal los seguros directos se clasifican en los *seguros de vida* y el resto, también denominados de *no vida*. Los seguros de vida cubren el riesgo de muerte y/o supervivencia de las personas y los riesgos complementarios de accidente y enfermedad (véase [8] para un mayor detalle en la clasificación de los ramos). En este trabajo nos ocupamos de la tarificación de los seguros *no vida*, y en especial del seguro del automóvil.

2.1. La prima o precio del seguro

Supongamos que disponemos de una cartera –o conjunto de pólizas– infinita cuyos elementos son idénticos, es decir, que observamos el mismo riesgo de un ramo de *no vida*. Para cada póliza observamos los datos referentes a la *siniestralidad*, recogida en las variables aleatorias *número de siniestros* (en general del período de un año), N, y sus correspondientes *cuantías*, X_i para $i = 1, \ldots, N$. El *coste total* por póliza, S, viene dado por la suma: $S = \sum_{i=1}^{N} X_i$. Si asumimos como hipótesis del proceso de riesgo la de equidistribución de las cuantías de los siniestros, la de independencia entre dichas cuantías y la de independencia entre la cuantía por siniestro y el número

[1] Ley 30/1995 de Ordenación y Supervisión de los Seguros Privados.

[2] El *principio de equidad* se refiere a que la prima se ajuste al riesgo de siniestralidad de cada póliza, es decir, que el asegurado pague según el riesgo que incorpora. Este criterio implica tener en cuenta los factores de riesgo que explican en mayor medida el comportamiento de la estructura aleatoria de siniestralidad.

[3] El *principio de suficiencia* se refiere a que en términos esperados las primas sean suficientes para cubrir todos los riesgos de la cartera considerada y que permitan hacer rentable, en condiciones de estabilidad a largo plazo, a la empresa aseguradora.

de siniestros, la esperanza del *coste total* por póliza, $E[S]$, la calculamos como el producto de la esperanza del número de siniestros por la esperanza de la cuantía de un siniestro: $E[S] = E[N] \times E[X]$. Esta cantidad se denomina *prima pura* con bases de segundo orden, P. Cambiamos el riesgo aleatorio que conlleva la póliza, que puede tomar cualquier valor positivo o nulo, por un valor cierto fijo que coincide con la esperanza matemática del coste total:

$$P = E[N] \times E[X]. \tag{1}$$

La prima pura (1) es la componente base del precio del seguro, con esta cantidad la entidad aseguradora acumula la cantidad suficiente para hacer frente a los siniestros previstos y esperados. Su cálculo debe estar basado en información estadística propia de cada ramo, y en su caso, de cada producto o modalidad de seguro. Por ello, es necesario partir de una base de datos de calidad, tanto a nivel individual de entidad aseguradora, como a nivel sectorial.

En la práctica no se dispone ni de una cartera infinita ni de suficiente información estadística, por lo que la *prima de riesgo* se corresponde con la esperanza matemática (1) más un recargo de seguridad o de solvencia. Además, hasta la obtención del precio final del seguro, añadimos la parte de gastos de gestión interna y externa de la entidad aseguradora, la parte de beneficios y los recargos externos a la prima e impuestos repercutibles (ver [8]). En este trabajo nos ocupamos únicamente del cálculo de la prima pura sin entrar en la aplicación de recargos de solvencia, para lo cual sería necesario tener en cuenta que las diferentes coberturas no son independientes entre sí.

2.2. Tarificación *a priori* o *class-rating*

El sistema de tarificación *a priori* asigna una prima a un riesgo que se incorpora en una cartera sin tener necesariamente experiencia sobre la siniestralidad que conlleva. Únicamente es necesario conocer determinadas características para asignar una siniestralidad esperada y con ella una prima.

Supongamos que disponemos de la experiencia de una cartera para una determinada cobertura de seguro en un período fijado. Para cada póliza, se han observado los datos referentes a las variables aleatorias de *siniestralidad* (número de siniestros y sus correspondientes cuantías), y una serie de factores potenciales de riesgo, los cuales pueden hacer referencia a características del objeto asegurado o a otros condicionamientos de este: características del asegurado, del tomador del seguro, condiciones socio-económicas que lo rodean, etc. Los principios técnicos en que se basa la tarificación *a priori* consisten en un proceso que debe solucionar las siguientes fases ([35], [19], [4], [6]):

1) La determinación de la estructura de tarifa que consiste en ([8]):

- La selección de las *variables de tarifa*: es la elección de los factores de riesgo o características que utilizaremos para distinguir a los asegurados con diferentes riesgos asociados ya que influyen en la siniestralidad. Los factores seleccionados pasan a ser las denominadas variables tarificadoras o de tarifa;
- La determinación de las *clases de tarifa*: es la elección de las clases o agrupaciones de clases de las variables tarificadoras que acaban discriminando a los diferentes grupos de riesgo en la tarifa final;
- La obtención de los *grupos de tarifa*: es la obtención de grupos homogéneos de riesgo, exclusivos y exhaustivos, formados a partir de las clases de tarifa anteriores;
- La inclusión de los gastos en la tarifa;
- Y el tratamiento especial y adecuado de los grandes riesgos.

