ReaR
ISNN 1989 4090
Revista electrónica de AnestesiaR
Marzo 2021

ORIGINAL

¿Exotérico o esotérico? Análisis multivariante.

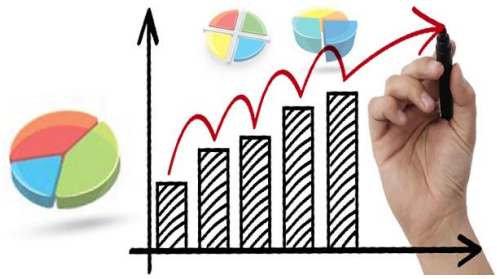
Molina Arias M.

Hospital Infantil Universitario La Paz, Madrid.

Resumen

El análisis multivariante es el conjunto de procedimientos estadísticos que estudian de forma simultánea varias características de un mismo sujeto o entidad, con el fin de analizar la interrelación que pueda existir entre todas las variables aleatorias que representan dichas características. Se describen los métodos principales agrupados en tres categorías: técnicas de rango completo y no completo, técnicas para reducir la dimensionalidad y métodos de clasificación y discriminación.

Introducción



El análisis multivariante es el conjunto de procedimientos estadísticos que estudian de forma simultánea varias características de un mismo sujeto o entidad, con el fin de analizar la interrelación que pueda existir entre todas las variables aleatorias que representan dichas características. Se describen los métodos principales agrupados en tres categorías: técnicas de rango completo y no completo, técnicas para reducir la dimensionalidad y métodos de clasificación y discriminación.

Hay días que vengo bíblico. Otros días vengo mitológico. Hoy vengo filosófico y hasta un poco masónico.

Y es que el otro día me dio por preguntarme cuál es la diferencia entre exotérico y esotérico, así que lo consulté con ese amigo de todos que tanto sabe de todo, nuestro amigo Google. Amablemente me explicó que ambos términos son parecidos y suelen explicar dos aspectos de una misma doctrina. El exoterismo hace referencia al saber que no se ve limitado a un determinado grupo de la comunidad que trata esos conocimientos, conocimientos que pueden ser divulgados y puestos al alcance de cualquiera. Por otra parte, el esoterismo hace referencia a los conocimientos que pertenecen a un orden más profundo y elevado, solo al alcance de unos pocos privilegiados especialmente educados para comprenderlos.

Y ahora, una vez comprendida la diferencia, os hago una pregunta un poco capciosa: la estadística multivariante, ¿es exotérica o esotérica? La respuesta, como es lógico, dependerá de cada uno, pero vamos a ver si es verdad que ambos conceptos no son contradictorios, sino complementarios, y podemos quedarnos en un justo término medio, al menos en la

comprensión de la utilidad de las técnicas multivariantes.

Estamos más habituados a utilizar técnicas de estadística univariante o bivariante, que permiten estudiar de forma conjunta un máximo de dos características de los individuos de una población para detectar relaciones entre ellas.

Sin embargo, con el desarrollo matemático y, sobre todo, de la capacidad de cálculo de nuestros ordenadores, cobran cada vez más importancia las técnicas de estadística multivariante o multivariada.

Podemos definir el análisis multivariado como el conjunto de procedimientos estadísticos que estudian de forma simultánea varias características de un mismo sujeto o entidad, con el fin de analizar la interrelación que pueda existir entre todas las variables aleatorias que representan dichas características. Permitidme que insista en los dos aspectos de estas técnicas: la multiplicidad de variables y el estudio de sus posibles interrelaciones.

Hay multitud de técnicas de análisis multivariante, abarcando desde los métodos puramente descriptivos hasta los que utilizan técnicas de inferencia estadística para obtener conclusiones de los datos y poder elaborar modelos que no son evidentes a simple vista observando los datos obtenidos. También nos permitirán desarrollar modelos de predicción de varias variables y establecer relaciones entre las mismas.

Algunas de estas técnicas son la extensión de sus equivalentes con dos variables, una dependiente y otra independiente o explicativa. Otras, no tienen nada equivalente parecido en la estadística de dos dimensiones.

Algunos autores clasifican estas técnicas en tres grandes grupos: los modelos de rango completo y no completo, las técnicas para reducir la dimensionalidad y los métodos de clasificación y discriminación. No os preocupéis si esto parece un galimatías, vamos a tratar de simplificarlo un poco.

