Segmentación y predicción en los modelos de tarificación

Caro Carretero, Raquel. rcaro@doi.icai.upcomillas.es
Departamento de Organización Industrial
Universidad Pontificia Comillas. ICAI

RESUMEN

El análisis multivariante ha experimentado una notable difusión en todas las ramas de la Ciencia en los últimos años. El avance de la informática ha permitido el tratamiento rápido de un número importante de variables y datos. Además, la aparición en el mercado de paquetes estadísticos, en PC, facilitan el trabajo y análisis de los resultados a usuarios no expertos en informática, sin necesidad de recurrir a lenguajes de programación complejos. De esta manera, *la estadística multivariante* pretende acercarse más a la realidad mediante el tratamiento conjunto de todas las variables observadas en el estudio de un fenómeno. En la Ciencia Actuarial, la estadística constituye la herramienta metodológica que permite al investigador analizar lo que aparentemente no tiene medida: la variabilidad en la prima. El análisis multivariante engloba los métodos y técnicas estadísticas que permiten estudiar y tratar, en bloque, un conjunto de variables medidas u observadas en una colección de individuos. Es importante tener en cuenta el hecho teórico de que todos los métodos multivariantes no son independientes entre sí, y que existen fuertes relaciones matemáticas entre ellos.

***Palabras claves*:**

La estadística multivariante; Ciencia Actuarial, SPSS; modelos lineales generalizados.

1. INTRODUCCIÓN

Ante la entrada de nuevas compañías seguida de un crecimiento en la captación de cuota de mercado, la situación del seguro de automóviles en España es el de una posición reactiva frente al líder; es decir, un intento de bajada de tarifas que da lugar a insuficiencia en las primas y como consecuencia, una pérdida de valor de la compañía. Se hace necesario, por tanto, una técnica de tarificación adecuada para *evitar procesos de antiselección* que producen una inestabilidad al equilibrio de la cartera y para *identificar segmentos rentables* en la cartera (nichos de mercado).En este sentido, el objetivo del este trabajo es presentar un resumen de los *métodos multivariantes* que pretenden acercarse más a esta realidad mediante el tratamiento conjunto de todas las variables observadas en el estudio de un fenómeno, centrándonos en aquellos que permiten **clasificar** (CHAID) y **predecir** (modelos lineales generalizados).

Quien ha llegado a entender y a aplicar los métodos estadísticos convencionales es consciente de que éstos analizan exhaustivamente el comportamiento de *una variable* en las muestras y en la población, y *de las leyes que pueden ligar dos variables entre sí*. La estadística univariante y la estadística bivariante clásicas, imprescindibles desde el punto de vista del análisis de las relaciones entre cada una de las variables explicativas y la variable que se desea explicar, no tienen en cuenta las correlaciones entre las variables, ni las interacciones entre las mismas, ni las relaciones no lineales. Pero en la vida real, una variable no suele depender sólo de otra y las relaciones dos a dos que puedan aparecer son insuficientes.

De esta manera, el análisis multivariante engloba los métodos y técnicas estadísticas que permiten estudiar y tratar, en bloque, un conjunto de variables medidas u observadas en una colección de individuos.

Es importante tener en cuenta el hecho teórico de que todos los métodos multivariantes no son independientes entre sí, y que existen fuertes relaciones matemáticas entre ellos. Por esta razón, muchos de los conceptos utilizados en la descripción de una prueba son posteriormente utilizados en otra.

Existe una dificultad añadida, común a todos los métodos multivariantes, la interpretabilidad de unos resultados que no siempre son únicos; que el matemáticamente mejor quizá no debe ser elegido y en muchas ocasiones obliga a sacrificar la optimización estadística en aras de una interpretación más útil o menos costosa.

Así, al enfrentarse a la realidad de un estudio, el investigador dispone habitualmente de muchas variables medidas u observadas en una colección de individuos; pretende estudiarlas conjuntamente, y acude al análisis multivariante. Se encuentra frente a una diversidad de técnicas y debe seleccionar la más adecuada a sus datos pero, sobre todo, a su objetivo científico.