2) El cálculo de la prima para cada grupo de tarifa, que consiste en la estimación de las primas de riesgo equitativas y suficientes que ajusten la siniestralidad para cada grupo de tarifa a partir de la esperanza del número de siniestros y de la cuantía por siniestro.

3) Por último, se realiza la adecuación de la tarifa al mercado competitivo teniendo en cuenta la competencia de mercado y los segmentos de población a los que va dirigida la cobertura.

Formamos pocos grupos si buscamos una tarifa resultante sencilla y aplicable, que diferencie mínimamente los riesgos de calidades diferentes, o formamos una agrupación más fina si el objetivo es mayor ajuste en la prima individual y más detalle en la tarifa final.

Los pasos de selección de variables tarificadoras, de determinación de las clases de tarifa y de obtención de los grupos de tarifa, dentro de la fase de determinación de la estructura de tarifa, están entre ellos estrechamente vinculados, y su resolución no es independiente en función de las metodologías de análisis estadístico utilizadas (ver [8]).

Para la realización de un proceso de tarificación *a priori* es conveniente que la experiencia en que se basa la tarifa pertenezca a un intervalo temporal lo más cercano posible al de actualización, y son necesarias revisiones periódicas con datos actualizados que repitan el proceso con todas sus fases, empezando por la selección de variables de tarifa ([13]). Si el período de observación es el año más reciente los siniestros pueden ser siniestros en curso o pendientes de reclamación, ya que ese año es el que tiene los siniestros más inmaduros. En tal caso es conveniente realizar, previamente al estudio de tarificación, una revisión de reservas de siniestros pendientes con el objetivo de obtener niveles actualizados de siniestralidad. Este es el caso del seguro obligatorio del automóvil cuando analizamos la cuantía por siniestro para la cobertura de responsabilidad civil de daños personales, ya que el período de maduración de dichos siniestros es largo.

2.3. Tarificación *a posteriori* o *experience-rating*

El sistema de tarificación *a posteriori*, en oposición al de la tarificación *a priori*, parte de una prima inicial para cada unidad de riesgo, individuo o grupo, que se modifica de acuerdo a la experiencia individual o colectiva para dar lugar a las primas de los períodos sucesivos. En un sentido amplio, la expresión *experience-rating* se aplica a todo problema de actualización de tarifas mediante la incorporación de nueva información.

La justificación de estos sistemas está en que dentro de cada clase de riesgo existe heterogeneidad debida a la influencia de ciertos factores de riesgo no considerados (conocidos o desconocidos) o a la incorrecta agrupación de las clases de los sí considerados. La heterogeneidad queda recogida por la siniestralidad de los períodos sucesivos. Al considerar la experiencia propia de cada póliza obtenemos un mayor grado de equidad en las primas de ejercicios posteriores. Incorporamos la información evolutiva mediante un sistema de bonificaciones y penalizaciones (sistema *bonus-malus*).

En el seguro del automóvil ([27], [30], [36], [37]) su aplicación suele tomar como referencia el número de siniestros declarados en un período de un año; si no se han declarado siniestros a las garantías computables se asciende por la escala de descuentos uno o más escalones. Por cada siniestro declarado a las garantías computables se desciende por la escala uno o más tramos, bien reduciéndose los descuentos, bien aplicando recargos. También se han realizado estudios basados en las cuantías de los siniestros declarados ([33]).

3. Tarificación y análisis multivariante

En este apartado, describimos el tipo de datos con el que realizar los estudios de tarificación y las técnicas de análisis multivariante empleadas para su resolución.

3.1. Datos: experiencia de siniestralidad y factores de riesgo

Una fase previa a todo el proceso de tarificación es la recopilación de los datos. Es imprescindible procesar la máxima información en torno al riesgo asegurado. La siniestralidad evoluciona en el tiempo por lo que es posible que a partir de un momento sea explicada por factores no tenidos en cuenta anteriormente.

Una vez recopilados los datos, el paso básico es la selección de las variables de tarifa que influyen por un lado en el número de siniestros, N, y, por otro, la selección de las variables de tarifa que influyen en la cuantía de un siniestro, X. El conjunto de factores de riesgo que explican ambas variables no tienen por qué coincidir. Su selección, hasta la formación de los grupos de tarifa, se realiza mediante métodos estadísticos de análisis multivariante de selección de predictores (ver [8]).

