embargo, y dado que se trata de una publicación técnica enmarcada en el seno de una entidad profesional, se pretende que este tipo de trabajos vayan ilustrados en todo caso por una aplicación práctica, con el fin de garantizar así una comunicación más directa y eficaz entre ambos enfoques, teórico y práctico.

Creemos que es necesario realizar un esfuerzo en esta línea para que los *Anales del IAE* continúen estando al servicio de todos los colegiados, suponiendo un foro de intercambio de puntos de vista que permita compartir el acervo de reflexiones y conocimientos adquiridos por todos los actuarios en el desarrollo de su actividad, cualquiera que sea la naturaleza de ésta.

Es posible que de esta forma, con el tiempo, lleguen a los *Anales del IAE* más artículos surgidos de la colaboración conjunta entre firmantes académicos y actuarios que ejercen la profesión, lo que, sin duda, resulta una mixtura enormemente enriquecedora para todo el colectivo.

En estos Anales correspondientes al año 2000, se incluyen una serie de trabajos que, por falta de espacio, no pudieron incluirse en números anteriores, por lo que agradecemos a los autores no sólo su valiosa contribución sino también su paciencia.

Por último, y para finalizar, quisiéramos reiterar nuestro agradecimiento expreso a D. Angel Vegas Montaner, anterior coordinador de los *Anales del IAE*, así como al Comité Científico, por su esmero y dedicación.

Rosa M. Mayoral Martínez
Coordinadora Comisión de Publicaciones

UNA ALTERNATIVA EN LA SELECCIÓN DE LOS FACTORES DE RIESGO A UTILIZAR EN EL CÁLCULO DE PRIMAS

Eva Boj del Val[1], Mª Mercè Claramunt Bielsa[2] y
Josep Fortiana Gregori[3]

RESUMEN

Se presenta un método de selección paso a paso de variables predictoras en el modelo general de regresión basada en distancias ([7],[8],[9]), aplicable cuando disponemos de una variable respuesta continua univariante y de un conjunto de predictores potenciales de tipo mixto, incluyendo variables continuas y categóricas ([4]). Éste se propone como herramienta alternativa para cubrir, dentro del proceso de tarificación *a priori* que se realiza en los seguros *no vida*, la fase de elección de variables de tarifa a partir del conjunto de factores potenciales de riesgo ([3]), obteniendo resultados comparables a los de otros métodos utilizados en este contexto. Se finaliza con unas ilustraciones que ponen de manifiesto tanto la aplicabilidad, como las propiedades básicas del procedimiento.

PALABRAS Y FRASES CLAVE: selección de predictores, regresión basada en distancias, distancias estadísticas, bootstrap, tarificación *a priori no vida*.

[1] Profesora Ayudante del Departament de Matemàtica Econòmica, Financera i Actuarial de la Universitat de Barcelona.
[2] Profesora Titular de Universidad del Departament de Matemàtica Econòmica, Financera i Actuarial de la Universitat de Barcelona.
[3] Profesor Titular de Universidad del Departament d'Estadística de la Universitat de Barcelona.

1. INTRODUCCIÓN

1.1. Generalidades

Un seguro es un servicio de seguridad ofrecido por una entidad económica o ente asegurador, por el cual el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura, a indemnizar, dentro de los límites pactados, el daño producido al asegurado o a satisfacer un capital, una renta u otras prestaciones convenidas. El riesgo de forma genérica lo definimos como el acontecimiento incierto, independiente de la voluntad exclusiva de las partes y cuya realización implica, normalmente, consecuencias desfavorables para el asegurado ([13]). Así, un contrato de seguro tiene dos componentes básicas e imprescindibles para su viabilidad: la existencia de un riesgo y el pago de un precio por su cobertura, es decir, la prima. La prima, que es función del riesgo asegurable y de los restantes factores que integran el coste de la empresa, es el precio del servicio más el margen explícito de beneficio y debe cumplir los principios de equidad y suficiencia de acuerdo con la naturaleza de los riesgos asumidos por el asegurador. El principio de equidad se refiere a que la prima o precio del seguro se ajuste al riesgo de siniestralidad de cada póliza. El principio de suficiencia se refiere a que en términos esperados las primas sean suficientes para cubrir todos los riesgos de la cartera considerada, es decir, que permitan hacer rentable, en condiciones de estabilidad a largo plazo, a la empresa aseguradora.

Notemos que el precio forma parte del servicio concebido como producto puesto a disposición del cliente y preparado para su introducción en el mercado. Los Actuarios, deben tener presente la técnica actuarial que hace referencia a los verdaderos costes de siniestralidad, gastos y márgenes que en todo momento han de dominar la estructura del precio, pero se comprende que, en su política de precios, existirá una componente de adecuación al mercado y al poder adquisitivo de cada segmento de población al que irá dirigido cada producto o cobertura. Así, el equilibrio entre las tendencias que defienden la asequibilidad del precio, por una parte, y su equidad y suficiencia, por otra, se hace primordial a la hora de diseñar la política del precio.

1.2. Tarificación

El objetivo de un proceso de tarificación es la obtención de unas primas o precios del seguro equitativas para cada riesgo, teniendo en cuenta la solvencia del asegurador. Las primas deben responder a los principios de libertad de competencia y han de estar fundadas en la equidad y suficiencia. Los principios técnicos en que se basa la elaboración de una tarifa constituirán el sistema de tarificación. Distinguimos, en el campo actuarial, dos sistemas de tarificación: tarificación *a priori* o *class-rating* y tarificación *a posteriori* o *experience-rating*.