Al observar muchas variables sobre una muestra se presume que una parte de la información recogida puede ser redundante. Así, existen métodos multivariantes de **reducción** de la dimensión que tratan de eliminarla. Por otro lado, como los individuos pueden presentar ciertas características comunes en sus respuestas, aparecen métodos multivariables que permiten su **clasificación** en grupos de cierta homogeneidad. Finalmente, podrá existir una variable cuya dependencia de un conjunto de otras sea interesante detectar para analizar su **relación** o, incluso, aventurar su **predicción** cuando las demás sean conocidas.

En la actividad aseguradora el riesgo es un factor fundamental; así pues su explicación y medición constituye una tarea de importancia capital para el actuario. El objetivo estará en obtener los niveles de rentabilidad y eficacia planteados, máxime, cuando la actividad aseguradora intenta determinar en el momento presente las primas a aplicar en el futuro a partir de la información del pasado. De ahí la importancia de un buen modelo predictivo que garantice cierta estabilidad en el tiempo.

De esta manera, como el riesgo se presenta como fenómeno puramente estocástico, es decir, como variable aleatoria, su estudio ha de basarse en técnicas que exploten al máximo la información que sobre el comportamiento de dicha variable se tenga. Con esta tarea se podrán establecer criterios que sirvan de base en la labor actuarial a la hora de fijar políticas de tarificación, cálculos de primas, recargos de seguridad, etc., así como la política de selección de riesgos en la cartera.

2. TÉCNICAS DE TARIFICACIÓN EN EL SEGURO DEL AUTOMÓVIL. SEGMENTACIÓN

Nos enfrentamos a un mercado muy saturado, donde el volumen de primas en el seguro de automóviles se va incrementado y representa casi el 50% respecto de las primas no-vida. Ante la entrada de compañías y canales de distribución nuevos (banca, teléfonos, etc.) y el crecimiento en la captación de cuota de mercado de los competidores, la situación del mercado del seguro de automóviles en España es la de una posición reactiva frente al líder, es decir, bajada de tarifas que da lugar a insuficiencia en las primas, así como insuficiencia en las reservas y como consecuencia de esto, pérdida de valor de la compañía. Se hace necesario, por tanto, una técnica de tarificación adecuada para evitar procesos de antiselección que conllevan una inestabilidad al equilibrio de la cartera y para identificar segmentos rentables en la cartera (nichos de mercado).

Desde hace muchos años, la tarifa en el seguro de automóviles se viene "segmentando", esto es, en el sentido de dividiendo, clasificando. De esta forma, la estructura clásica tarifaria en nuestro país, se divide, primero por tipo de vehículo, luego dentro de cada tipo, por potencia, por tonelaje, por cilindrada o por número de plazas, y posteriormente por grupos, sin olvidar los factores como el uso, la edad del conductor o la antigüedad del carné de conducir. Sin embargo en este caso nos referimos a *segmentación* cuando lo hacemos en el sentido de método de análisis multivariante.

Se debe utilizar un análisis multivariante porque la tarificación clásica no tiene en cuenta las interrelaciones que se producen entre las variables. Para entender este aspecto es muy significativa la relación que existe entre la edad y antigüedad del carnet de conducir. Antes los recargos por edad y antigüedad se establecían independientemente, de tal forma que a un conductor de 20 años se le establecía un recargo por ser joven y además otro por ser inexperto. Con la segmentación multivariante, al tener en cuenta las interrelaciones, cada colectivo soporta la prima que le corresponde.

2.1. Fundamentos básicos de la segmentación utilizando el módulo CHAID de SPSS.

Ante la lógica imposibilidad de abordar técnicas laboriosas sin la ayuda del proceso automático de datos, se puede elegir para la realización de aplicaciones prácticas, uno de los programas estadísticos más reconocidos actualmente, el SPSS. Este paquete estadístico nos permite la interpretación de las salidas de ordenador ante la ingente información que se le proporciona.