Experiencia de siniestralidad

El objetivo es clasificar las pólizas según el riesgo que incorporan, por lo que la variable aleatoria que recoge la *siniestralidad* (coste total, S, número de siniestros, N, o cuantía de un siniestro, X) es la variable dependiente. El número de siniestros, N, es una variable discreta numérica que toma valores en los naturales. Dependiendo de la metodología estadística utilizada, a veces es tratada como cuantitativa, y a veces como categórica ordinal.

Factores de riesgo

Los *factores potenciales de riesgo*, $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_P)$, son los predictores, mediante alguno de ellos (variables de tarifa) explicamos la estructura de riesgo. Son características "medibles" que observamos y que tienen una posible relación de causa con la siniestralidad objeto de estudio. Es necesario que tengan una definición y que puedan ser representados a través de variables.

En general el conjunto de factores de riesgo es de tipo mixto (mezcla de variables cuantitativas y cualitativas), y es importante disponer de los datos en sus escalas originales. En ocasiones nos encontramos predictores continuos discretizados de antemano (por ejemplo la edad del conductor en intervalos de edad). Esto implica una pérdida de información al pasar a una escala de medida menor. Resulta imposible obtener los datos originales, cuantitativos, de los ya codificados como discretizaciones, pues no sabemos qué valor tomó la variable dentro de cada grupo, sólo sabemos entre qué valores osciló. En este caso no somos capaces de deshacer la agrupación original para realizar otra que proporcione mejores resultados.

El pequeño esfuerzo que representa para una entidad aseguradora la correcta gestión inicial de datos (aparte de la gestión obligada para la ESA[4] que requiere UNESPA[5]) supone una mejora significativa en el largo y costoso proceso de tarificación que ha de servir a largo plazo para la obtención de mayores beneficios. El mercado español es competitivo y ello implica un reto para las compañías que confían en métodos simples de tarificación que no tienen en cuenta la composición de sus carteras.

Datos ponderados:

Cuando nos centramos en un período de observación pueden haber pólizas que no han estado en vigor durante el período completo porque se han incorporado a lo largo del período o porque han vencido a lo largo del mismo y no han renovado. Para reflejar este hecho una posibilidad es extrapolar el resultado de siniestralidad a todo el período, y otra es ponderar la siniestralidad según el tanto por uno de período en que la póliza ha estado viva con respecto al tiempo total ([5]).

Por otro lado, debido al gran volumen de las carteras de seguros ($n \geqslant 10^4$) es usual trabajar con datos agregados. En este caso, los pesos se corresponden con el número de pólizas para el número medio de siniestros y el coste total medio, y con el número de siniestros para la cuantía media por siniestro. Para agregar los n datos consideramos que todos los factores son categóricos (discretizando de una manera adecuada a los continuos (ver [8])) y construimos una tabla cruzada de todos ellos. Pasamos a tener tantas observaciones diferentes, m, como el producto del número de clases de todos los factores menos las combinaciones de celdas vacías.

En el tratamiento del número de siniestros con datos agregados (o en general ponderados) tenemos dos posiblidades ([15]). Una es la de trabajar directamente con la variable aleatoria *número total de siniestros* por celda (o combinación), N_u para $u = 1, \ldots, m$ y otra más habitual es trabajar con la *frecuencia de siniestralidad*. La frecuencia de siniestralidad se obtiene dividiendo la variable aleatoria número total de siniestros por celda, N_u, por el número de expuestos o pólizas de dicha celda, e_u: $\frac{N_u}{e_u}$ para $u = 1, \ldots, m$. La ponderación, w_u, de la frecuencia de siniestralidad de cada celda es $w_u = e_u$.

En el tratamiento de las cuantías de los siniestros con datos agregados también tenemos las dos posibilidades anteriores: trabajar directamente con la variable aleatoria *coste total de los siniestros* por celda, S_u para $u = 1, \ldots, m$; o trabajar con *la cuantía media* por celda construida dividiendo la variable aleatoria coste total de los siniestros por celda, S_u, por el número de siniestros de dicha celda, n_u: $\frac{S_u}{n_u}$ para $u = 1, \ldots, m$. La ponderación de la cuantía media de cada celda es $w_u = n_u$.

Podemos realizar dos procesos de selección de predictores diferentes, uno para el número de siniestros y otro para las cuantías de dichos siniestros. Los conjuntos de variables de tarifa así obtenidos pueden o no coincidir. Como resultado de la tarificación obtenemos una predicción para el número esperado de siniestros por póliza y una para la cuantía esperada de un siniestro, cuyo producto nos proporciona la prima pura (1).

