PANORAMA GENERAL DE LOS MÉTODOS MULTIVARIADOS

Aníbal Angulo

Los métodos multivariados son extraordinariamente útiles puesto que ayudan a los investigadores a dar sentido a grandes conjuntos de datos, complicados y complejos, que constan de una cantidad de variables medidas en grandes números de unidades experimentales. La importancia y la utilidad de los métodos multivariados aumentan cuando se incrementa el número de variables que se están midiendo y el número de unidades experimentales que se están evaluando.

A menudo, el objetivo primario de los análisis multivariados es resumir grandes cantidades de datos en relativamente pocos parámetros. El tema subyacente de muchas técnicas multivariadas es la simplificación.

El interés de los análisis multivariados es encontrar relaciones entre 1) las variables respuesta, 2) las unidades experimentales, y 3) tanto las variables respuesta como las unidades experimentales. Se podría decir que existen relaciones entre las variables respuesta cuando, en realidad, algunas de las variables están midiendo una cantidad común.

Podrían existir relaciones entre las unidades experimentales si algunas de ellas son semejantes entre sí. Muchas técnicas multivariadas tienden a ser de naturaleza exploratoria en lugar de confirmatoria. Es decir, muchos métodos multivariados tienden a motivar hipótesis y no probarlas. Considere una situación en la cual un investigador tiene 50 variables medidas sobre más de 3000 unidades experimentales. Los métodos estadísticos tradicionales suelen exigir que un investigador establezca algunas hipótesis, reúna algunos datos y, a continuación, use esos datos para comprobar o rechazar esas hipótesis. Una situación alternativa que se da frecuentemente es un caso en el cual un investigador dispone de una gran cantidad de datos y se pregunta si pudiera haber una información valiosa en ellos. Las técnicas multivariadas suelen ser útiles para examinar los datos en un intento por saber si hay información que valga la pena y sea valiosa en esos datos.

Una distinción fundamental entre los métodos multivariados es que algunos se clasifican como “técnicas dirigidas por las variables”, en tanto que otros se clasifican como “*técnicas dirigidas por los individuos*”.

Las técnicas dirigidas por las variables son aquellas que se enfocan primordialmente en las relaciones que podrían existir entre las variables respuesta que se están midiendo.

Las técnicas dirigidas por los individuos son la que se interesan principalmente en las relaciones que podrían existir entre las unidades experimentales o individuos que se están midiendo, o en ambos. Algunos ejemplos de este tipo de técnica se encuentran en el análisis discriminante, el análisis de agrupaciones y el análisis multivariado de la varianza (manova): (Multivariate Analysis of Variance)

Muchos métodos multivariados ayudan a los investigadores a crear nuevas variables que tengan propiedades deseables, entre éstas indicamos las siguientes técnicas.

Análisis de Componentes Principales (PCA).

Cuando un investigador está empezando a pensar acerca del análisis de un nuevo conjunto de datos, debe considerar varias preguntas acerca de ellos. Las preguntas importantes incluyen: 1) ¿Existen algunos aspectos en los datos que resultan extraños? 2) ¿Se puede suponer que los datos están distribuidos normalmente? 3) ¿Hay algunas anormalidades en los datos? 4) ¿Existen datos outliers[6](atípicos)?

La razón más importante para realizar un Análisis de Componentes Principales es usarlo como herramienta para cribar los datos de variables múltiples. Se pueden crear nuevas variables, llamadas calificaciones de componentes principales, que se pueden usar como entrada para programas de trazado de gráficas y situación de datos y, con frecuencia, un examen de las presentaciones gráficas resultantes revelarán las anormalidades en los datos que se está planeando analizar. Además, se pueden analizar por separado las mediciones de los componentes principales para ver sí se cumplen las

[6] Las unidades experimentales cuyas variables medidas parecen incoherentes o son atípicas con relación a las mediciones realizadas en las otras unidades experimentales suelen llamarse datos outliers.

hipótesis relativas a la distribución, como la normalidad de las variables y la independencia de las unidades experimentales. A menudo, se requieren esas hipótesis para que sean válidos ciertos tipos de análisis estadísticos.

El Análisis de Componentes Principales suele ser bastante útil para los investigadores que desean realizar la división en subgrupos de las unidades experimentales, de modo que unidades experimentales similares pertenezcan al mismo subgrupo. En este caso, se pueden usar las calificaciones de los componentes principales como entrada para los programas de agrupación, lo que suele incrementar la eficacia de estos programas, reduciendo al mismo tiempo el costo de su uso. Además, se recomienda usar las mediciones de los componentes principales para ayudar a validar los resultados de los programas de agrupación.