Para poder hablar de las **TÉCNICAS DE RANGO COMPLETO Y NO COMPLETO**, creo que habrá que explicar primero a qué rango nos estamos refiriendo.

Aunque no vamos a entrar en ello ni de lejos, todos estos métodos encierran en su interior técnicas de cálculo matricial. Ya sabéis, las matrices, un conjunto de números en dos dimensiones (las que vamos a tratar aquí) que forman filas y columnas y que pueden sumarse y multiplicarse entre sí, además de otras operaciones.

Se define el rango de una matriz como el número de filas o columnas que son linealmente independientes (da igual filas o columnas, el número es el mismo). El rango puede valer desde 0 hasta el mínimo número de filas o de columnas. Por ejemplo, una matriz de 2 filas y 3 columnas podrá tener un rango de 0 a 2. Una matriz de 5 filas y 3 columnas podrá tener un rango de 0 a 3. Ahora imaginad una matriz de dos filas, la primera 1 2 3 y la segunda 3 6 9 (tiene 3 columnas). Su rango máximo sería 2 (el número menor de filas y de columnas) pero, si os fijáis, la segunda fila es la primera multiplicada por 3, así que solo hay una linealmente independiente, por lo que su rango es igual a 1.

Pues bien, se dice que una matriz es de rango completo cuando su rango es igual al más grande posible para una matriz de sus mismas dimensiones. El tercer ejemplo que os he puesto sería una matriz de rango no completo, ya

que una matriz de 2×3 tendría un rango máximo de 2 y el de nuestra matriz es de 1.

Una vez entendido esto, vamos con los métodos de rango completo y no completo.

El primero que veremos es el **modelo de regresión lineal múltiple**. Este modelo, extensión del de regresión lineal simple, se utiliza cuando tenemos una variable dependiente y una serie de variables explicativas, todas ellas cuantitativas, y se cumple que se pueden relacionar de forma lineal y que las explicativas conforman una matriz de rango completo.

De forma similar a la regresión simple, esta técnica nos permite predecir los cambios de la variable dependiente en función de las variables explicativas. La fórmula es similar a la de la regresión simple, pero incluyendo todas la variables independientes explicativas, así que no os voy a aburrir con ella. No obstante, dado que os he castigado con los rangos y las matrices, dejadme que os diga que, en términos matriciales, se puede expresar de la siguiente manera:

$$Y = X\beta + e_i$$

donde X es la matriz de rango completo de las variables explicativas. La ecuación incluye un término de error que se justifica por la posible omisión en el modelo de variables explicativas relevantes o de errores de medida.

Para complicar las cosas, imaginad que tratásemos de correlacionar simultáneamente varias variables independientes con varias dependientes. En este caso no nos sirve la regresión múltiple y tendríamos que recurrir a la técnica de **correlación canónica**, que permite realizar predicciones de varias variables dependientes en función del valor de varias explicativas.

Si recordáis de la estadística bivariante, el análisis de la varianza (ANOVA) es la técnica que nos permite estudiar el efecto sobre una variable dependiente cuantitativa de las variables explicativas cuando estas son categorías de una variable cualitativa (a estas categorías las llamamos factores). En este caso, como cada observación puede pertenecer a uno y solo uno de los factores de la variable explicativa, la matriz X será de rango no completo.

Una situación un poco más complicada se produce cuando las explicativas son una variable cuantitativa y uno o más factores de una cualitativa. En estas ocasiones recurrimos a un modelo lineal generalizado denominado análisis de la covarianza (ANCOVA).

Trasladando lo que acabamos de decir al reino de la estadística multivariante, tendríamos que utilizar la extensión de estas técnicas. La extensión del ANOVA cuando hay más de una variable dependiente que no se puede combinar en una sola es el **análisis multivariante de la varianza** (MANOVA). Si coexisten factores de variables cualitativas con variables cuantitativas, recurriremos al **análisis multivariante de la covarianza** (MANCOVA).