3.2. Modelos de predicción

Los primeros artículos sobre modelos estadísticos en el seguro del automóvil se centran en el número de siniestros prestando menos atención a las

[4]Estadística sectorial del Seguro de Automóviles.
[5]Unión Española de Entidades Aseguradoras y Reaseguradoras.

cuantías. Históricamente el debate principal ha estado basado en si se debía utilizar un modelo aditivo o uno multiplicativo para establecer la relación entre el número medio de siniestros y los factores de riesgo. Los dos modelos son casos particulares del MLG: el modelo aditivo clásico de función de enlace identidad y de error Normal ([28], [29], [**?**]), es decir, el modelo clásico de regresión; y el modelo multiplicativo, de función de enlace logarítmica, combinado con una distribución de Poisson ([38]). En [38] encontramos al MLG presentado como una extensión del modelo clásico de regresión lineal, y como la formalización probabilística motivada en los inicios por el modelo de Bailey y Simons ([2],[3]) entre otros.

El modelo de predicción más sencillo es el modelo clásico de regresión lineal por mínimos cuadrados ordinarios. Pero diversas limitaciones de este modelo (las deficiencias de ajuste en muchas situaciones reales, la restricción a respuestas gausianas o análogas) motivan a buscar modelos más generales que permitan seleccionar la función de predicción de una familia más amplia que la de las funciones lineales, permitiendo alguna no linealidad. En este apartado nos detenemos en dos formas importantes de introducir esta no linealidad correspondientes a los MLG y a la RBD.

Ambos modelos de predicción son extensiones en direcciones distintas del modelo de mínimos cuadrados ordinarios, y están contenidos en la clase mucho más grande de los modelos no lineales. Fijada una clase, encontramos el elemento de ella más adecuado a unos datos concretos de acuerdo con algún criterio de proximidad, como mínimos cuadrados o máxima verosimilitud. Los modelos más sencillos son más robustos, mientras que cuando se permite una familia grande existe el peligro de caer en la sobreparametrización.

Los **Modelos Lineales Generalizados ([31], [20])** han sido muy aplicados en tarificación ([26], [15], [7], [8]).

Cuando tarificamos respecto de la cuantía de los siniestros es apropiado utilizar una distribución Gamma o una Gaussiana Inversa preferiblemente a una Normal ya que estas distribuciones no toman valores negativos y tienen asimetría positiva; y cuando tarificamos respecto del número de siniestros es apropiado utilizar una distribución de Poisson, una Binomial o una Binomial Negativa. Véase [31] pp. 337, 400 y 413-414 para un estudio detallado sobre la función de enlace y la distribución de error óptimas para unos datos determinados y una selección de predictores concreta. Los datos empleados en la referencia anterior son datos actuariales de cuantías de siniestros extraidos de [1] referentes a la cobertura de daños propios, estos mismos datos se han empleado en [11] con RBD.

En general, el MLG más apropiado para unos datos es el que proporciona una menor desvianza –la medida que en estos modelos generaliza la suma de cuadrados residual. Las diferentes maneras de reducirla son ([32]): variar la función de enlace, variar la distribución del error, y/o variar las variables de tarifa incluidas en el modelo. Lo usual, aunque no óptimo, es fijar una función de enlace y una distribución del error y *a posteriori* realizar la selección de variables de tarifa mediante un proceso de selección de predictores haciendo uso del estadístico de test F basado en desvianzas ([15], [34], [38]). Los programas informáticos que ofrecen las consultorías a las entidades aseguradoras suelen implementar únicamente y de manera cerrada el MLG de distribución del error Gamma para las cuantías y de Poisson para el número de siniestros, pudiendo elegir entre el link identidad para componer una tarifa aditiva y el logarítmico para una multiplicativa. El MLG resultante permite cubrir la fase de selección de variables de tarifa y también el cálculo de la prima pura (1). Esto no ocurre si la técnica estadística utilizada para la selección de variables de tarifa se basa por ejemplo en análisis *cluster* o análisis discriminante (ver [8]).

La **Regresión Basada en Distancias** fue propuesta por Cuadras en [16]. Detalles y aportaciones posteriores pueden verse en [17], [18], [8] y [12].

La RBD es una extensión del modelo clásico de regresión: la información aportada por las variables de tarifa queda reflejada en una matriz de distancias, intuitivamente la respuesta se proyecta en el espacio euclídeo obtenido mediante escalado multidimensional métrico de esta distancia. Con predictores cuantitativos y métrica l^2 se obtiene como caso particular el modelo de mínimos cuadrados ordinarios.