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

Partimos de la experiencia de una cartera para un determinado riesgo en un período fijado (en general 1 año), en la que se ha observado para cada individuo la siniestralidad y una serie de características o factores potenciales del riesgo observado. Se trata de realizar un proceso que pasará por las siguientes fases ([10],[23]):

a) Determinación de la estructura de tarifa, resolviendo:

- Selección de las variables tarificadoras: elección de los factores de riesgo o características que utilizaremos para distinguir a los asegurados con diferentes riesgos asociados, puesto que influirán en la siniestralidad. Los factores seleccionados pasarán a ser variables tarificadoras;

- Determinación de las clases de tarifa: elección de las clases o agrupaciones de clases de las variables tarificadoras anteriormente seleccionadas, que acabarán discriminando a los diferentes grupos de riesgo en la tarifa final;

- Obtención de los grupos de tarifa: formación de grupos homogéneos de riesgo, exclusivos y exhaustivos, formados a partir de las clases de tarifa anteriores;

- Inclusión de los gastos en la tarifa;

- Tratamiento adecuado de los grandes riesgos.

b) Cálculo de un nivel adecuado de prima para cada grupo de tarifa: estimación de las primas de riesgo (equitativas y suficientes), que ajusten la siniestralidad para cada grupo de tarifa, estudiando las distribuciones de la media del número de siniestros y del coste medio por siniestro, y obteniendo la prima pura de la clasificación, con la posibilidad de adaptarlo a un modelo aditivo, multiplicativo u otro.

c) Y por último, la implementación de la tarifa en un mercado competitivo: será la adecuación de la tarifa a la práctica. A parte de la justificación teórica de la selección de los factores de riesgo es necesario tener presente la competencia de mercado y los segmentos de población a los cuales va dirigida la cobertura. Hay factores de riesgo que por su propia naturaleza supondrían una discriminación indeseada y el producto no sería aceptado y, sin embargo, hay otros que invitan a ser elegidos por la relación intuitiva que merecen con el riesgo y que serían fácilmente aceptados.

Es usual y conveniente seleccionar un número limitado de factores, ya que a mayor número de factores de riesgo, más complicada y cara resultará la administración para el asegurador; con más factores la correlación entre estos incrementará rápidamente; los modelos estadísticos empleados sufrirán problemas de sobreparametrización; y además, sólo seremos capaces de utilizar factores medibles.

Dejando a un lado el aspecto de mercado y centrando la atención en la parte técnica actuarial, dado el objetivo de equidad y suficiencia en las primas, buscaremos la formación de grupos de riesgo homogéneos determinados por combinaciones de clases de tarifa, que tendrán internamente una siniestralidad esperada similar y por lo tanto poca dispersión entorno a su valor esperado. Podrá interesarnos formar pocos grupos si buscamos una tarifa resultante sencilla y aplicable, que diferencie sólo mínimamente los riegos de calidades diferentes, o podrá interesarnos formar una agrupación más fina, es decir, con más grupos, si el objetivo es mayor ajuste en la prima individual y más detalle en la tarifa final.

Es conveniente que la experiencia en que se basará la tarifa pertenezca a un intervalo temporal lo más cercano posible al momento de actualización, y serán necesarias revisiones periódicas con datos actualizados que repetirán el proceso con todas sus fases, pues no hay que dejar de lado a los predictores actualmente no aptos, ni a otros factores potenciales no considerados anteriormente, ya que los riesgos podrían evolucionar de tal forma que se vieran influenciados por estos ([22]).

Cae por su propio peso la importancia de realizar correctamente los pasos de la fase inicial de determinación de la estructura de tarifa para una correcta realización de la tarifa final, formando parte esta fase del informe técnico actuarial en la justificación de las primas resultantes para la Dirección General de Seguros.

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

Se parte de una prima inicial para cada unidad de riesgo (individuo o grupo), que se va modificando en períodos sucesivos de acuerdo con la experiencia individual o colectiva. 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 si considerados, que pondrá de manifiesto la siniestralidad con el transcurso del tiempo. Al considerar esta experiencia obtendremos un mayor grado de equidad en las primas de los ejercicios posteriores. Una manera de incorporar la información evolutiva de los riesgos es realizar un sistema de bonificaciones y penalizaciones (*bonus-malus*) de acuerdo con los resultados obtenidos ([18],[21]).

Cabe notar, que en este caso también sería interesante realizar un estudio de cuáles serán los factores de riesgo influyentes en la siniestralidad de la prima inicial de que parte el estudio, pues aunque se suponen dados, de esta forma la heterogeneidad que se intenta corregir con las bonificaciones y penalizaciones será menor y éstas a su vez más leves.

1.3. Factores de riesgo

En general, para la realización de un proceso de tarificación *a priori* de un ramo de *no vida*, dispondremos de la experiencia de siniestralidad de una cartera de riesgos compuesta por n pólizas, para un período determinado (en general 1 año). De cada póliza tendremos, para el período de observación, la siniestralidad y una serie de características o factores potenciales de riesgo.

La variable aleatoria Y que recogerá la siniestralidad de una póliza, podrá venir representada por:

N: variable aleatoria número de siniestros para el período de observación ,

X: variable aleatoria cuantía de un siniestro dentro del período de observación,

Z: variable aleatoria cuantía total de todos los siniestros para el período de observación, $Z = X_1 + X_2 + \cdots + X_N$.