Las técnicas AID (Automatic Interaction Detection) son una familia de técnicas de análisis de conglomerados (cluster analysis) para variables categóricas. Su versión informática más popular es el módulo de segmentación CHAID (Chi-square Automatic Interaction Detection) del paquete estadístico SPSS.

Las técnicas AID fueron desarrolladas inicialmente por Sonquist y Morgan en 1964. Su objetivo era dividir secuencialmente la muestra en subgrupos mutuamente exclusivos mediante una serie de divisiones binarias. Cada división era determinada a partir de la selección del mejor predictor posible, y la combinación de categorías del predictor que maximizará la reducción de la parte de la varianza no explicada de la variable criterio. Los programas actuales permiten divisiones que no necesariamente han de ser binarias. El resultado de un análisis AID es una serie de subgrupos, representados habitualmente mediante un diagrama de árbol, los cuales maximizan la diferencia entre las categorías de la variable criterio.

CHAID tiene la ventaja frente a otros métodos de análisis multivariante que es muy intuitivo al presentarse en forma de árbol. **CHAID** es una técnica de segmentación que divide la población en grupos o segmentos homogéneos respecto a la variable que deseamos explicar, a la vez que los grupos resultantes difieren significativamente entre sí. **CHAID** basa sus decisiones en la significación del análisis del estadístico Chi-cuadrado.

En nuestro caso, los segmentos obtenidos en la segmentación, son creados a partir de la introducción simultánea de las variables que mejor explican desde el punto de vista estadístico *la siniestralidad*: edad, potencia del vehículo, zona, sexo, o la antigüedad del permiso de conducir. La importancia de cada variable en el árbol dependerá de su nivel de significación en cada nodo.

2.2. ¿Cuando utilizar CHAID?

Las situaciones en las que **CHAID** puede ser más ventajoso son aquellas en las que se dispone de una variable criterio categórica y variables predictoras también categóricas. Se ejecuta habitualmente sobre ficheros agregados. Si las variables predictoras son continuas hay que categorizarlas o dividirlas en intervalos, aunque si además la variable predictora es dicotómica puede ser más adecuado utilizar la regresión logística. Para utilizar las técnicas AID es necesario disponer de tamaños de muestra grandes, ya que la muestra se subdividirá en múltiples subgrupos, y cabe el riesgo de encontrar grupos vacíos o poco representativos si no se dispone de suficientes sujetos en cada combinación de categorías. El problema es similar al que aparece cuando se efectúan tablas de clasificación con criterios múltiples.

De esta manera, **CHAID** puede quedar definido como "una técnica de clasificación para datos categóricos, basada en una estructura de árbol" y en términos generales puede considerarse como un procedimiento para dividir la población, en dos o más grupos diferenciados. Basándose en las categorías del "mejor" predictor de una variable dependiente. El procedimiento continua con la división de cada uno de estos grupos iniciales en subgrupos menores, basándose en otras variables predictoras. Este proceso continúa hasta que no es posible encontrar otras variables predictoras estadísticamente significativas (o hasta que se cumpla un criterio especificado previamente).

2.3. Utilidades de la segmentación

1.- Se puede utilizar para buscar algún colectivo en el que nos interese diferenciar primas, o buscar nichos rentables.

2.- Como método de tarificación, la prima de la cartera debe ser suficiente para el conjunto, no olvidando recargar cuando sea necesario.

3.- Como criterio de selección de nuestra cartera

2.4. Problemas fundamentales en la segmentación

1.- Dificultad en el tratamiento informático dentro de las entidades. Se necesitan recursos económicos, técnicos y humanos. Es imprescindible, una base de datos con información válida y suficiente para empezar cualquier tipo de análisis.

2.- Estabilidad de la cartera; la segmentación por si sola no garantiza necesariamente una cartera rentable. El equilibrio técnico de la entidad se consigue a través de la correcta fijación de las primas en cada segmento y, de la consecución de una prima suficiente para toda la cartera, en el sentido de que al bonificar determinados segmentos no olvidemos que hay que recargos otros. La segmentación, no debe de ser entendida exclusivamente como la búsqueda del segmento rentable, sino como método de cálculo de primas en el sentido actuarial.