Un resumen del procedimiento es como sigue: partimos de una respuesta continua, **y,** centrada, y un conjunto de predictores **F** (que pueden ser de tipo mixto), mediante una métrica euclídea ([14], [24]) calculamos la matriz de distancias al cuadra-

do, $\mathbf{\Delta}^2$, y la de productos escalares $\mathbf{G} = -\frac{1}{2}\mathbf{J}\mathbf{\Delta}^2\mathbf{J}$, donde $\mathbf{J}$ es la matriz de centrado $\mathbf{J} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. Una descomposición $\mathbf{X}\mathbf{X}^T = \mathbf{G}$ (que existe por la condición euclídea) da una configuración euclídea centrada $\mathbf{X}$, sobre la que aplicamos mínimos cuadrados ordinarios. La predicción, $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ donde $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, puede obtenerse directamente sin explicitar $\mathbf{X}$ como $\mathbf{H} = \mathbf{G}^+\mathbf{G}$ siendo $\mathbf{G}^+$ la pseudo-inversa de Moore-Penrose de $\mathbf{G}$ ([21]). Calculamos la predicción para un nuevo individuo $\{n+1\}$ haciendo uso de la fórmula de interpolación de Gower ([22], [25]) y obtenemos la fórmula de predicción $\widehat{\mathbf{y}}_{n+1} = \frac{1}{2}(\mathbf{g} - \mathbf{d})\mathbf{G}^+\mathbf{y}$, donde $\mathbf{g}$ es el vector fila que contiene la diagonal de $\mathbf{G}$ y $\mathbf{d}$ es el vector fila que contiene las distancias del nuevo individuo al resto, $\mathbf{d} = (d^2_{n+1,1}, \ldots, d^2_{n+1,n})^T$. En el caso general de predictores de tipo mixto podemos utilizar el coeficiente de similaridad de Gower ([23], [24]).

Hemos construido implícitamente, mediante la distancia, una función no lineal $\mathbf{X} = \phi(\mathbf{F})$ que relaciona las variables observadas $\mathbf{F}$ con los predictores euclídeos latentes $\mathbf{X}$.

Esta no linealidad hace que la selección de variables de tarifa sea más complicada que en el modelo clásico, pues no es aplicable el test F usual. En [12] proponemos una solución adaptando un método de *bootstrap no paramétrico* para la estimación de los p-valores de un estadístico F generalizado. Al igual que el MLG, la RBD permite cubrir todas las fases de la tarificación hasta la estimación de las primas puras.

4. Tarificación del seguro del automóvil en España en función de las coberturas

El seguro obligatorio del automóvil cubre la responsabilidad civil a terceros respecto a los daños personales y materiales ocasionados a terceras personas como consecuencia de un hecho de la circulación. Ambas coberturas tienen frecuencia de siniestralidad y cuantía media de un siniestro muy diferentes, por lo que diferenciamos la información con el objetivo de calcular la esperanza del coste total por póliza, (1) por separado. Por el mismo motivo tratamos cada cobertura adicional al seguro obligatorio del automóvil hasta llegar a un *todo riesgo* (responsabilidad civil suplementaria, daños propios, rotura de lunas, incendio, robo, ...) también por separado. Las *primas puras totales por póliza* las calculamos como la suma aritmética de las primas puras correspondientes a cada cobertura ([26]). Si denotamos por λ^c al número esperado de siniestros y por m^c a la cuantía esperada de un siniestro respecto a la cobertura c (siendo c: daños materiales, daños personales, daños propios, etc), la *prima pura total* es:

$$PT = \sum_c \lambda^c \times m^c. \tag{2}$$

En el proceso de tarificación realizamos la selección de variables de tarifa por separado para cada cobertura y además, dentro de cada cobertura, seleccionamos el conjunto de predictores respecto del número de siniestros y respecto de la cuantía de un siniestro. Los conjuntos de variables de tarifa pueden o no coincidir. Como resultado obtenemos una predicción para el número esperado de siniestros por póliza y una para la cuantía esperada de un siniestro para cada cobertura, lo que nos permite calcular la prima pura total (2). El estudio de tarificación puede llevarse a cabo con los datos de la cartera o haciendo uso de las estadísticas basadas en datos sectoriales.