Por hipótesis de la *Teoría Clásica del Riesgo*, según el principio de equivalencia o de la prima pura, $P = E[Z] = E[N]_N \cdot E[X]_X = \overline{n} \cdot \overline{x}$. Por lo que la esperanza de la cuantía total podrá expresarse como el producto de la esperanza del número y de la cuantía, y será el valor cierto que sustituirá a la variable aleatoria cuantía total de los siniestros Z. Dada tal relación, estaremos interesados en estudiar el comportamiento aleatorio de la tres variables, de forma independizada, así como la relación de dependencia entre X y N. En realidad, el interés de estudio estará centrado en la variable cuantía total Z, pero nos será más complicado procesar la información total, por lo que en la mayoría de ocasiones estudiaremos por separado el comportamiento de la cuantía y del número.

Los factores de riesgo serán características medibles que habremos observado y que tendrán una posible relación de causa con la siniestralidad objeto de estudio. Estos podrán hacer referencia tanto a características del objeto asegurado como a otros condicionamientos

de éste: características del asegurado, del tomador, condiciones socio-económicas que lo rodean, etc. Tan sólo es necesario que puedan tener alguna definición y ser representadas como variables, además de ser tenidas en cuenta de modo expreso como datos codificados disponibles para posibles estudios. Por ejemplo en los seguros de *vida*, los dos factores de riesgo considerados principalmente son la edad y el sexo del asegurado, un número limitado en comparación con los ramos de *no vida*; en los seguros de *no vida* de responsabilidad civil obligatoria de automóviles: la categoría del vehículo, la potencia del vehículo, la zona de circulación, el uso del vehículo, la antigüedad del permiso de conducir, la edad del conductor habitual, la marca, modelo y versión del vehículo; en los seguros de *no vida* de robo a comercios: el capital asegurado, el tipo de comercio (de deportes, de arte, de moda y confección, supermercados, agencias de transportes), etc.

Una fase previa a todo el proceso de tarificación incluido el paso de selección de variables tarificadoras sería "la selección o recopilación de posibles factores potenciales de riesgo", aunque la información buscada quedara implícita en preguntas suaves que el asegurado (o tomador en su caso) estuviera dispuesto a responder. Es imprescindible conocer y procesar la máxima información en torno al riesgo asegurado, ya que como la siniestralidad va evoltucionando en el tiempo, factores que hoy no son incluidos en la tarifa (porque no explican suficientemente el riesgo, o porque nunca habían sido considerados como posibles factores), puedan serlo en un futuro ([6]).

En general, las variables consideradas en un estudio pueden ser clasificadas de diferentes formas:

- Según el objetivo y la interpretación del análisis:
  - Variables respuesta o dependientes
  - Variables intermedias: son las que son tratadas como respuesta para algunas variables y como explicativas para otras
  - Variables explicativas o independientes

- Según la estructura de los posibles valores:
  - Variables cuantitativas
  - Variables cualitativas:
    - Nominales: las diferentes categorías no tienen ningún tipo de ordenación
    - Ordinal: las diferentes categorías tienen implícita alguna ordenación natural

Respecto al objetivo e interpretación del análisis, en el caso de la tarificación *a priori no vida*, se trata de clasificar las *n* pólizas según el riesgo que incorporan, por lo que la variable aleatoria *Y* que recoge la siniestralidad jugará el papel de variable dependiente. Los factores de riesgo serán las variables independientes o explicativas, pues a través de ellas seremos capaces de explicar la estructura de riesgo de la siniestralidad. A las variables explicativas también se les llama variables predictoras o predictores a secas ya que servirán para la predicción de la variable dependiente, a la que por el mismo motivo se la denomina variable respuesta.

Es usual que los datos disponibles de una cartera incorporen predictores categóricos (tanto nominales, como ordinales), que se correspondan con variables de naturaleza categórica (por ejemplo el sexo del conductor o el color del coche). Pero además, en muchas ocasiones, encontraremos predictores continuos discretizados de antemano (por ejemplo la edad del conductor o la antigüedad del carnet) con la pérdida de información que este proceso de discretización supone. Esto tiene dos consecuencias:

- para la selección de variables tarificadoras: se hace necesario utilizar procesos de selección con modelos capaces de contemplar predictores de tipo mixto (mezcla de cualitativos y cuantitativos) en el caso de existencia de variables continuas en combinación con categóricas, pues si alteramos la naturaleza de los datos discretizando variables continuas de antemano, las técnicas estadísticas proporcionarán resultados erróneos, debido a que la mayoría de ellas no son robustas en el sentido de que pequeños cambios en los datos conducen a resultados diferentes;

- para la formación o en su caso agrupación de las clases de tarifa: si no guardamos los datos originales de los predictores continuos y únicamente disponemos de los valores agrupados según un determinado criterio, no podremos deshacer esta agrupación original, para realizar otras que quizás proporcionan mejores resultados. Resulta imposible obtener los datos originales continuos de los ya codificados como discretizaciones, pues no sabremos qué valor tomó la variable dentro de cada grupo, sólo sabremos entre qué valores osciló.

En conclusión a este comentario, notamos que el pequeño esfuerzo que representa la correcta gestión inicial de datos (caso de no ser estadísticas comunes realizadas por entidades de interés, por tener dificultades añadidas), llevará a una mejora significativa al largo y costoso proceso de tarificación que servirá, a largo plazo a la empresa aseguradora, en la obtención de mayores beneficios y mejor gestión de los riesgos de su cartera.

