EL ANÁLISIS DISCRIMINANTE

Por Tevni Grajales G.

El análisis discriminante como la regresión logística son técnicas multivariantes de dependencia como la regresión lineal múltiple. La diferencia inicial consiste en que la variable dependiente o criterio es una variable no métrica (categórica). A continuación una breve discusión de esta técnica. En primer lugar se especifican los objetivos que se procuran cuando se usa la técnica, después se presentan los supuestos que deben ser considerados para proceder en el uso de la técnica y por último se comenta, de manera suscinta, el proceso usado.

Objetivos de la Técnica:

1. Distinguir entre diversos grupos mutuamente excluyentes. Como puede ser entre buenos y malos clientes de un empresa; alumnos responsables e irresponsables en una institución.

- Para distinguir o clasificar las observaciones de una investigación.
- Para detectar el por qué de las diferencias.
- Para pronosticar a qué grupo pertencerá una persona de acuerdo a sus características.

2. Identificar las variables que son importantes para distinguir entre los grupos o fin de desarrollar un procedimiento para predecir la membresía de aquellos casos que no han sido estudiados. Como puede ser el caso de responder a una solicitud de préstamo de un cliente en una empresa, o la solicitud de empleo por parte de un estudiante de la institución.

Se debe tener en cuenta que:

- Los grupos de la variable dependiente deben ser mutuamente excluyentes.
- No es confiable el grado de error si se pronostica con los mismos datos (use un % de la población o muestra observada).
- Se pueden tener n variables independientes siempre y cuando sean medidas métricas.
- La variable dependiente no debe ser métrica sino categórica (nominal) para formar grupos.

Supuestos que fundamentan la técnica:

Para que se logre una *Función Linear Discriminante* óptima se supone que:

- Cada grupo debe ser una muestra de una población normal multivariada.
- Las matrices de covarianza poblacional deben ser iguales. (Se entiende por varianza la dispersión de los datos con respecto a su media y covarianza la dispersión de los datos tomando en cuenta las dos variables como si fueran una sola). Se utiliza la M de Box para probar la igualdad de las matrices de covarianza de la variables independientes entre los grupos que forman la variable dependiente. Si la M de Box resulta significativa no se puede sustentar la hipótesis que señala que las matrices de covarianza poblacional son iguales y por consiguiente no hay lugar al uso de la técnica.

La Selección de los Casos para el Estudio

Para discriminar entre grupos es necesario contar con un conjunto de casos que hayan sido claramente identificados como componentes indudables de uno de los grupos.

Por lo que resulta de suma importancia seleccionar de manera adecuada los casos que servirán como criterios de selección. De modo que el objetivo de esta fase es identificar en la muestra los casos que habiendo sido ubicados de manera apropiada y precisa como pertenecientes a uno de los grupos de clasificación, dispongan de información completa respecto a las variables que serán usadas para definir los grupos (variables predictoras).

Viera (1997) realizó un estudio entre 835 maestros de secundaria del estado de Nuevo León con el fin de medir su grado de satisfacción en el trabajo y algunas variables predictoras del mismo. Como parte del trabajo se incluyeron escuelas estatales, transferidas y particulares. Una vez terminado el estudio, se consiguió autorización de parte de la investigadora para realizar análisis estadísticos posteriores, a partir de la base de datos elaborada para el estudio. Y a continuación se presenta un estudio discriminante con el 50% de la muestra a fin de elaborar una ecuación discriminante que permita clasificar al maestro según el tipo de escuela en la que trabaja.

Entre otras cosas se procedió a identificar las características más importantes que comparten los maestros de cada una de estas escuelas con el fin de determinar una función discriminante. A fin de verificar la adecuación de los datos a los supuestos correspondientes a esta técnica, se procedió a observar la distribución de cada una de las variables y posteriormente se probó la igualdad de la covarianza entre los grupos usando la M de Box lo cual dio como resultado una M = 13.41647 con una F6,69687 =2.20453 (P.0395) resultado que condujo a cuestionar el hecho que las matrices de covarianza poblacional sean iguales. Si la significatividad estadística de M es mayor que el nivel crítico (por ejemplo .01) se sustenta la igualdad de varianzas. Pero si la prueba muestra una significatividad estadística se consideran los grupos como diferentes y se violenta el supuesto. En el caso particular, siendo que la probabilidad es mayor a .01 se procede con la técnica.

