Análisis Discriminante

Curso: IN540

Profesor: Marcelo Henriquez

Auxiliar: José Miguel Carrasco

Definición

¨ Cómo técnica de análisis de dependencia:

Pone en marcha un modelo de causalidad en el que la variable endógena es una variable NO MÉTRICA y las independientes métricas.

¨ Cómo técnica de análisis de clasificación:

Ayuda a comprender las diferencias entre grupos. Explica, en función de características métricas observadas, porqué los objetos/sujetos se encuentran asociados a distintos niveles de un factor.

Las diferencias vienen por

- ¨ El análisis de regresión:
  En la regresión, la endógena es métrica

- ¨ El análisis ANOVA: En el ANOVA, la endógena es métrica y las exógenas NO MÉTRICAS (al contrario que en el discriminante)

- ¨ El LOGIT - PROBIT: Idéntica al discriminante en el objetivo pero apoyada en técnicas de estimación paramétrica idénticas a la regresión y no en análisis de descomposición de la varianza:

- (1) DV: Más adecuada para factores sólo binarios
- (2) DV: Más compleja de cálculo - interpretación
- (3) V: Se ve menos afectada por incumplimientos de supuestos teóricos necesarios a priori (normalidad, por ejemplo)
- (4) V: Permite incorporar explicativas no métricas en forma de ficticias
- (5) Los resultados admiten explotación en términos de probabilidad

Análisis discriminante descriptivo (ejemplo)

- (Objetivo)
  Se desea caracterizar el perfil de los compradores de un determinado producto en un determinado establecimiento.

- (Diseño)
  Para ello, se diseña una muestra con 100 compradores y 100 no compradores y se toman datos de renta, edad y cercanía al establecimiento de venta.

- (Resultado)
  El análisis discriminante establecerá la importancia relativa de cada uno de estos atributos en la decisión de compra permitiendo orientar mejor la política promocional o de distribución del producto.

Análisis Discriminante Predictivo (ejemplo)

- (Objetivo)

  Se desea prever el riesgo de morosidad relativa a los préstamos personales en una entidad bancaria.

- (Diseño)

  Se explota el fichero histórico de clientes morosos - no morosos y se observan variables cuantitativas potencialmente explicativas: renta total, edad, créditos adicionales, años de estabilidad laboral, etc...

- (Resultado)

  Aplicando el modelo estimado con el fichero histórico, el análisis permitirá anticipar el riesgo de morosidad de nuevos clientes.

ETAPAS DE UN ANÁLISIS DISCRIMINANTE

- Selección de variables dependiente e independientes
- Selección del tamaño muestral
- División de la muestra
- Chequeo de las hipótesis de partida
- Estimación del modelo
- Validación de las funciones discriminantes
- Contribución de las variables a la capacidad discriminante de las funciones
- Valoración de la capacidad predictiva
- Utilización funciones

Recordar

Los grupos deben ser mutuamente excluyentes.

Las variables explicativas:
(1) no deben ser excesivas
(2) deben atender siempre al objetivo conceptual
(3) pueden someterse a un test univariante de diferencia de medias o un test ANOVA

Selección de método

- ¨ Método simultáneo o por etapas:

(1) estimación en una sola etapa (número reducido de variables, interés por el conjunto)

(2) estimación polietápica: selección de menos a más, analizando las interacciones de las variables discriminantes (amplio número de variables, dudas sobre el modelo teórico)

- ¨ Método cálculo : Método de Fisher, D de Mahalanobis, ….

Método de Fisher

- ¨ Variable Y (Función Discriminante): combinación lineal de las variables originales "X" que:

(1) Presente la mínima variación INTRA grupal

(2) Presente la máxima variación ENTRE grupal

- ¨ La función discriminante no será única: si se parte de una clasificación en "g" grupos, se obtendrán varios conjuntos de parámetros, es decir, varias funciones discriminantes (Menor de “g-1” o “p”)