En el seguro del automóvil, y dependiendo de la cobertura, los factores tenidos en cuenta son:

- *Factores relativos al vehículo asegurado*: valor, antigüedad, categoría, clase, tipo, marca, modelo, número de plazas, potencia, peso, o relación potencia / peso, color, etc

- *Factores relativos al conductor*: edad, sexo, antigüedad del carnet, estado civil, profesión, número de hijos, posibilidad de conductores ocasionales, resultado de la experiencia en el pasado, etc

- *Factores relativos a la circulación*: zona de circulación, uso del vehículo, kilómetros anuales, etc

En la práctica, la calidad del ajuste, tanto si utilizamos el MLG o la RBD depende de la cobertura y de si analizamos el número o las cuantías de los siniestros. Por ejemplo, para la cobertura de daños materiales es usual obtener un buen ajuste cuando analizamos el número de siniestros con predictores como la edad del conductor, la antigüedad del carnet, la potencia del vehículo,... En cambio, cuando analizamos las cuantías de los siniestros en la cobertura de daños personales es difícil obtener un buen ajuste ya que las cuantías dependen en gran medida de elementos que la entidad aseguradora no puede controlar *a priori*.

4.1. Escenario actual del seguro del automóvil en España

Para una correcta tarificación del seguro del automóvil debemos tener presentes las pecularidades que dicho seguro tiene en España. En los últimos años han habido cambios gracias a la utilización a nivel sectorial de las nuevas tecnologías de la informática y de las telecomunicaciones por parte de UNESPA y gracias al servicio transaccional prestado por TIREA[6]. En unos años los servicios han crecido de forma importante (1994: CICOS[7], 2002: SDM[8], 2000: SINCO[9], 2001: ESA,...). Nos referimos a [9] para un mayor detalle.

La Estadística sectorial del Seguro de Automóviles, ESA

UNESPA ha realizado estadísticas comunes en papel hasta 1997. Estas estadísticas han permitido a las entidades aseguradoras disponer de información estadística detallada a nivel sectorial. Actualmente, la ESA, desde el año 2001, ha retomado el análisis estadístico. El servicio ESA está compuesto por una base de datos con el total de expuestos y siniestros aportados por las entidades aseguradoras colaboradoras. En esta nueva estadística sólo las entidades aseguradoras que colaboran de forma activa tienen acceso a los resultados agregados de las explotaciones estadísticas, cosa que implica un mayor grado de compromiso por parte de las entidades aseguradoras y una mayor validez en los resultados. Las entidades aseguradoras no participantes reciben tan sólo un documento-resumen con información breve y genérica.

Dada la magnitud de las nuevas tecnologías, las consultas de las estadísticas descriptivas resultantes de la explotación se realizan de forma on-line, existiendo la posibilidad de descargarlas en ficheros de tipo Excel. Es posible solicitar dos tipos de explotación: explotación global, con el total de datos cargados por la totalidad de las entidades aseguradoras colaboradoras, y explotación individual, para cada una de las entidades aseguradoras participantes. De cada explotación, se puede obtener un resumen ejecutivo, o un informe con información de detalle. El resumen ejecutivo proporciona un resumen para cada una de las garantías contempladas, por año de estadística y tipo de vehículo. La información de detalle se divide en dos grandes bloques: por un lado todo lo referente a responsabilidad civil, y por otro, daños propios, incendio, lunas, robo, ocupantes, retirada de carnet y defensa Jurídica. Además se recogen aparte los casos especiales de flotas y mercancías peligrosas. Como novedad se incorpora información sobre el segundo conductor.

Con el fin de actualizar la información, en esta estadística se han modificado algunas de las variables de tarifa utilizadas históricamente. Por ejemplo, la situación del riesgo se mide por la provincia de la póliza y no por el lugar de ocurrencia del siniestro. También se han actualizado las categorías y usos de los vehículos, y adicionalmente se utiliza la información técnica de cada vehículo aportada por el código de Base SIETE[10]. El código base SIETE incluye codificadas las principales características técnicas de todos los vehículos susceptibles de ser asegurados en España y radica en Centro Zaragoza. Concretamente, las variables de tarifa por las que se pueden obtener resultados son: la edad, la antigüedad del carnet y el sexo del conductor, la provincia y la antigüedad de la póliza, el tipo (y sus subcategorías), el uso, la antigüedad y el valor del vehículo.

El servicio ESA (que funciona desde el año 2001, y por lo tanto es un servicio bastante reciente) era un servicio totalmente necesario para llenar el vacío de información técnica del riesgo elemental producido en el Sector Asegurador desde la última elaboración de la Estadística Común de Automóviles del año 1997. De cara a la tarificación es importante resaltar que en la ESA las variables de tarifa utilizadas y sus clases han sido actualizadas. A diferencia de las estadísticas comunes anteriores que incluían informes actuariales, los resultados de la ESA son meramente descriptivos, pero permiten a las entidades aseguradoras descargar las consultas en ficheros con los que realizar sus propios estudios.