Notemos que, si un predictor seleccionado como variable de tarifa es continuo, acabaremos discretizándolo durante el proceso de tarificación para obtener las clases de tarifa basándonos en algún criterio, pues si asignáramos a cada valor continuo una categoría, nada más que un individuo nuevo tomara otro valor, ya no seriamos capaces de clasificarlo en ningún grupo de tarifa, a no ser que el predictor no interviniera en ninguna partición. Pero tal discretización debería ser realizada posteriormente a la obtención del conjunto de variables de tarifa.

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 del proceso de tarificación *a priori no vida*, están entre ellos estrechamente vinculados, y su resolución no es independiente en función de los métodos estadísticos utilizados. En muchas ocasiones, la mejor tarifa no tiene por qué ser la que ofrezca una tabla cruzada de las variables tarificadoras más relevantes, sino la combinación sólo de algunas de

las clases de algunos de los factores que realmente presenten interacciones significativas en la explicación de la estructura de riesgo.

Es sabido que los métodos de selección dependen de parámetros que el investigador fija, que indican el grado de discriminación a partir del cual consideraremos que el factor de riesgo será influyente o no en la estructura de riesgo, por lo que la selección o no de una variable será un tanto subjetiva. No existe, por tanto, un método óptimo, por lo que en general todos ellos deben tenerse en cuenta con sus correspondientes ventajas e inconvenientes a la hora de tomar la decisión. Al menos la mayor parte de ellos deberían coincidir en la decisión tomada.

2. PROCESO DE SELECCIÓN

2.1. Nomenclatura

La nomenclatura que vamos a utilizar desde este momento, valedera tanto para la selección de predictores en un problema de cualquier índole, como en la aplicación actuarial de selección de variables de tarifa, es la siguiente:

- $t = [0, t)$: período de observación considerado (en general 1 año)
- n: número de individuos (o pólizas con el mismo riesgo en el caso de una cartera actuarial)
- Y: vector de realizaciones de la variable respuesta, vector $nx1$ continuo univariante (en el caso actuarial es la variable que recoge la siniestralidad objeto de estudio)
- $X^1, X^2, \ldots, X^P$: P vectores $nx1$ de realizaciones de las variables explicativas, dependientes o predictoras (en el caso actuarial serán los factores potenciales de riesgo)

- $X=(X^1, X^2, \ldots, X^P)$: matriz de predictores nxP de tipo mixto (mezcla de variables cualitativas y categóricas)
- $X^{(1)}, X^{(2)}, \ldots, X^{(K)}$: K variables seleccionadas procedentes de las variables explicativas sin alterar su estructura en los valores, con una nueva numeración que indicará el orden de preferencia en el proceso de selección, independientemente de la numeración inicial, $1 \le K \le P$ (en el caso actuarial sería el conjunto de variables de tarifa seleccionadas en un primer estadio de entre el conjunto de factores potenciales de riesgo anteriores)
- $X^{(\,)} = \left(X^{(1)}, X^{(2)}, \ldots, X^{(K)}\right)$: matriz nxK de tipo mixto, conjunto resultante de predictores seleccionados

2.2. Elementos previos

2.2.1. El modelo DB

El modelo utilizado es el de regresión basada en distancias ([5], [6],[7]). Está basado en análisis de distancias y, de forma genérica, consiste en proyectar la respuesta en el subespacio generado por componentes principales, X, las cuales se obtienen mediante *metric scaling*, y pasan a jugar el papel de variables predictoras en el modelo:

$$\hat{Y} = X\hat{B} = P_X Y$$

donde

$$\hat{B} = (X'X)^{-1} X'Y$$

$$P_X = X(X'X)^{-1} X'$$

Para su correcta aplicación sólo es requerido definir una adecuada función de distancias entre individuos con la propiedad euclídea en el sentido *del multidimensional scaling*.

No son requeridas hipótesis funcionales en el modelo, ni distribucionales en los datos. No hay límite establecido en una variable respuesta ([9]), aunque nosotros aquí realizaremos el estudio utilizando el modelo con una respuesta continua univariante ([4]). Por su diseño y características, soporta el uso de datos de tipo mixto, incluyendo variables continuas y categóricas, directamente, por lo que tanto los predictores como las respuestas, no necesitan transformaciones previas en la estructura de sus valores. Y el tratamiento de datos faltantes no ofrece dificultad.

Cabe notar, que si la función de distancias utilizada en el modelo es la distancia euclidea ordinaria, y los predictores son continuos, las predicciones obtenidas coinciden con las ofrecidas por el modelo clásico de regresión lineal múltiple por mínimos cuadrados; si la distancia es el *matching coefficient*, y los predictores son categóricos, las predicciones coinciden también con las de la regresión clásica siempre que los predictores utilizados en esta última sean variables *dummies* que representen las clases de cada factor y sus interacciones, en éste caso, las predicciones son las medias resultantes de la tabla cruzada de todos los factores.

2.2.2. Coeficiente de asociación

Definimos la variabilidad geométrica de una matriz de distancias, asociada con la matriz de productos escalares G, por $\Delta^{(2)} = g1'+1g'-2G$, como

$$V(\Delta) = \frac{1}{2n^2} 1'\Delta^{(2)}1 = \frac{1}{n} trG$$

En estos términos podemos generalizar el coeficiente de correlación múltiple como

$$R^2_{Y,X} = \frac{V(\hat{\Delta}_Y)}{V(\Delta_Y)}$$

donde el numerador proviene de la estimación $\hat{Y} = DBR(X)$, y el denominador directamente de Y.