Al observar las respuestas de los 389 maestros observados, 218 eran de escuelas estatales, 136 de transferidas y 35 de particulares. Además se observó que 31 casos no ofrecieron la información requerida.

La Verificación de la Diferencia entre los Grupos

Resulta necesario verificar que en efecto los grupos son diferentes, es decir que son excluyentes entre sí. Esto se puede observar por medio de:

- La diferencia entre las medias de los grupos. Se entiende que si los grupos son excluyentes, sus medias deben ser significativamente diferentes.
- Lambda de Wilks que resulta ser la razón entre la suma de los cuadrados dentro de los grupos y la suma total de los cuadrados. Un valor Lambda de uno ocurre cuando la media de todos los grupos observados es igual. Lo que significa que a medida que el valor Lambda se acerca a cero, la mayor parte de la variabilidad es atribuible a la diferencia entre las medias de los grupos.

Otro aspecto muy importante a considerar es el grado de interdependencia existente entre las variables dado que los análisis multivariados son muy afectados por la interdependencia.

Esto se puede observar en el SPSS mediante una matriz de correlación denominada pooled within-grupos en la cual es importante identificar las variables cuyas correlaciones son grandes, casos que deben constituir una notable minoría; de otra manera es muy probable que no se cuente con material apropiado para realizar el trabajo propuesto.

La Estimación de los Coeficientes b

Una vez determinada la legitimidad de las variables, los casos y los grupos en estudio, se debe proceder a analizar las variables en conjunto.

El análisis discriminante, como otros procedimientos estadísticos multivariados, procura resumir en un índice, toda la información contenida en diversas variables independientes. Para lograr dicha finalidad se buscará estimar el peso de las variables , que permita lograr la mejor separación posible entre grupos.

La ecuación lineal discriminante es similar a la regresión múltiple:

$$D = b\,o + b\,1X1 + b\,2X2 + b\,3X3 + ... + b\,pXp$$

Siendo X los valores de las variables predictoras y b los coeficientes estimados a partir de los datos. En el análisis discriminante se busca separar lo más posible (diferenciar) a los promedios de cada grupo pero respecto a la dispersión de los datos. Entre mayor dispersión de datos exista, será mayor el empalme de los datos, por lo que en al valorar la función, es necesario tomar en cuenta la dispersión o variación de los datos, además de la distancia entre los promedios.

A fin de que la función lineal discriminante pueda separar o diferenciar entre los grupos, estos deben diferir en sus valores D. Por lo tanto los coeficientes (b) se escogen de manera que los valores de la función discriminante difieran tanto como sea posible entre los grupos. {Máxima S $(m_1 - m_2)^2 / s^2$ }.

El Cálculo de los Valores Discriminantes

Para cada uno de los casos estudiados se puede calcular un puntaje (valor) discriminante D haciendo uso de los coeficientes de la función. Se los multiplica por el correspondiente valor de la variable en cada caso y luego se suman los

productos.

Según el ejemplo que aparece en el anexo la función 1 operaría de la siguiente manera considerando las dos variables incluidas (tiempo de servicio y satisfacción).

$$D = -2.10 + .97(tise) - .38(satisf)$$

Una vez determinado el puntaje (valor) discriminante D, resulta posible obtener una regla para clasificar cada caso en uno de los grupos. Con ese fin se usa la regla de Bayes, por medio de la cual se estima la probabilidad de que un caso en particular con un puntaje discriminante D pertenezca a un grupo en particular. En este procedimiento se distingue tres tipos de probabilidades:

La probabilidad previa o P(Gi) Calculada en una muestra representativa de la población como la proporción observada de los casos correspondientes a cada grupo.

La probabilidad condicional o P(D/Gi) La se cual calcula si los valores D están distribuidos de manera normal para cada grupo y los parámetros de las distribuciones pueden ser estimados. Esta representa la probabilidad de obtener un valor D particular si el caso observado es miembro de un grupo determinado.

