BIBLIOTECA DE PROGRAMAS

EL PROGRAMA BMDO4M DE ANALISIS DISCRIMINANTE.

1. Introducción al problema de la discriminación y clasificación de 2 grupos (T^2 de Hottelling, D^2 de Mahalanobis).

Es bien sabido que a la hora de discriminar entre dos situaciones o dos grupos de individuos sobre los que se ha medido una única variable respuesta, el test de la t de student(*) para la diferencia de medias es un buen modelo para la discriminación siempre que la variable respuesta sea normal y se pueda suponer igualdad de varianzas en los 2 grupos. Bien distinta es la situación en la que la variable respuesta no es unidimensional, es decir hemos medido más de una variable en cada individuo y queremos discriminar entre los dos grupos en base a estas mediciones. Un método potente para hacerlo es el análisis discriminante siempre que se pueda admitir la normalidad de las variables estudiadas y la homogeneidad de las matrices de covarianza para los dos grupos.

2. Modelo matemático

Sea $x^t=(x_1,...x_m)$ la variable aleatoria normal m-dimensional que observamos. Supongamos que hemos estraido muestras de tamaño n_1 del 1º grupo y de tamaño n_2 del 2º grupo designamos por $\bar{x}_1$ y $\bar{x}_2$ los centroides de dichos grupos.

Se trata de encontrar $b^t=(b,b_2...b_m)$ de forma que b^tx sea la óptima combinación lineal de las variables observadas a la hora de discriminar entre los 2 grupos.

Desde un punto de vista intuitivo el óptimo b que podamos encontrar será aquel que maximice la distancia entre las proyecciones sobre el eje definido por b de los centroides de los grupos y minimice la dispersión dentro de cada grupo, donde

$((\bar{x}_1 - \bar{x}_2)^tb)^2$ es la distancia entre las proyecciones de los centroides sobre el eje definido por b.

(*) Vease el programa BMDP3D para la utilización de este test.

$b^t Wb$ es la dispersión dentro de las clases donde W es la matriz de var-cov. obtenida a partir de las muestras en los 2 grupos.

El criterio de óptimo queda reducido a calcular

$$\underset{b}{\text{Max}} \quad \frac{((\bar{x}_1-\bar{x}_2)^t b)^2}{b^t W b}$$

cuya solución es

$$b \text{ es prop. a } W^{-1}(\bar{x}_1-\bar{x}_2)$$

y entre todos los posibles tomamos aquel que tenga norma 1 ($: b^t b=1$).

Si llamamos $F = \frac{(\bar{x}_1+\bar{x}_2)^t}{2} \cdot b$ que es el punto medio entre las proyecciones de los 2 centroides sobre el eje definido por B, la regla de decisión o clasificador que asigna alguno de los dos grupos a nuevas observaciones viene dado por

$$c(x) = \begin{cases} \text{Grupo 1 si } b^t x \leq F \\ \text{Grupo 2 si } b^t x > F \end{cases}$$

la distancia entre las dos clases tomará la forma:

$$d^2(c_1, c_2) = D^2 = (\bar{x}_1 - \bar{x}_2)^t W^{-1} (\bar{x}_1 - \bar{x}_2)$$

que generalizada nos da la distancia de Mahalanobis entre dos puntos cualesquiera

$$d(x,y) = (x-y)^t W^{-1} (x-y)$$

siendo $\frac{n_1 \cdot n_2}{n_1+n_2} D^2 = t^2$, el estadístico T^2 Hottelling, que bajo las condiciones de normalidad y homogeneidad de varianza, en la hipótesis de igualdad de centroides ($\bar{x}_1 = \bar{x}_2$) se cumple:

$$\frac{n-m-1}{m(n-2)} \; T^2 \; : \; F_{m,\, n-m-1}$$

siendo $n = n_1 + n_2$

test que nos sirve para contratar la diferencia entre las clases.

