Qué es análisis estadístico multivariado

HERNAN GARCIA

El análisis multivariado se refiere a un conjunto de métodos los cuales pueden analizar simultáneamente la relación existente entre variables correlacionadas. Kendell (1980) define "Análisis multivariado como la rama del análisis estadístico concerniente con las relaciones de conjuntos de variables dependientes".

Cuando se analizan varias características o variables de un mismo individuo o cuando éste es sometido a varios tratamientos, estas variables por lo general están correlacionadas. Una serie de análisis estadísticos univariados realizados separadamente para cada característica puede conducir a interpretaciones erróneas de los resultados puesto que se ignora la correlación o interdependencia entre variables.

En las ciencias sociales a veces es preciso combinar varias preguntas para representar una idea, por ejemplo la clase social a menudo se representa mejor por un conjunto de preguntas que incluyan el ingreso, la educación y la ocupación. Cuando se crean variables que son el resultado de la combinación de varias preguntas las técnicas univariadas hacen confuso el análisis y no permiten extraer toda la información del conjunto de datos.

Las técnicas multivariadas son una herramienta poderosa para analizar los datos en términos de muchas variables y permiten

extraer la máxima información posible del conjunto de datos. En la actualidad existen paquetes estadísticos tales como el: SAS (Sistema de Análisis Estadístico), SPSS (Paquete Estadístico para las Ciencias Sociales), STATGRAPHICS, etc., que permiten utilizar estas técnicas.

En el campo multivariado pueden utilizarse diferentes enfoques, tanto por los distintos tipos de situaciones que se presentan al obtener los datos, como por el objetivo específico del análisis. Los más importantes son:

SIMPLIFICACION DE LA ESTRUCTURA O REDUCCION DE LOS DATOS.

El objetivo es encontrar una manera simplificada de representar el universo de estudio. Esto puede lograrse mediante la transformación de un conjunto de variables interdependientes en otro conjunto de variables independientes o en otro conjunto de menor dimensión. Las técnicas que se utilizan con mayor frecuencia son el análisis por componentes principales y el análisis factorial.

CLASIFICACION

Este tipo de análisis permite ubicar las observaciones dentro de grupos o bien concluir que los individuos están dispersos aleatoriamente en el multiespacio. También pueden agruparse variables. Las técnicas empleadas son los métodos de clasificación jerárquicos y no jerárquicos y análisis discriminante.

INVESTIGACION DE LA DEPENDENCIA ENTRE VARIABLES

Para ello se seleccionan del conjunto ciertas variables (una o más) y se estudia su dependencia de las restantes. Entre los métodos para detectar dependencia comprenden el análisis de regresión múltiple, análisis de correlación canónica y análisis discriminante.

ANALISIS DE LA INTERDEPENDENCIA

El objetivo es analizar la interdependencia entre variables, la cual abarca desde la independencia total hasta la colinealidad cuando alguna de ellas es combinación lineal de las otras. Entre las técnicas para analizar la interdependencia entre variables o individuos se incluyen el análisis de factores, clasificación, el análisis de correlación canónica, el análisis por componentes principales.

FORMULACION Y PRUEBA DE HIPOTESIS

A partir de un conjunto de datos es posible encontrar modelos que permitan formular hipótesis en función de parámetros estimables. La prueba de este nuevo modelo requiere una nueva recopilación de datos a fin de garantizar la necesaria independencia y validez de las conclusiones. Una de las técnicas empleadas es la manova.

A continuación se describen brevemente algunas de las técnicas multivariadas más utilizadas.

METODO DE COMPONENTES PRINCIPALES

Mediante este método se obtienen componentes o combinaciones lineales de las variables originales que permiten simplificar el universo de estudio, centrándose en las componentes que sintetizan la máxima variabilidad residual. Los objetivos más importantes son:

- Generar nuevas variables que puedan expresar la información contenida en el conjunto original de datos.
- Reducir la dimensionalidad del problema que se está tratando, como paso previo para futuros análisis.
- Eliminar cuando sea posible, algunas de las variables originales ya sea porque ellas aportan poca información o porque una

variable contiene parte de información ya suministrada por otra u otras variables.

El análisis por componentes principales debe ser aplicado cuando se desea conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o características de los elementos.

ANALISIS FACTORIAL

Este es un término genérico para varias técnicas que pretenden explicar la correlación de un conjunto grande de variables en términos de un conjunto reducido de variables subyacentes denominadas factores. Al reducir el número de variables, los procedimientos tratan de retener tanto de la información como sea posible y de hacer de las variables restantes tan significativas y tan fáciles de manipular como sea posible. El propósito del análisis de factores es generar una comprensión de la estructura fundamental de las preguntas, variables u objetos y combinarlos en nuevas variables.

El análisis de factores permite generar varias soluciones para un mismo conjunto de datos, cada solución es generada por un esquema de rotación de factores, es decir, cada rotación tiene una interpretación diferente y esto se lo hace en términos de cargas o puntajes de factores. Cuando no se realiza ninguna rotación el análisis que se emplea es el de componentes principales.