2.2.3. Estadístico de contraste

Supongamos que queremos comparar dos modelos,

$$\text{M1: } \hat{Y} = DBR(X^1, X^2, \cdots, X^k)$$
$$\text{M2: } \hat{Y} = DBR(X^1, X^2, \cdots, X^{k+1})$$

es decir, queremos constatar que la variable, de forma genérica, X^{k+1}, contribuye de forma relevante en la explicación de la variable respuesta. Para ello, definimos el estadístico de contraste siguiente ([4]):

$$Q(X^{k+1}) = \frac{V^{k+1}(\hat{\Delta}_Y) - V^k(\hat{\Delta}_Y)}{V(\Delta_Y) - V^{k+1}(\hat{\Delta}_Y)}$$

Puesto que no conocemos su distribución, haremos uso de la metodología *bootstrap*, con tal de hallar el *p-valor* asociado a la muestra original. Con él seremos capaces de estudiar la significación de la variabilidad añadida. Los métodos de predicción basados en distancias son especialmente adecuados para el empleo de dicha metodología ([12]), pues el hecho que todas las interdistancias entre individuos de un remuestreo aparezcan ya en la matriz de distancias inicial, permite "simular" y vectorizar estos remuestreos por medio de matrices de multiplicidades, lo que es de gran economía computacional.

Cabe notar que, el coeficiente de asociación definido en el apartado anterior y el estadístico de contraste actual, sólo dependen de la función de distancias entre observaciones elegida.

Además, con predictores continuos y distancia euclidea ordinaria la variabilidad geométrica se reduce a la conocida varianza muestral; el coeficiente de asociación se reduce al coeficiente de determinación; y el estadístico de contraste al estadístico F que corrientemente se aplica para la selección de variables en regresión clásica, exceptuando la modificación por grados de libertad, no necesaria para el cálculo del

p-valor mediante *bootstrap*. Así, el proceso de selección que pasamos a detallar se reduce, bajo estas condiciones, al *stepwise* clásico.

2.3. Proceso

El mecanismo de introducción y de eliminación de predictores estudiado es análogo al del *stepwise* clásico: en cada etapa en una primera fase se selecciona, al menos temporalmente, el predictor de mayor asociación con la respuesta, y en una segunda fase, se deshecha, si es necesario, algún predictor del conjunto resultante. Se parte por lo tanto de que ninguna variable ha sido seleccionada inicialmente.

Puede realizarse la particularización hacia un proceso de introducción progresiva tan sólo suprimiendo en cada paso la fase de eliminación; y hacia un proceso de eliminación progresiva si suponemos que todos los predictores candidatos forman inicialmente parte del modelo y aplicamos sólo en cada etapa la fase de eliminación, en ambos casos, realizando los correspondientes contrastes de significación.

El proceso se inicia con la regresión de Y con cada uno de los predictores potenciales, y la selección para entrar, al menos temporalmente, el X^p que nos dé mayor R^2_{Y,X^p}, al que nomenclaremos $X^{(1)}$:

$$\hat{Y} = DBR(X^p),\ p = 1,2,\ldots,P \Rightarrow R^2_{Y,X^p},\ p = 1,2,\ldots,P;$$

$$X^{(1)} = \left\{ X^p \,/\, \underset{p\in\{1,2,\ldots,P\}}{Sup} R^2_{Y,X^p} \right\}$$

Comenzamos una numeración entre paréntesis según el orden de preferencia en la selección que no tiene nada que ver con la inicial que es a modo de etiqueta. De forma genérica, supongamos que, de un paso anterior, se ha seleccionado el conjunto de variables formado por $\{X^{(1)}, X^{(2)},\cdots, X^{(k)}\}$ provenientes del conjunto inicial de P variables. Las fases de selección y de eliminación se realizan como sigue:

- *Fase de introducción*: de entre las variables $\{X^1, X^2,\cdots, X^P\}\setminus\{X^{(1)}, X^{(2)},\ldots, X^{(k)}\}$ se elige como candidata a entrar $X^{(k+1)}$, al menos temporalmente, la que ofrezca mayor asociación con la respuesta teniendo en cuenta el efecto de las k variables seleccionadas en el paso anterior, esto es:

$$X^{(k+1)} = \left\{ X^p \,/\, \underset{p\in\{1,2,\ldots,P\}\setminus\{(1),(2),\ldots,(k)\}}{Sup} R^2_{Y,X^{(1)}X^{(2)}\ldots X^{(k)}X^p} \right\}$$

- *Fase de eliminación*: miramos si alguna de las variables del conjunto $\{X^{(1)}, X^{(2)},\cdots, X^{(k)}, X^{(k+1)}\}$ se ha vuelto no significativa, mediante el estadístico ya definido en los elementos previos:

$$Q(X^{(l)}) = \frac{V^{\{X^{(1)},X^{(2)},\ldots,X^{(k+1)}\}}(\hat{\Delta}_Y) - V^{\{X^{(1)},X^{(2)},\ldots,X^{(k+1)}\}\setminus X^{(l)}}(\hat{\Delta}_Y)}{V(\Delta_Y) - V^{\{X^{(1)},X^{(2)},\ldots,X^{(k+1)}\}}(\hat{\Delta}_Y)},$$

obtenemos los correspondientes valores de probabilidad mediante el método bootstrap:

$q^{(l)} = F\left[Q(X^{(l)})\right] \equiv q_{Y,X^{(l)},X^{(1)}\ldots X^{(l-1)}X^{(l+1)}\ldots X^{(k+1)}}$.