3.- Lleva los resultados a la tarifa actual

3. LOS MODELOS LINEALES GENERALIZADOS

Los modelos lineales generalizados son una clase de modelos estadísticos que generaliza a los modelos lineales clásicos (regresión lineal, modelos de análisis de la varianza), incluyendo otros modelos útiles en el análisis estadístico. Tales modelos pueden ser los modelos log-lineal para el análisis de datos en forma de cuantías, modelos probit y logit para datos en forma de proporciones y modelos para datos continuos con errores estándar proporcionales constantes.

Un aspecto importante de la generalización es la presencia en todos los modelos de un predictor lineal basado en una combinación lineal de variables explicativas o de "estímulo". Las variables pueden ser continuas o categóricas (o incluso una mezcla de ambas). La existencia de un predictor lineal significa que los conceptos de la regresión clásica y los modelos de análisis de la varianza, en tanto que se refieren a la estimación de parámetros en un predictor lineal, nos lleva directamente a una clase amplia de modelo.

Los modelos lineales generalizados comparten propiedades, como es la linealidad, además tienen un algoritmo común para la estimación de parámetros por máxima

verosimilitud. Utilizan mínimos cuadrados ponderados con una variable dependiente ajustada y no requieren hipótesis preliminares de los valores de los parámetros. Estas propiedades comunes permiten estudiar a los modelos lineales generalizados como una única clase.

En este sentido, cabe preguntarse si es posible realizar una tarificación adecuada a través de un modelo lineal generalizado que tenga en cuenta las variables frecuencia de siniestralidad y coste de los siniestros. Para su aplicación práctica se puede utilizar igualmente el paquete estadístico SPSS.

4. CONCLUSIONES

Las estadísticas univariante y bivariante clásicas, imprescindibles desde el punto de vista del análisis de las relaciones entre cada una de las variables explicativas y la variable que se desea explicar, no tienen en cuenta las correlaciones entre las variables, ni las interacciones entre las mismas, ni las relaciones no lineales. No obstante, en la vida real, una variable no suele depender sólo de otra.

De esta manera, se debe utilizar un análisis multivariante que permite estudiar y tratar, en bloque, un conjunto de variables observadas en una colección de individuos. La tarificación clásica no tiene en cuenta las interrelaciones que se producen entre las variables. En este sentido, la prima resultante al utilizar métodos multivariantes es distinta a la obtenida en la tarificación clásica.

Cabe destacar una serie de diferencias entre las técnicas de segmentación y los modelos lineales:

1.- Son técnicas complementarias, no sustitutivas

2.- Las técnicas de segmentación permiten:

*detectar factores importantes

*determinar el número adecuado de niveles en cada factor

*detectar interacciones

Sin embargo para predecir la prima final, es más idóneo el empleo de modelos lineales porque:

* ayudan a detectar distribuciones cruzadas: existe una desigual distribución en la cartera respecto otros factores

* analizan el efecto aislado de cada factor

* dan mayor credibilidad de los resultados

* detectan interacciones

5. REFERENCIAS BIBLIOGRÁFICAS

[1] CARRASCO, J.L. y HERNÁN, M.A. (1993). Estadística Multivariante en las Ciencias de la vida. Fundamentos, métodos y aplicación. Editorial Ciencia 3, S.L. Madrid.

[2] COUTTS, S. (1984). Motor Premium Rating. Insurance, Mathematics and Economics 3, pp.73-96

[3] McCULLAGH, P. y NELDER, J.A. (1989). Generalized Linear Models. Second Edition Chapman and Hall. London.

[4] MANLY, Bryan F.J. (1990). Multivariate Statistical Methods. A primer. 4th ed. Chapman and Hall. Bristol.

[5] RENCHER, A.C. y SCHAALJE, G. B. (2008). Linear Models in Statistics, 2nd Edition. Wiley-Interscience.

[6] SPSS Inc (2008). SPSS® for Windows. Chicago.