El segundo grupo de técnicas multivariantes son las que tratan de **REDUCIR LA DIMENSIONALIDAD.**

En algunas ocasiones tenemos que manejar un número de variables tan elevado que resulta complejo organizarlas y llegar a alguna conclusión útil. Ahora bien, si tenemos la suerte de que las variables estén correlacionadas entre sí, la información que aporte el conjunto será redundante, ya que la que den unas variables incluirá la que ya aportan otras variables del conjunto.

En estos casos resulta útil reducir la dimensión del problema disminuyendo el número de variables a un conjunto más pequeño de variables no correlacionadas entre sí y que recojan la mayor parte de la información incluida en el conjunto original. Y decimos la mayor parte porque, como es obvio, cuanto más reduzcamos el número, más información perderemos.

Las dos técnicas fundamentales que utilizaremos en estos casos son el análisis de componentes principales y el análisis factorial.

El **análisis de componentes principales** toma un conjunto de p variables correlacionadas y las transforma en uno nuevo de variables no correlacionadas, al que denominamos componentes principales. Estas componentes principales nos permiten explicar las variables en términos de sus dimensiones comunes.

Sin entrar en detalle, se elabora una matriz de correlaciones y una serie de vectores que nos proporcionarán las nuevas componentes principales, ordenadas de mayor a menor según la varianza de los datos originales que explique cada componente. Cada componente será una combinación lineal de las variables originales, algo similar a una recta de regresión.

Imaginemos un caso muy sencillo con seis variables explicativas (X1 a X6). La componente principal 1 (CP1) puede ser igual, por decir algo, a 0,15X1 + 0,5X2 – 0,6X3 + 0,25X4 – 0,1X5 – 0,2X6 y, además, explicar el 47% de la varianza. Si la CP2 resulta que explica el 30% de la varianza, con CP1 y CP2 tendremos controlado el 77% con un conjunto de datos más fácil de manejar (pensemos si en lugar de 6 variables tenemos 50). Y no solo eso, si representamos gráficamente CP1 frente a CP2, podemos ver si se produce algún tipo de agrupamiento de la variable en estudio según los valores de las componentes principales.

De esta manera, si tenemos suerte y unas pocas componentes recogen la mayor parte de la varianza de las variables originales, habremos reducido la dimensión del problema. Y aunque, en ocasiones, esto no es posible, siempre nos puede servir para encontrar agrupaciones en los datos definidos por un gran número de variables, lo cual nos enlaza con la siguiente técnica, el **análisis factorial**.

Sabemos que la varianza total de nuestros datos (la que estudia el análisis de componentes principales) es la suma de tres componentes: la varianza común o compartida, la varianza específica de cada variable y la varianza debida al azar y los errores de medición. Una vez más, y sin entrar en detalles, el método del análisis factorial parte de la matriz de correlaciones para aislar únicamente la varianza común y tratar de encontrar una serie de dimensiones subyacentes comunes, llamadas factores, que no son observables viendo el conjunto original de variables.

Como vemos, estos dos métodos son muy parecidos, por lo que existe mucha confusión sobre cuándo se debe utilizar uno y cuándo otro, máxime teniendo en cuenta que el análisis de componentes principales puede ser el primer paso en la metodología del análisis factorial.

Ya lo hemos dicho, el análisis de componentes principales trata de explicar la máxima proporción posible de la varianza total de los datos originales, mientras que el objetivo del estudio del análisis factorial es explicar la covarianza o correlación que existe entre sus variables. Por tanto, habitualmente se utilizará el análisis de componentes principales para buscar combinaciones lineales de las variables

originales y reducir un conjunto de datos extenso a otro más reducido y manejable, mientras que recurriremos al análisis factorial cuando busquemos un nuevo conjunto de variables, generalmente más reducido que el original, y que represente lo que tienen en común las variables originales.

Avanzando en nuestro arduo camino de hoy, para aquellos esforzados que todavía sigáis leyendo, vamos a tratar los **MÉTODOS DE CLASIFICACIÓN Y DISCRIMINACION**, que son dos: el análisis de conglomerados y el análisis discriminante.

El **análisis de conglomerados** trata de reconocer patrones o formas para resumir la información contenida en las variables iniciales, que se agrupan en función de su mayor o menor homogeneidad. En resumen, buscamos grupos que sean mutuamente excluyentes, de forma que los elementos sean los más parecidos posible a los de su grupo y lo más diferentes posible a los de los otros grupos.