CLASIFICACION

Una manera de analizar y estudiar un conjunto de individuos es clasificándolos en subconjuntos de acuerdo con algún objetivo predeterminado. En forma esquemática la clasificación trata el problema de particionar un conjunto en subconjuntos, tales que la diferencia entre elementos de un mismo subconjunto sea mínima y

sea máxima para los elementos de diferentes subconjuntos. La formulación matemática y estadística a este problema se hace mediante métodos como: modelos probabilísticos, teoría de grafos o criterios de optimización y algoritmos.

El uso de las técnicas de clasificación está subtendido por algunas ideas generales concernientes a las observaciones. Podrá tratarse de descubrir una partición que realmente exista (esta existencia es conjeturada antes del análisis estadístico o es revelada después del análisis). Inversamente, la partición puede ser empleada como instrumento para explorar los datos. Este último caso, es una generalización de histogramas unidimensionales, que con el objeto de facilitar el análisis, las observaciones son agrupadas en clases homogéneas.

Las técnicas de clasificación recurren a métodos algorítmicos y no a cálculos formalizados usuales. Básicamente se consideran dos tipos de métodos de clasificación: los métodos jerárquicos y los métodos no jerárquicos.

Las técnicas de clasificación jerárquica presentan una estructura de árbol, es decir, todos los individuos forman una clase, luego aparecen formando dos, tres, etc. clases y finalmente cada individuo forma una clase. Si se parte de n clases formadas cada una por un individuo y se van agrupando por pasos sucesivos hasta formar una sola clase, las técnicas se denominan aglomerativas, en caso contrario se denominan técnicas divisivas.

En los métodos de clasificación no jerárquicos el número de clases se establece a priori y el algoritmo de clasificación asigna los individuos a las clases, partiendo de algunos valores iniciales (puntos semillas) y buscando optimizar algún criterio establecido de antemano.

ANALISIS DISCRIMINANTE

Mientras que la meta del análisis de factores es generar dimensiones que maximicen la interpretación y expliquen la varianza, la meta del análisis discriminante es generar dimensiones que discriminen o separen los objetos tanto como sea posible, es decir, identifica grupos o conglomerados de atributos sobre los cuales difieren los objetos. Por ejemplo, el interés podría estar en determinar la forma como un usuario de un servicio difiere de un no usuario. Tal como en el análisis de factores, cada dimensión se basa en una combinación de los atributos fundamentales.

ANALISIS DE CORRELACION CANONICA

Este análisis es de los más utilizados y de los que más aplicaciones tiene. Consiste en buscar las máximas correlaciones posibles entre conjuntos de variables.

El análisis de correlación canónica tiene ciertas propiedades similares al análisis de componentes principales, sin embargo, éste considera relaciones dentro de un conjunto de variables, la correlación canónica lo hace entre dos conjuntos de variables. Una manera de ver este análisis es como una extensión de la técnica de regresión múltiple, que busca encontrar las relaciones entre las variables independientes y la variable dependiente.

El tipo de problemas para los cuales es útil el análisis de correlación canónica se lo puede ilustrar con el siguiente ejemplo. Un investigador desea explorar las posibles relaciones entre el conjunto de variables que caracterizan el logro académico de los estudiantes de cierta institución. La respuesta a esta pregunta se da mediante combinaciones lineales de las variables de la personalidad que están más correlacionadas con las combinaciones lineales de las variables del logro académico.

En la actualidad hay una variedad considerable de métodos multivariados que se han desarrollado con la evolución de los computadores. Algunos de estos son: escala multidimensional, análisis de series cronológicas, análisis path, mapas multidimensionales, correspondencia binaria y correspondencia múltiple.

BIBLIOGRAFIA

[1] Anderberg, M. CLUSTER ANALYSIS FOR APPLICATIONS. Academic Press. Londres, 1973.

[2] Baustista, L. y Ramos, J. ANALISIS DE DATOS DE ENCUESTAS Y TABULADOS. Universidad Nacional. Bogotá. 1988.

[3] Kendall, M.G. MULTIVARIATE ANALYSIS. Griffin. Londres. 1980.

[4] Lebart, L., Morineau, A. y Warwick, K. MULTIVARIATE DESCRIPTIVE STATICAL ANALYSIS. John Wiley. New York. 1984.

[5] Mardia, K., Kent, J. y Bibby, J. MULTIVARIATE ANALYSIS. Academic Press. Orlando. 1979.

[6] Morrison, D.F. MULTIVARIATE STATICAL METHODS. McGraw-Hill. New York. 1976.

[7] Seber, G. MULTIVARIATE OBSERVATIONS. John Wiley. New York. 1984.

DEPARTAMENTO DE MATEMATICAS Y ESTADISTICA
UNIVERSIDAD DE NARIÑO
PASTO - COLOMBIA