De todos los $q^{(l)}, (l) \in \{(1),(2),\ldots,(k+1)\}$ escogemos

$$q^{(m)} = Mín\{q^{(l)}\}_{l=1,2,\ldots,k+1}$$

$\rightarrow$ Si $q^{(m)} > 1-\alpha^*$, ninguna variable predictora es eliminada;

$\rightarrow$ Si $q^{(m)} \le 1-\alpha^*$, el predictor $X^{(m)}$ es eliminado:

- si $(m) \neq (k+1)$ vamos al paso siguiente con el conjunto $\{X^{(1)},\ldots, X^{(m-1)}, X^{(m+1)},\ldots, X^{(k+1)}\}$,
- si $(m) = (k+1)$, el proceso termina con el conjunto $\{X^{(1)}, X^{(2)},\ldots, X^{(k)}\}$.

El proceso parará cuando ocurra alguno de los siguientes casos: la variable introducida en un mismo paso sea también eliminada; en dos pasos se obtenga el mismo conjunto de predictores seleccionados; no queden más predictores significativos a entrar en el modelo; o ya tengamos el número deseado.

Destacamos que para la realización del proceso: no son requeridas hipótesis distribucionales en los datos, ni funcionales en el modelo; solamente se requiere definir una adecuada función de distancias entre observaciones, con la propiedad euclidea en el sentido del *multidimensional scaling*, de la cual dependerán el test y las medidas de asociación utilizadas durante el proceso; no son necesarias transformaciones en los predictores, pues el modelo de regresión basado en distancias admite directamente datos de tipo mixto; y el tratamiento de datos faltantes no ofrece dificultad.

3. APLICACIONES NUMÉRICAS

Para las 3 ilustraciones del método suponemos los siguientes parámetros pre-establecidos: nivel de significación permitido $\alpha^* = 0.05$, número máximo deseado de predictores $k^* = 2$. En las tres aplicaciones tenemos una variable respuesta continua univariante, para el detalle sobre el significado de las variables predictoras correspondientes ver [3].

3.1. Aplicación 1

Esta aplicación numérica ha sido realizada con n = 38 individuos, la variable respuesta Y es continua, y disponemos de 2 predictores ambos de naturaleza continua X^1, X^2, (los datos originales pueden encontrarse en [11]).

➢ Paso 1:

- *Fase de introducción*:

Variable	Coeficiente
X^1	$R^2_{Y,X^1} = 0.03385$
X^2	$R^2_{Y,X^2} = \mathbf{0.45739}$

la primera variable seleccionada, al menos temporalmente, es $X^{(1)} = X^2$

- *Fase de eliminación*: la probabilidad acumulada obtenida para testear si $X^{(1)}$ contribuye en la explicación de la variabilidad es $q_{Y,X^{(1)}} = 0.999997 > q^* = 0.95$, por lo que aceptamos la variable en el modelo

➢ Paso 2:

- *Fase de introducción*: $X^{(2)} = X^1$, al menos temporalmente
- *Fase de eliminación*:

Variable contrastada	Probabilidad acumulada
$X^{(2)}$	$q_{Y,X^{(2)}.X^{(1)}} = \mathbf{1}$
$X^{(1)}$	$q_{Y,X^{(1)}.X^{(2)}} = \mathbf{1}$

la probabilidad acumulada mínima es $q^{(m)} = q_{Y,X^{(2)}.X^{(1)}} = q_{Y,X^{(1)}.X^{(2)}} = 1 > q^* = 0.95$, por lo que ninguna variable es eliminada.

Puesto que no hay más variables a entrar en el modelo y ya tenemos el número de predictores deseado, el proceso finaliza con el conjunto $\{X^1, X^2\} = \{X^{(2)}, X^{(1)}\}$.

En este proceso hemos utilizado la distancia euclidea ordinaria como función de distancias entre individuos, obteniendo como caso particular el proceso *stepwise* clásico utilizado en regresión por mínimos cuadrados por ser ambos predictores de tipo continuo ([4]).

3.2. Aplicación 2

Esta aplicación numérica ha sido realizada con n = 32 individuos, la variable respuesta Y es continua, y disponemos de 5 predictores continuos, $C1, C2, C3, C4, C5$, 2 predictores binarios, $B1, B2$, y 3 predictores categóricos $Q1, Q2, Q3, X^1, X^2$, (los datos originales pueden encontrarse en [17]).

➢ Paso 1:

- *Fase de introducción*:

Variable	Coeficiente
C1	$R^2_{Y,C1} = 0.97883$
C2	$R^2_{Y,C2} = 0.97485$
C3	$R^2_{Y,C3} = 0.81982$
C4	$R^2_{Y,C4} = \mathbf{0.98916}$
C5	$R^2_{Y,C5} = 0.94784$
B1	$R^2_{Y,B1} = 0.69862$
B2	$R^2_{Y,B2} = 0.59477$
Q1	$R^2_{Y,Q1} = 0.73246$
Q2	$R^2_{Y,Q2} = 0.42915$
Q3	$R^2_{Y,Q3} = 0.44453$

la primera variable seleccionada, al menos temporalmente, es $X^{(1)} = C4$

- *Fase de eliminación*: la probabilidad acumulada obtenida para testear si $X^{(1)}$ contribuye en la explicación de la variabilidad es $q_{Y,X^{(1)}} = 1 > q^* = 0.95$, por lo que aceptamos la variable en el modelo

➢ Paso 2:

- *Fase de introducción*:

Variable	Coeficiente
C1	$R^2_{Y,C1 \cdot X^{(1)}} = 0.9894$
C2	$R^2_{Y,C2 \cdot X^{(1)}} = 0.9894$
C3	$R^2_{Y,C3 \cdot X^{(1)}} = 0.9894$
C5	$R^2_{Y,C5 \cdot X^{(1)}} = \mathbf{0.99027}$
B1	$R^2_{Y,B1 \cdot X^{(1)}} = 0.98918$
B2	$R^2_{Y,B2 \cdot X^{(1)}} = 0.98937$
Q1	$R^2_{Y,Q1 \cdot X^{(1)}} = 0.98918$
Q2	$R^2_{Y,Q2 \cdot X^{(1)}} = 0.9894$
Q3	$R^2_{Y,Q3 \cdot X^{(1)}} = 0.9894$

la segunda variable a entrar en el modelo, al menos temporalmente, es $X^{(2)} = C5$

- *Fase de eliminación*:

Variable contrastada	Probabilidad acumulada
$X^{(2)}$	$q_{Y,X^{(2)} \cdot X^{(1)}} = \mathbf{1}$
$X^{(1)}$	$q_{Y,X^{(1)} \cdot X^{(2)}} = \mathbf{1}$

la probabilidad acumulada mínima es $q^{(m)} = q_{Y,X^{(2)}.X^{(1)}} = q_{Y,X^{(1)}.X^{(2)}} = 1 > q^* = 0.95$, por lo que ninguna variable es eliminada.

El proceso finaliza con el conjunto $\{C4, C5\} = \{X^{(1)}, X^{(2)}\}$ puesto que habíamos establecido $k^* = 2$, el número máximo de predictores a entrar. En caso contrario, la siguiente variable a entrar en el modelo hubiera sido *Q1*.

En este proceso hemos utilizado el coeficiente de similaridad de *Gower* ([14]) como función de distancias entre individuos. Con esta aplicación pretendemos mostrar la viabilidad del método en el soporte de datos de tipo mixto. Se han comparado los resultados de la selección, con los de otros procesos existentes, corroborando que la mayoría de ellos elegían entre las primeras opciones las dos variables seleccionadas.

3.3. Aplicación 3

Esta aplicación numérica ha sido realizada con $n = 401$ individuos. La variable respuesta *Y*, cuantía de los impagos ocasionados a una entidad financiera por aquellos clientes que no pudieron hacer frente al pago de la deuda contraída en un determinado momento, es continua. Disponemos de 2 predictores, F^1 y F^2, de tipo categórico: F^1 es el estado civil con tres categorías, por lo tanto categórica nominal, y F^2 es la antigüedad en el puesto laboral (discretizada de antemano en también tres clases sin posibilidad de deshacer el cambio), por lo tanto categórica ordinal. Podemos encontrar una explicación más detallada de los datos y un estudio de la significación de las variables mediante coeficientes de credibilidad en [1].

➢ Paso 1:

- *Fase de introducción*:

Variable	Coeficiente
F^1	$R^2_{Y,F^1} = 0.033$
F^2	$R^2_{Y,F^2} = \mathbf{0.457}$

la primera variable seleccionada, al menos temporalmente, es $X^{(1)} = F^2$, en este caso la antigüedad en el puesto laboral

- *Fase de eliminación*: la probabilidad acumulada obtenida para testear si $X^{(1)}$ contribuye en la explicación de la variabilidad es $q_{Y,F^{(1)}} = 0.965083 > q^* = 0.95$, por lo que aceptamos la variable en el modelo

➢ Paso 2:

- *Fase de introducción*: $X^{(2)} = X^1$, al menos temporalmente
- *Fase de eliminación*:

Variable contrastada	Probabilidad acumulada
$X^{(2)}$	$q_{Y,X^{(2)}.X^{(1)}} = \mathbf{0.98126}$
$X^{(1)}$	$q_{Y,X^{(1)}.X^{(2)}} = 0.99997$

la probabilidad acumulada mínima es $q^{(m)} = q_{Y,X^{(2)}.X^{(1)}} = 0.98126 > q^* = 0.95$, por lo que la variable no es eliminada.

Puesto que no hay más variables a entrar en el modelo y ya tenemos el número de predictores deseado, el proceso finaliza con el conjunto $\{F^1, F^2\} = \{X^{(2)}, X^{(1)}\}$.

En este proceso hemos utilizado el *matching coefficient* como función de distancias entre individuos, obteniendo así un procedimiento similar al propuesto por *Hallin* [16], excepto por el criterio de selección en cada paso ([3]).

4. SUMARIO

Se ha presentado una técnica de análisis multivariante consistente en un método de selección de variables predictoras paso a paso, aplicable cuando el modelo es el de regresión basada en distancias ([7],[8],[9]) y, disponemos de una variable respuesta continua univariante y de un conjunto de predictores potenciales de tipo mixto, incluyendo variables continuas y categóricas ([4]).

Ésta se propone como herramienta alternativa para cubrir la fase de selección de variables de tarifa, de entre el conjunto de factores potenciales de riesgo, que se realiza en el proceso de tarificación *a priori* de los seguros *no vida* ([3]). En tal caso, la variable respuesta se corresponde con la experiencia de siniestralidad (número, cuantía o cuantía total de los siniestros), en general del período de 1 año, y los predictores a seleccionar con los posibles factores potenciales del riesgo a explicar. Notemos que, en esta fase previa de selección, no asumimos ninguna hipótesis respecto a la estructura de la tarifa.