Respecto a las garantías correspondientes a la responsabilidad civil obligatoria en que se puede desglosar la estadística, éstas son: responsabilidad civil de daños materiales, responsabilidad civil de

[6] Tecnologías de la Información y Redes para las Entidades Aseguradoras.
[7] Centro Informático de Compensación de Siniestros.
[8] Servicio de gestión de Siniestros de Daños Materiales.
[9] Fichero histórico de SINiestralidad de COnductores.
[10] Sistema Informativo de Especificaciones Técnicas.

daños corporales, responsabilidad civil total y, como novedad en las dos últimas estadísticas, responsabilidad civil de siniestros tramitados por el sistema CICOS y que son acreedores.

5. Aportaciones recientes y perspectivas de futuro

- En el apartado 3.1 hemos visto que es habitual trabajar con datos ponderados, por lo que para completar la aplicabilidad de la RBD en la tarificación construimos la versión heteroscedástica del modelo de predicción y adaptamos a este caso la metodología bootstrap de selección de predictores incluido el estadístico de test que se desarrolla para el caso homocedástico en [12]. Parte de estos resultados se han presentado ya en el *XXIX Congreso Nacional de Estadística e investigación Operativa* celebrado en Tenerife en mayo de 2006 ([11]). En este trabajo se presenta la formulación y una aplicación con los datos de [1] referentes a cuantías de siniestros para la cobertura de daños propios. Obtenemos mejores resultados de ajuste con la RBD que con el MLG. Como era de esperar para esta cobertura, las variables de tarifa resultantes hacen referencia a características del vehículo.

- Otro grupo de desarrollos se refiere a problemas numéricos o computacionales de la RBD. Dentro de este ámbito hemos presentado el trabajo *Implementing PLS for distance-based regression* ([10]) en el que adaptamos los Mínimos Cuadrados Parciales a la RBD con aplicación carteras de seguros del mercado Español moderadamente grandes, $n \sim 10^4$.

- Por otro lado, fijado un conjunto de predictores, obtenemos una mayor flexibilidad si seleccionamos de forma adaptativa la distancia de una familia de métricas euclídeas indexadas por un hiperparámetro θ. De este modo el ajuste del modelo resultante es óptimo dentro de una familia de métricas dada. Unos primeros resultados pueden encontrarse en [21].

- Finalmente, una extensión de la RBD es la RBD-MLG, que consiste en incorporar una función de enlace que relacione la esperanza de la respuesta con el predictor lineal, en este caso una combinación lineal de los predictores euclideos $\mathbf{X}$.

Agradecimientos

Este trabajo está financiado, en parte, por el Ministerio de Ciencia y Tecnología y FEDER, proyecto MTM2006-09920/ que tiene de título *Análisis Multivariante No Lineal: Técnicas No Paramétricas y Basadas en Distancias* y del que es investigador principal el Dr. Pedro Delicado Useros del Departamento de Estadística e Investigación Operativa de la Universidad Politécnica de Cataluña. Agradezco las sugerencias de mis compañeros M^a Mercè Claramunt y Josep Fortiana.

Referencias

[1] Baxter L.A., Coutts S. M., and Ross G.A.F. (1980). Applications of Linear Models in Motor Insurance. *Transactions of the 21st International Congress of Actuaries*, **2**, 11-29.

[2] Bailey R.A., and Simon L.J. (1960). Two studies in automobile insurance ratemaking. *Actuarial Studies in Non-Life Insurance Bulletin*, **1:4**, 192-217.

[3] Bailey R.A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society*, **50**, 4-11.

[4] Boj E., Claramunt M. M., and Fortiana J. (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 6**, 11-35.

[5] Boj E., Claramunt M.M., and Fortiana J. (2001). Weighted metric scaling applied to automobile insurance data when the exposure base is not the unit. *Proceedings of the Fifth International congress on Insurance: Mathematics and Economics*. Penn State University.

[6] Boj E., Claramunt M.M., and Fortiana J. (2001). Herramientas estadísticas para el estudio de perfiles de riesgo. *Anales del Instituto de Actuarios Españoles*, **Tercera Época 7**, 59-89.

[7] Boj E., Claramunt M.M., Fortiana J., and Vidiella A. (2002). The use of distance-based regression and generalised linear models in the rate making process. An empirical study. *Mathematics Preprint Series*, Institut de Matemàtica de la Universitat de Barcelona 305.

[8] Boj E., Claramunt M.M., and Fortiana J. (2004). Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación. *Cuadernos de la Fundación MAPFRE Estudios*, **88**.

[9] Boj E., Claramunt M.M., Fortiana J., and Vegas A. (2005a). Bases de datos y estadísticas del seguro de automóviles en España: influencia en el cálculo de primas. *Estadística Española*, **47**, 539-566.