3. Descripción del Programa BMD04M

Este programa calcula un análisis discriminante para 2 grupos teniendo como salida:

1) Medias de cada grupo y diferencia de medias.
2) Matriz de productos cruzados(*).
3) Inversa de la matriz de productos cruzados.
4) Coeficiente de la función discriminante (**).
5) D^2 de Mahalanobis, estadístico F asociado .
6) Medias, desviaciones típicas y varianzas en cada uno de los grupos de las proyecciones de las observaciones sobre el eje discriminante.
7) Valores de las proyecciones de las observaciones sobre el eje discriminante ordenadas de mayor a menor.

Entrada de datos

Si X_{ijk} $\quad i = 1,2$ (grupo)
$\quad j = 1, \ldots n_i$ (observación en el grupo)
$\quad k = 1, \ldots m$ (variable que se observa)

representa a nuestras observaciones, la forma de introducir los datos es:

1º grupo	1º tarjeta	X_{111}	X_{112}	$\cdots$	X_{11m}
	2º "	X_{121}	X_{122}	$\cdots$	X_{12m}
	.				
	.				
	.				
	n_1 "	$X_{1n_1 1}$	$X_{1n_1 2}$		$X_{1n_1 m}$
2º grupo	n_1+1 "	X_{211}	X_{212}		X_{21m}
	.				
	.				
	.				
	n_1+n_2 "	$X_{2n_2 1}$	$X_{2n_2 2}$		$X_{2n_2 m}$

teniendo que tener todas las tarjetas el mismo formato.

(*) que sería $(n_1 + n_2 - 2)$ W; siendo W la matriz de var.-cov. de las observaciones a la que hemos hecho referencia en el apartado anterior.

(**) (son los componentes del vector b, que mencionabamos en el apartado anterior).

Este programa mediante la adecuada utilización de la tarjeta SELECT permite hacer distintos análisis discriminantes sobre los grupos depediendo estos de las variables elegidas para realizarlos.

Procedimiento de Cálculo

Si designamos nuestras observaciones por

X_{ijk} $\quad i = 1,2$ (grupo)

$j = 1,2,...n_i$ (replicación por grupo)

$k = 1,2,...m$ (variables)

Paso 1.

Calcula $\bar{X}_i = (\bar{X}_{i.1},... \bar{X}_{i.m}) \quad i = 1,2$
centroide de cada grupo

Paso 2.

$\bar{X}_1 - \bar{X}_2 = (\bar{X}_{1.1} - \bar{X}_{2.1}, ..., \bar{X}_{1.n} - \bar{X}_{2.m})$
diferencia de centroides

Paso 3.

Las matrices S^1, S^2

$$S^i_{1,n} = \sum_{j=1}^{n_i} (X_{ij1} - X_{i.1})(X_{ijn} - X_{i.n})$$

Paso 4.

La Matriz de Productos Cruzados $A = S^1 + S^2$ donde $A = (n_1 + n_{2-2})$ W siendo W la matriz de varianza covarianza obtenida a partir de los 2 grupos.

Paso 5.

Cálculo de A^{-1}

Paso 6.

Coeficientes de la función discriminante

$(b,...., b_m) = (\bar{X}_1 - \bar{X}_2)^t A^{-1}$

Paso 7.

$$D^2 = (n_1+n_2-2)\ (\bar{X}_1-\bar{X}_2)^t\ A^{-1}\ (\bar{X}_1-\bar{X}_2)$$

Paso 8.

Estadístico F asociado a D^2

$$F_{(m,n_1+n_2-1-m)}: \frac{n_1 n_2\ (n_1+n_2-m-1)}{m(n_1+n_2)\ (n_1+n_2-2)}\ D^2$$

Paso 9.

medias, desviación típica y varianza de los valores de las proyecciones de las observaciones sobre el eje discriminante

$$Z_{ix} = b_1\ X_{ix1} + \ldots + b_m\ X_{ixm}$$

Paso 10.

Proyecciones de las observaciones sobre el eje discriminante ordenadas de mayor a menor etiquetadas con el número de grupo.

Bibliografía

Anderson,R.C.+Bancroft,T.A.:"*Statistical theory in research*" M'Graw Hill,1955.

Morrison: "*Multivariate statistical methods*" M'Graw Hill,1967.

Green Paul E.: "*Analyzing multivariate DATA*" The Dryden Press, 1978.

BMD.: "*Biomedical Computer Programs*" University of California Press, 1975.