Para la ejecución del proceso no son requeridas hipótesis distribucionales en los datos, ni funcionales en el modelo; sólo se requiere definir una adecuada función de distancias entre observaciones con la propiedad euclidea en el sentido del *multidimensional scaling*, de la cual dependerán el test y las medidas de asociación utilizadas; no son necesarias transformaciones previas en los predictores, pues el modelo de regresión basado en distancias admite directamente datos de tipo mixto; y el tratamiento de datos faltantes no ofrece dificultad. Los programas informáticos están realizados por los autores y no tienen límite establecido ni en número de variables ni de individuos, solamente dependen de la capacidad computacional del ordenador utilizado.

Cabe notar que, en el caso de distancia euclidea ordinaria y predictores continuos, el método se reduce al *stepwise* clásico utilizado con regresión lineal múltiple por mínimos cuadrados ([4]).

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. Encontramos como herramientas alternativas procesos de selección basados en modelos de regresión ([5],[15],[19A,19B]), análisis discriminante ([2]), análisis cluster ([20]), etc.

Con este trabajo, pretendemos justificar la utilización de técnicas de análisis estadístico multivariante como paso necesario en el proceso de tarificación *a priori no vida*, de manera que las entidades aseguradoras entiendan que no deben escoger factores de riesgo sólo de las estadísticas comunes y de las pautas comerciales y de mercado, sino de su propia experiencia, pues esto les ha de conducir a una mejor gestión de los riesgos de su cartera.

BIBLIOGRAFÍA

[1] 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.

[2] BEUTHE M. & PH. VAN NAMEN (1975): "*La seléction des assurés et la détermination des primes d'assurances par l'analyse discriminante*". Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker, vol. 75, no 2, pp. 137-156.

[3] BOJ E. & M. M. CLARAMUNT (1999): "*Selection of predictors in rate making*". III International Congress on Insurance: Mathematics & Economics. London 19, 20 and 21 July 1999.

[4] BOJ E., CLARAMUNT M. M. Y J. FORTIANA (2000): "*Selección de predictores en el modelo basado en distancias*". XXV Congreso Nacional de Estadística e Investigación Operativa. Vigo del 4 al 7 de abril del 2000.
[5] BOJ E., CLARAMUNT M. M., FORTIANA J. & A. VIDIELLA (2000): "*A comparison of distance-based regression and generalized linear models in the rate making process. An empirical study*". IV International Congress on Insurance: Mathematics & Economics. Barcelona 24, 25 and 26 July 2000.
[6] BROCKMAN M. J. & T. S. WRIGHT (1992): "*Statistical Motor Rating: Making Effective Use of your Data*". Journal of the Institute of Actuaries 119, III, pp. 457-543.
[7] CUADRAS C. M. (1989): "*Distance Analysis in discrimination and classification using both continuous and categorical variables*". En: Statistical Data Analisys and Inference (Y. Dodge ed.). Elsiever Science Publisher. North-Holland. Amsterdam. pp. 459-474.
[8] CUADRAS C. M. AND C. ARENAS (1990): "*A distance based model for prediction with mixed data*". Communications in Statistics, Theory Meth., 19. pp. 2261-2279.
[9] CUADRAS C. M. AND J. FORTIANA (1998): "*Generalized distance-based regression*". Classification & Psychometric Soc. Joint Meeting, June 1998.
[10] 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, pp. 7-17.
[11] DRAPER N. R. & H. SMITH (1981): "*Applied Regression Analysis*" (second edition). John Wiley & Sons, New York, pp. 519-520.
[12] FORTIANA J. Y C. M. CUADRAS (1994): "*Estimación bootstrap de la distancia entre poblaciones*". XXI Congreso Nacional de Estadística e Investigación Operativa. Calella 18-21 de abril de 1994.
[13] GARRIDO J. J. (1987): "*Teoría general y derecho español de los seguros privados*". En: Tratado general de los seguros. Tomo I, vol. II.
[14] GOWER J. C. (1971): "*A general coefficient of similarity and some of its properties*". Biometrics, 27, pp. 857-874.
[15] HABERMAN S. & RENSHAW, A. E. (1998): "*Actuarial applications of Generalized Linear Models*". En: Statistics in Finance,

Hand D. J. & S. D. Jacka, Arnold applications of statistics, London· Sydney· Auckland, 1998, pp. 42-65.
[16] HALLIN P. M. (1977): "*Méthodes statistiques de construction de tarif*". Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker, vol. 77, no 2, pp. 160-175.
[17] HENDERSON H. V. & P. F. VELLEMAN (1981): "*Building multiple regression models interactively*". Biometrics 37, pp. 391-411.
[18] LEMAIRE J. (1995): "*Bonus-malus system in automobile insurance*". Kluwer-Nijhof Publishing, Boston, MA.
[19A] LEMAIRE J. (1977): "*Selection procedures of regression analysis applied to automobile insurance*". Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker, vol. 77, no 2, pp. 143-160.
[19B] 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, vol. 79, no 1, pp. 65-72.
[20] LOIMARANTA K., JACOBSSON J. & H. LONKA (1980): "*On the Use of Mixture Models in Clustering Multivariate Frequency Data*". Transactions of the 21 st International Congress of Actuaries, 2, T147-161.
[21] NIETO U. Y J. VEGAS (1993): "*Matemática Actuarial*". Ed. Mapfre. Madrid.
[22] PITKÄNEN P. (1975): "*Tariff Theory*". Astin Bulletin 8:2, (1975), pp. 204-228.
[23] VAN EEGHEN J., E. K. GREUP AND J. A. NIJSSEN (1983): "*Rate Making*". Survey of Actuarial Science 2, Nationale-Nederlanden N. V., Rotterdam.