[10] Boj E., Claramunt M.M., Grané A., and Fortiana J. (2005b). Implementing PLS for Distance-Based Regression: Computational issues. *PLS AND RELATED METHODS. Proceedings of the PLS'05 International Symposium*, 153-158.

[11] Boj E., Claramunt M.M., and Fortiana J. (2006). Regresión basada en distancias en presencia de heteroscedasticidad. *XXIX Congreso Nacional de Estadística e investigación Operativa*, 279-280, Mayo de 2006, Tenerife.

[12] Boj E., Claramunt M.M., and Fortiana J. (2007). Selection of Predictors in Distance-Based Regression. *Communications in Statistics – Simulation and Computation*, **36:1**, (to appear).

[13] Booth P., Chadburn R., Cooper D., Haberman S., and James D. (1999). *Modern Actuarial Theory and Practice*. Chapman & Hall. Boca Raton (California).

[14] Borg I., and Groenen P. (2005). *Modern multidimensional scaling: theory and applications*. Second Edition. Springer-Verlag. New York.

[15] Brockman M.J., and Wright T.S. (1992). Statistical Motor Rating: Making Effective Use of your Data. *Journal of the Institute of Actuaries*, **119:3**, 457-543.

[16] Cuadras C.M. (1989). Distance Analysis in Discrimination and Classification using both Continuous and Categorical Variables. In: Dodge, Y. (ed.), Recent Developments in *Statistical Data Analysis and Inference*. Elsevier Science Publisher, North-Holland, Amsterdam, 459-474.

[17] Cuadras C.M., and Arenas C. (1990). A distance-based model for prediction with mixed data. *Communications in Statistics – Theory and Methods*, **19**, 2261-2279.

[18] Cuadras C.M., Arenas A., and Fortiana J. (1996). Some Computational Aspects of a Distance-Based Model for prediction. *Communications in Statistics – Simulation and Computation*, **25**, 593-609.

[19] de Wit G.W. (1986). Risk Theory, a Tool for Management. In: M. Goovaerts et al. eds., *Insurance and Risk Theory*. Reidel. Dordrecht-Boston, MA, 7-17.

[20] Dobson A.J. (2001). *An Introduction to Generalized Linear Models*. Second Edition. Chapman & Hall. London.

[21] Esteve A. (2003). *Distancias estadísticas y relaciones de dependencia entre conjuntos de variables*. Tesis Doctoral. Universidad de Barcelona.

[22] Gower J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika*, **53**, 325-338.

[23] Gower J.C. (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 857-874.

[24] Gower J.C., and Legendre P. (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, **3**, 5-48.

[25] Gower J.C., and Hand D.J. (1996). *Biplots*. London: Chapman and Hall.

[26] Haberman S., and Renshaw A.E. (1996). Generalized Linear Models and Actuarial Science. *The Statistician*, **45:4**, 407-436.

[27] Heras A., Vilar J.L., and Gil J.A. (2002). Asymptotic fairness of Bonus-Malus System and optimal scales of premiums. *The Geneva Papers on Risk and Insurance Theory*, **27**, 61-82.

[28] Lemaire J. (1977). Selection procedures of regression analysis applied to automobile insurance. *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker*, **77:2**, 143-160.

[29] 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**, 65-72.

[30] Lemaire J. (1995). *Bonus-malus system in automobile insurance*. Kluwer-Nijhof Publishing. Boston, MA.

[31] McCullagh P., and Nelder J.A. (1989). *Generalized Linear Models*. Second Edition. Chapman & Hall. London.

[32] Millenhall S.J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society*, **86**, 393-487.

[33] Morillo I. (2001). *Sistemas de Bonus-Malus: Comparaciones y alternativas*. Tesis Doctoral. Universidad de Barcelona.

[34] Stroinski K.J., and Currie I.D. (1989). Selection of Variables for Motor Insurance Rating. *Insurance: Mathematics and Economics*, **8**, 35-46.

[35] van Eeghen J., Greup E.K., and Nijssen J.A. (1983). *Rate Making*. Survey of Actuarial Science, **2**, Nationale-Nederlanden N. V. Rotterdam.

[36] Vegas A. (1993). Fundamentos técnicos del sistema Bonus-Malus. *Previsión y Seguro*, **22**, 141-167.

[37] Vilar J.L., Gil J.A., and Heras A. (2004). Estudio de la estructura de una cartera de pólizas y de la eficiencia de un sistema Bonus-Malus. *Cuadernos de la Fundación MAPFRE Estudios*, **84**.

[38] Zehnwirth B. (1994). Ratemaking: from Bailey and Simon (1960) to Generalised Linear Regression Models. *Casualty Actuarial Society Winter Forum*, 615-659.