La parte más famosa del análisis de conglomerados es, sin duda, su representación gráfica, con árboles de decisión y dendrogramas, en los que se van separando de forma jerárquica grupos homogéneos cada vez más diferentes a los más alejados entre las ramas del árbol.

Pero, en lugar de querer segmentar la población, vamos a suponer que ya tenemos una población segmentada en un número de clases, k. Supongamos que tenemos un grupo de individuos definidos por un número p de variables aleatorias. Si queremos saber a qué clase de la población puede pertenecer un determinado individuo, recurriremos a la técnica del **análisis discriminante**.

Imaginemos que tenemos un nuevo tratamiento que es muy caro, así que solo queremos indicarlo en los pacientes que estemos seguros de que van a cumplir bien el tratamiento. Así, nuestra población está segmentada en cumplidores y no cumplidores. Nos sería muy útil seleccionar un conjunto de variables que nos permitiesen discriminar a qué clase puede pertenecer una persona concreta e, incluso, cuáles de estas variables son las que discriminan mejor entre los dos grupos. Así, mediremos las variables en el candidato al tratamiento y, utilizando lo que se conoce como criterio o regla de discriminación, lo asignaremos a uno u otro grupo y procederemos en consecuencia. Eso sí, no nos olvidemos, siempre habrá una probabilidad de equivocarse, por lo que nos interesará encontrar la regla discriminante que minimice la probabilidad de error de discriminación.

El análisis discriminante puede parecernos similar al análisis por conglomerados, pero, si lo pensamos, la diferencia es clara. En el análisis discriminante los grupos están previamente definidos (cumplidores o no cumplidores, en nuestro ejemplo), mientras que en el análisis por conglomerados buscamos grupos que no son evidentes: analizaríamos los datos y descubriríamos que hay pacientes que no se toman la pastilla que les mandamos, algo que ni se nos había pasado por la cabeza (además de nuestra ignorancia, demostraríamos nuestra inocencia).

Y aquí lo vamos a dejar por hoy. Hemos sobrevolado desde gran altura sobre el escarpado paisaje de la estadística multivariante y espero que haya servido para trasladarla del campo de lo esotérico al de lo exotérico (¿o era al revés?). No hemos entrado en la metodología específica de cada técnica, ya que podríamos haber escrito un libro

entero. Con entender qué es cada método y para qué sirve, más o menos, creo que tenemos bastante ganado. Además, los paquetes estadísticos los llevan a cabo, como siempre, sin esfuerzo.

Tampoco penséis que hemos tratado todos los métodos que se han desarrollado para el análisis multivariante. Existen otros muchos, como el análisis conjunto y el escalamiento multidimensional, muy utilizados en publicidad para determinar los atributos de un objeto que son preferidos por la población y cómo influyen en la percepción que tienen sobre el mismo. También podríamos perdernos entre otras técnicas más nuevas, como el análisis de correspondencias, o los modelos de probabilidad lineal, como el análisis logit y el probit, que son combinaciones de regresión múltiple y análisis discriminante, o los modelos de ecuaciones simultáneas o estructurales. Pero esa es otra historia…

Bibliografía

– Técnicas de la reducción de la dimensión a través de R. En: Marqués Asensio F, ed. R en profundidad. Programación, gráficos y estadística. RC Libros. Madrid, 2017; 359-94. (PDF)

– Técnicas de clasificación y segmentación a través de R. En: Marqués Asensio F, ed. R en profundidad. Programación, gráficos y estadística. RC Libros. Madrid, 2017; 395-410. (PDF)

– Guàrdia J, Benítez S. Introducción a la estadística multivariante. En: Peró Cebollero M, Leiva Ureña D, Guàrdia Olmos J, Solanas Pérez A, eds. Estadística aplicada a las ciencias sociales mediante R y R-Commander. Ibergarceta Publicaciones SL. Madrid, 2012; 499-563. (HTML)

– García Pérez A, ed. Métodos avanzados de estadística aplicada. Técnicas avanzadas. Universidad Nacional de Educación a Distancia. Madrid, 2005. (HTML)

Correspondencia al autor

Manuel Molina Arias
mma1961@gmail.com
Servicio de Gastroenterología.
Hospital Infantil Universitario La Paz.
Madrid. España.

Aceptado para el blog en agosto de 2020