Análisis por Factores

El análisis por factores (FA: Factor Analysis) es una técnica que se emplea frecuentemente para crear nuevas variables que resuman toda la información de la que podría disponerse en las variables originales. El análisis por factores también se usa para estudiar las relaciones que podrían existir entre las variables medidas en un conjunto de datos. Un objetivo básico de esta técnica es determinar si las variables respuesta exhiben patrones de relaciones entre sí, tales que esas variables se pueden dividir en subconjuntos de modo que las variables en un subconjunto estén fuertemente correlacionadas con cada una de las otras y que las variables en subconjuntos diferentes tengan bajas correlaciones entre sí. Por tanto, el FA se usa con frecuencia para estudiar la estructura de correlación de las variables en un conjunto de datos. Una semejanza entre el FA y el PCA es que este último también se puede usar para crear nuevas variables que no están correlacionadas entre sí. Esas variables se llaman clasificación de factores.

Una ventaja que parece tener el FA sobre PCA, es que cuando se están creando nuevas variables, generalmente, las nuevas variables creadas por el FA son mucho más fáciles de interpretar que las creadas por el PCA. Si un investigador desea crear un conjunto más pequeño de nuevas variables que se puedan interpretar y que resuman la mayoría de la información existente en las variables medidas entonces se recomienda usar el Análisis por Factores.

Análisis Discriminante

El análisis discriminante (DA Discriminant Analysis) se usa principalmente para clasificar individuos o unidades experimentales en dos o más poblaciones definidas de manera única. Para desarrollar una regla discriminante que clasifique las unidades experimentales en una de varias categorías posibles, el investigador debe tener una muestra aleatoria de unidades experimentales de cada grupo posible de clasificación. Entonces, el DA proporciona los métodos que permitirán a los investigadores establecer reglas que se puedan emplear para clasificar otras unidades experimentales en uno de los grupos de clasificación.

Análisis Discriminante Canónico

El análisis discriminante canónico (CDA: Canonical Discriminant Analysis) es un procedimiento con el que se crean nuevas variables que contienen toda la información útil para la discriminación de la que se dispone en las variables originales. A menudo, estas nuevas variables conducen a reglas más sencillas para clasificar las unidades experimentales en los diferentes grupos.

Regresión Logística

Con frecuencia se usa la regresión logística para modelar la probabilidad de que una unidad experimental caiga en un grupo particular, con base en la información medida en la propia unidad. Estos modelos se pueden usar con fines de discriminación. En el caso de las tarjetas de crédito Bancario, se puede modelar la probabilidad de que un individuo con ciertas características demográficas sea un buen riesgo de crédito; después de desarrollar este modelo, se puede usar para predecir la probabilidad de que un nuevo solicitante caiga en un cierto grupo de riesgo. Los individuos cuya probabilidad pronosticada para el grupo de “buen riesgo” sea mayor que 0.5 se determinan como buenos riesgos de crédito.

Análisis por agrupación

Suponga que un arqueólogo descubre un gran escondite de fragmentos de alfarería y toma pequeñas muestras representativas de cada fragmento, y que cada muestra se puede analizar y que se pueden averiguar las cantidades relativas de

diferentes elementos químicos, como zinc, magnesio, hierro, etc. El Arqueólogo quiere separa los fragmentos en montones distintos, de modo que los que queden en cada uno de esos montones provenga de la misma pieza de alfarería. Resulta claro que esto puede ser una tarea difícil porque el arqueólogo no sabe cuántos montones de fragmentos resultarán, cuantos fragmentos quedaran en cada montón o si algunos fragmentos en realidad pertenezcan al mismo montón. El análisis por agrupación es un método multivariado que puede ayudar a resolver este problema.

El análisis por agrupación (CA: Cluster Analysis) es semejante al discriminante en el sentido de que se usan para clasificar individuos o unidades experimentales en subgrupos definidos de manera única. Este análisis se puede emplear cuando el investigador cuenta con una muestra aleatoria previamente obtenida de cada uno de los subgrupos definidos de manera única. El análisis por agrupación trata los problemas de clasificación cuando no se sabe de antemano de cuáles subgrupos se originan las observaciones.

Análisis Multivariado de la varianza

El análisis multivariado de la varianza (MANOVA) es una generalización del análisis univariado de la varianza (ANOVA: Analysis of Variance), una técnica usada para comparar las medias de varias poblaciones en una sola variable medida.