La probabilidad posterior o P(Gi/D) La probabilidad condicional de D dado el grupo, provee una idea de los valores correspondientes a los miembros de un grupo en particular. Pero cuando no se conoce a qué grupo pertenece, se requiere estimar cuánto se asemeja a la membresía en los diferentes grupos a partir de la información disponible. A esto es lo que se denomina probabilidad posterior P(Gi/D) y se puede estimar a partir de P(D/Gi) y P(Gi) usando la regla de Bayes. El caso en observación se clasificará en base a su puntaje discriminante D, en el grupo para el cual la probabilidad posterior se más grande.

Coeficientes de Funciones Discriminantes

Existen diferentes conjuntos de coeficientes según el tipo de función lineal, de los cuales mencionaremos dos:

Coeficientes de la función lineal discriminante también conocidos como coeficientes de la función canónica discriminante siendo que son idénticos a los que se obtienen por medio de un análisis canónico de correlación. Se trata de pesos discriminantes son determinados por la estructura de varianza de las variables originales a través de los grupos del a variable dependiente (criterio). Las variables independientes con un poder discriminante grande por lo general presentan pasos grandes y las que tienen poco poder discriminante presentan pesos pequeños. Aunque la existencia de multicolinearidad entre las variables independientes puede conducir a una excepción de la regla.

Coeficientes de la función lineal discriminante de Fisher o coeficientes de clasificación que pueden ser usados directamente para clasificar. Se obtiene un conjunto de coeficientes para cada uno de los grupos y se asigna cada caso al grupo para el cual tiene el puntaje discriminante D más grande. Se trata de un método por el cual se determina una función lineal para cada grupo en la variable dependiente. La clasificación logra calculando un puntaje para cada observación en la función de clasificación de cada grupo y posteriormente asignado cada caso al grupo en el que tiene el puntaje más alto. Esto difiere de lo que se conoce como puntaje Z de discriminación el cual se obtiene para cada función discriminante.

Las resultados de la clasificación son iguales para ambos métodos si se usan todas las funciones canónicas discriminantes (Kshirsager & Arseven, 1975; Green, 1979).

La prueba de hipótesis para el modelo discriminante: si la diferencia en los centroides (los promedios de cada grupo de las variables independientes) o las distancias que presentan los promedios de las variables involucradas en el modelo son significativas. Si en efecto las diferencias son significativas, esto quiere decir que el modelo es bueno para discriminar. Se

puede usar la prueba Mahalanobis-D^2 en la que se obtiene un valor F calculado para comparar con un F de la tabla la hipótesis nula dice: el modelo discriminante no es bueno para discriminar o clasificar observaciones.

Si la función discriminante es estadísticamente significativa y su capacidad para clasificar es aceptable, el investigador puede dedicarse a realizar interpretaciones sustantivas de sus resultados. Este proceso involucra el examen de la función discriminante para determinar la importancia relativa de cada una de las variables independientes en su capacidad para discriminar entre los grupos. Los tres métodos para determinar la importancia relativa son
1. Los pesos discriminantes estandarizados.La forma tradicional de interpretar la funcion discriminante consiste en observar el signo y la magnitud de los pesos discriminantes estandarizados asignados a cada una de las variables en la función. Cuando se ignora el signo, cada peso representa la contribución relativa de su variable a la función. Las variables independientes con mayor peso contribuyen más al poder discriminante de la función. La forma de interpretar esto es análoga a los pesos beta del análisis de regresión y por consiguiente está sujeto a las mismas críticas.
2. La carga discriminante (estructura de correlaciones) miden la correlacion lineal simple entre cada variable independiente y la función discriminante. La carga discriminante refleja la varianza que la variable independiente comparte con la función discriminante y puede ser interpretada como carga factorial al evaluar la contribución relativa de cada variable independiente respecto a la función discriminante.
3. Los valores F parciales. Cuando se utiliza el método stepwise se obtiene una forma adicional para interpretar el poder discriminante relativo de las variables independientes por medio del uso de la F parcial. Esto se logra examinando el tamaño absoluto de los valores F significativos y ordenándolos en de mayor a menor. Grandes valores F indican mayor poder discriminatorio. Esto ofrece la oportunidad de observar el nivel de significatividad de cada una de las variables.

Los resultados del Anexo D incluyen una clasificación de resultados que muestra el porcentaje de casos que logran ser clasificados correctamente en su grupo..

Anexos disponibles por solicitud a tevgra@umontemorelos.edu.mx
Altius
tgrajales.net