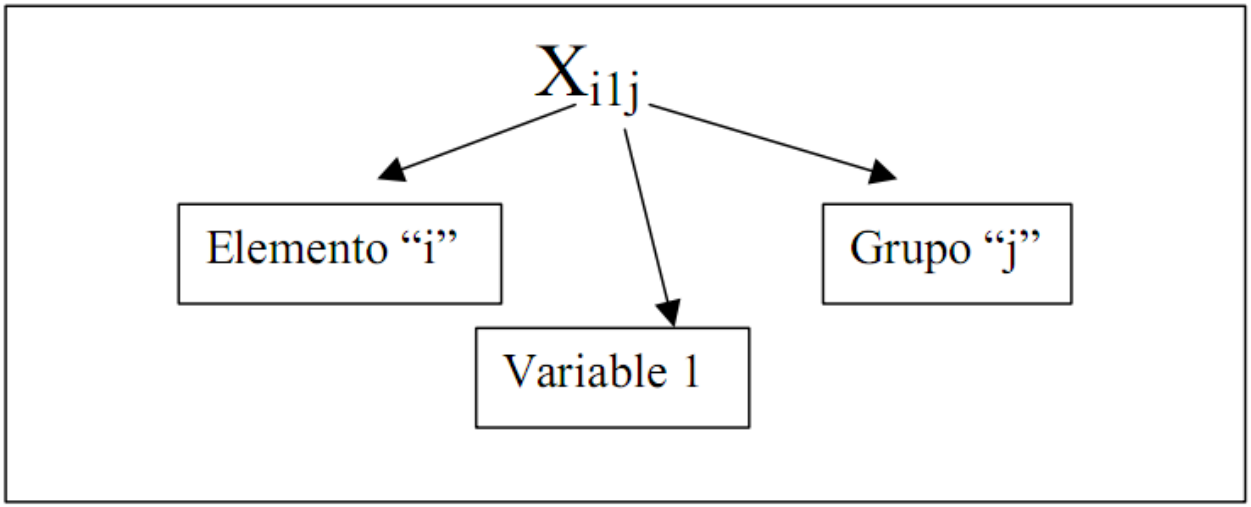
Método de Fisher

Elemento	Grupo	**X1**	**X2**	**X3**		**Y=f(x1,x2,x3)**
1	A	**25**	**25**	**23**	→	10
2	A	**15**	**14**	**26**		12
3	A	**14**	**13**	**21**	→	11
4	B	**25**	**18**	**41**		5
5	B	**65**	**14**	**18**	→	4
6	B	**15**	**18**	**48**		4

Método de Fisher

(Obtención de la/s funciones discriminantes)
(Planteamiento para “g” grupos y “p” variables)

Partimos de:
- “g“ grupos que representamos con el subíndice j=1,2.....g
- “p” variables
- “n” elementos para cada una de ellas que representamos con i=1,2.....n



Método de Fisher

Conforme a esta nomenclatura definimos las siguientes matrices:
MATRIZ DE OBSERVACIONES PARA EL ELEMENTO “i” DEL GRUPO “j”

$$X_{ij} = \begin{pmatrix} X_{i1j} \\ X_{i2j} \\ . \\ . \\ X_{ipj} \end{pmatrix} \forall \begin{matrix} i = 1,2,....n_j \\ j = 1,2.....g \end{matrix}$$

Método de Fisher

MATRIZ DE MEDIAS DEL GRUPO "j"

$$X_{ij} = \begin{pmatrix} X_{i1j} \\ X_{i2j} \\ . \\ . \\ X_{ipj} \end{pmatrix} \forall \begin{matrix} i = 1,2,....n_j \\ j = 1,2.....g \end{matrix}$$

MATRIZ DE MEDIAS TOTALES

$$\overline{X} = \begin{pmatrix} \overline{X}_{\bullet 1\bullet} \\ \overline{X}_{\bullet 2\bullet} \\ . \\ . \\ \overline{X}_{\bullet p\bullet} \end{pmatrix}$$

Método de Fisher

Definidas estas matrices la variación Entre e Intra será:

$$E = \sum_{j=1}^{g} n_j \cdot (\bar{X}_{\bullet j} - \bar{X})(\bar{X}_{\bullet j} - \bar{X})'$$

[matriz de orden (p x p)]

$$I = \sum_{j=1}^{g} \sum_{i=1}^{n_j} \cdot (X_{ij} - \bar{X}_{\bullet j})(X_{ij} - \bar{X}_{\bullet j})'$$

[matriz de orden (p x p)]

Método de Fisher

De esta forma, la ratio a maximizar sería:

$$F = \frac{V.Entre}{V.Intra}$$

Sin embargo nuestro objetivo es encontrar los parámetros "b" de la combinación lineal:
Y = b' X
que maximicen este ratio por lo que debemos expresar estas V.Intra y V.Entre en función de los parámetros "b" de este modelo.

Método de Fisher

Dicho de otro modo, lo que queremos es maximizar la V. Entre y minimizar la V.Intra para la variable discriminante “y”. Puede demostrarse que:

$$SCE_y = b'Eb$$

$$SCI_y = b'I\ b$$

Por lo que, lógicamente, el ratio a maximizar puede expresarse como:

$$F = \frac{SCE_y / g - 1}{SCI_y / n - g}$$

Método de Fisher

que obviando los grados de libertad supone:

$$max(\lambda) = max \frac{b' Eb}{b' Ib}$$

Esta operación arroja varias soluciones del conjunto de parámetros “b” lo que significa que para un determinado conjunto de datos siempre encontraremos más de una solución. (El menor de “g-1” o “p”).

Método de Fisher –

Validación Resultado

Autovalores: En el método de Fisher, la obtención de las distintas funciones se deriva de un proceso de obtención de raíces y vectores propios de una forma cuadrática. La suma de cuadrados entre grupos de cada función discriminante, viene definida por un autovalor l(i).

Ratio Autovalor / Suma autovalores: capacidad discriminante relativa, pero no absoluta.

Test Bartlett: El test de Bartlett, distribuido como una chi cuadrado de p(g-1) grados de libertad contrasta secuencialmente la hipótesis de todos los autovalores como 0:

$$B = \left[n - 1 - \frac{p+g}{2} \right] \sum_{\varphi=1}^{r} ln\,(1+\lambda_{\varphi})$$

Correlación canónica función - variable clasificación original:
Coeficientes elevados anticipan adecuada capacidad discriminante

Método de Fisher –

Valoración de la capacidad predictiva

Los contrastes de significación no informan sobre la capacidad predictiva del modelo

Cálculo de la Puntuación de Corte Óptima

Cálculo de la Puntuación de Corte Óptima modificada para el caso de grupos de tamaño desigual representativos de la estructura de la población (muestreo aleatorio).

Construcción de la “Matriz de Confusión”

Análisis de casos individuales (detección de nuevas variables a incluir en el análisis)

Método de Analisis Factorial SPSS

- Ahora

Análisis Discriminante

Curso: IN540

Profesor: Marcelo Henriquez

Auxiliar: José Miguel Carrasco