Cuando se miden varias variables en cada unidad experimental, podría producirse un ANOVA sobre cada variable medida, usando una variable a la vez; por ejemplo si se miden 30 variables, un investigador podría producir 30 análisis separados, uno para cada variable. Sin embargo, esto no es inteligente, por desgracia la mayoría de los experimentos se están analizando con la aplicación de análisis de una variable a la vez.

Los estadísticos promueven dos objeciones principales para los análisis separados de cada variable medida. Una objeción es que las poblaciones que se están comparando pueden ser diferentes en alguna variable, pero no en otras. A menudo el investigador se encuentra confuso en cuanto a cuales poblaciones son diferentes en realidad y cuáles son semejantes. Los análisis multivariados de la varianza pueden ayudarlos a comparar varias poblaciones al considerar, simultáneamente todas las variables medidas.

Una segunda objeción es que se tiene protección inadecuada contra errores tipo I cuando se realizan análisis de una variable a la vez. Recuerde lo visto en los cursos

introductorios de estadística, lo que ocurre en error tipo I siempre que se rechaza una hipótesis verdadera. Cuanto más variables analice un investigador, mayor es la probabilidad de que por lo menos una de las variables que se están analizando y de hallar por lo menos uno de estos análisis estadísticamente significativos, es decir, tiende a uno.

Es evidente que el gran riesgo de cometer errores de tipo I deba inquietar a los experimentadores. El investigador debe estar confiado cuando afirma que dos o más poblaciones tienen medias diferentes con respecto a una variable medida y debe confiar que conduzcan a análisis similares sobre conjuntos semejantes de datos.

Debe realizarse un MANOVA siempre que se están comparando entre sí dos o más poblaciones diferentes sobre un número grande de variables respuesta.

Si un MANOVA muestra diferencias significativas entre las medias de las poblaciones, entonces el investigador puede confiar en que verdaderamente existen diferencias reales. En este caso, resulta razonable considerar el análisis de una variable a la vez para detectar dónde ocurren en realidad las diferencias.

Si el ANOVA no revela diferencias significativas entre las medias de las poblaciones, entonces el investigador debe tener precaución extrema al interpretar los análisis de una variable a la vez. Esos análisis pueden identificar como positivos falsos.

Análisis de Variables Canónicas

El análisis de variables canónicas (CVA: Canonical Variates Analysis) es un método en el que se crean nuevas variables con conjunción con los análisis multivariados de la varianza. Estas nuevas variables son útiles porque ayudan a los investigadores a determinar dónde ocurren las diferencias importantes entre las medias de las poblaciones, cuando se están comparando poblaciones sobre muchas variables diferentes mediante el uso simultáneo de todas las variables medidas. En ocasiones, las variables canónicas pueden surgir diferencias importantes que, del contrario podrían pasarse por alto.

Análisis de Correlación Canónica.

El análisis de correlación canónica es una generalización de la correlación múltiple en los problemas de regresión. Se requiere que las variables respuesta se dividan en dos grupos. La asignación de las variables en estos dos grupos siempre debe motivarse por la naturaleza de las variables respuesta y nunca por una inspección de los datos. Por

ejemplo, una asignación legítima sería aquella en la que las variables en uno de los grupos sean fáciles de obtener y no caras para medirse, mientras que las que se encuentren en el otro grupo sean difíciles de obtener y caras para medirse. Una cuestión básica que se espera responder con el análisis de correlación canónica es si se pueden usar las variables que se encuentran en uno de los grupos para predecir las variables en el otro. Cuando se puede, entonces este análisis intenta resumir las relaciones entre los dos conjuntos de variables, mediante la creación de nuevas variables a partir de cada uno de los dos grupos de variables originales.

Bibliografía.

Alvarez Cáceres, R (1995) Estadística Multivariante y no Paramétrica Aplicada a las Ciencias de Salud.

Anderberg M. J. Clusters Analysis for Applications N. 4. Academic Press 1973

Dallas E. Jhonson Métodos Multivariados Aplicados al Análisis de Datos.

M. Ángeles Cea, D Ancona Análisis Multivariante Teoría y Práctica en el Investigación Social

----- O -----

Estadístico

Es alguien que es bueno con los números pero carece de personalidad suficiente como para ser Contador.

Un Estadístico podría meter su cabeza en un horno y sus pies en hielo, y decir que en promedio se encuentra bien.

Tres Estadísticos salen de cacería. En esto ven a un ciervo pastando, tranquilamente. El primer estadístico apunta, dispara, y la bala pasa dos metros y diez centímetros a la derecha del ciervo. El segundo estadístico apunta, dispara y la bala pasa dos metros y diez centímetros a la izquierda del ¡Bien, le hemos dado!

Índice