Tema 6: Análisis Discriminante

Introducción

Supongamos que un conjunto de objetos está ya clasificado en una serie de grupos, es decir, se sabe previamente a qué grupos pertenecen. El Análisis Discriminante se puede considerar como un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos, y.las variables independientes son continuas y determinan a qué grupos pertenecen los objetos. Se pretende encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos.

Un segundo objetivo es construir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados con un cierto grado de riesgo.

Es necesario considerar una serie de restricciones o supuestos:

Se tiene una variable categórica y el resto de variables son de intervalo o de razón y son independientes respecto de ella.

Es necesario que existan al menos dos grupos, y para cada grupo se necesitan dos o más casos.

El número de variables discriminantes debe ser menor que el número de objetos menos 2: $x_1, ..., x_p$, donde $p < (n-2)$ y n es el número de objetos.

Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.

El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 (con q grupos, $(q-1)$ funciones discriminantes).

Las matrices de covarianzas dentro de cada grupo deben ser aproximadamente iguales.

Las variables continuas deben seguir una distribución normal multivariante.

Modelo matemático

A partir de q grupos donde se asignan a una serie de objetos y de p variables medidas sobre ellos $(x_1, \ldots, x_p)$, se trata de obtener para cada objeto una serie de *puntuaciones* que indican el grupo al que pertenecen $(y_1, \ldots, y_m)$, de modo que sean funciones lineales de $x_1, \ldots, x_p$

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1p}x_p + a_{10} \\ &\cdots\cdots\cdots \\ y_m &= a_{m1}x_1 + \cdots + a_{mp}x_p + a_{m0} \end{aligned}$$

donde $m = \text{mín}(q-1, p)$, tales que discriminen o separen lo máximo posible a los q grupos. Estas combinaciones lineales de las p variables deben maximizar la varianza *entre* los grupos y minimizar la varianza *dentro* de los grupos.

Descomposición de la varianza

Se puede descomponer la variabilidad total de la muestra en variabilidad dentro de los grupos y entre los grupos.

Partimos de

$$Cov\left(x_j, x_{j'}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ij'} - \bar{x}_{j'}\right)$$

Se puede considerar la media de la variable x_j en cada uno de los grupos $I_1, \ldots, I_q$, es

decir,

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} x_{ij}$$

para $k = 1, \ldots, q$.

De este modo, la media total de la variable x_j se puede expresar como función de las medias dentro de cada grupo. Así,

$$\sum_{i \in I_k} x_{ij} = n_k \bar{x}_{kj},$$

entonces

$$\begin{aligned} \bar{x}_j &= \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} x_{ij} = \\ &= \frac{1}{n} \sum_{k=1}^{q} n_k \bar{x}_{kj} = \sum_{k=1}^{q} \frac{n_k}{n} \bar{x}_{kj}. \end{aligned}$$

Así,

$$Cov\,(x_j, x_{j'}) = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (x_{ij} - \bar{x}_j)\,(x_{ij'} - \bar{x}_{j'})$$

Si en cada uno de los términos se pone:

$$\begin{aligned} (x_{ij} - \bar{x}_j) &= (x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j) \\ (x_{ij'} - \bar{x}_{j'}) &= (x_{ij'} - \bar{x}_{kj'}) + (\bar{x}_{kj'} - \bar{x}_{j'}) \end{aligned}$$

al simplificar se obtiene:

$$\begin{aligned} Cov\,(x_j, x_{j'}) &= \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (x_{ij} - \bar{x}_{kj})\,(x_{ij'} - \bar{x}_{kj'}) + \sum_{k=1}^{q} \frac{n_k}{n} (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj'} - \bar{x}_{j'}) = \\ &= d(x_j, x_{j'}) + e(x_j, x_{j'}). \end{aligned}$$

Es decir, la covarianza *total* es igual a la covarianza *dentro* de grupos más la covarianza *entre* grupos. Si denominamos como $t(x_j, x_{j'})$ a la covarianza total entre x_j y $x_{j'}$ (sin distinguir grupos), entonces lo anterior se puede expresar como

$$t(x_j, x_{j'}) = d(x_j, x_{j'}) + e(x_j, x_{j'}).$$

En notación matricial esto es equivalente a

$$T = E + D$$

donde:

T = matriz de covarianzas total

E = matriz de covarianzas entre grupos

D = matriz de covarianzas dentro de grupos.

Extracción de las funciones discriminantes

La idea básica del Análisis Discriminante consiste en extraer a partir de $x_1, \ldots, x_p$ variables observadas en k grupos, m funciones $y_1, \ldots, y_m$ de forma

$$y_i = a_{i1}x_1 + \cdots + a_{ip}x_p + a_{i0}$$

donde $m = \text{mín}(q-1, p)$, tales que $corr(y_i, y_j) = 0$ para todo $i \neq j$.

Si las variables $x_1, \ldots, x_p$ están tipificadas, entonces las funciones

$$y_i = a_{i1}x_1 + \cdots + a_{ip}x_p$$

para $i = 1, \ldots, m$, se denominan funciones *discriminantes canónicas*.

Las funciones $y_1, \ldots, y_m$ se extraen de modo que

(i) y_1 sea la combinación lineal de $x_1, \ldots, x_p$ que proporciona la mayor discriminación posible entre los grupos.

(ii) y_2 sea la combinación lineal de $x_1, \ldots, x_p$ que proporciona la mayor discriminación posible entre los grupos, después de y_1, tal que $corr(y_i, y_2) = 0$.

En general, y_i es la combinación lineal de $x_1, \ldots, x_p$ que proporciona la mayor discriminación posible entre los grupos después de y_{i-1} y tal que $corr(y_i, y_j) = 0$ para $j = 1, \ldots, (i-1)$.

Procedimiento matricial

Se sigue un método parecido al análisis factorial, así se busca una función lineal de $x_1, \ldots, x_p$: $y = a'x$, de modo que

$$Var(y) = a'Ta = a'Ea + a'Da$$

es decir, la variabilidad entre grupos más la variabilidad dentro de grupos.

Queremos maximizar la variabilidad entre los grupos para discriminarlos mejor y esto equivale a hacer

$$\text{máx}\left(\frac{a'Ea}{a'Ta}\right),$$

es decir, maximizar la varianza entre grupos en relación al total de la varianza.

Si consideramos la función

$$f(a) = \frac{a'Ea}{a'Ta}$$

Se observa que f es una función *homogénea*, es decir,

$$f(a) = f(\mu a)$$

para todo $\mu \in \mathbb{R}$.

El hecho de que la función sea homogénea implica que calcular máx $\left(\frac{a'Ea}{a'Ta}\right)$ equivale a calcular

$$\text{máx}\,(a'Ea)$$

tal que

$$a'Ta = 1$$

Como este es el esquema habitual de los multiplicadores de Lagrange, se define

$$L = a'Ea - \lambda(a'Ta - 1)$$

y se calcula su derivada:

$$\frac{\partial L}{\partial a} = 0.$$

$$\begin{aligned} \frac{\partial L}{\partial a} &= 2Ea - 2\lambda Ta = 0 \Rightarrow \\ Ea &= \lambda Ta \Rightarrow \\ (T^{-1}E)a &= \lambda a \end{aligned}$$

Por tanto, el autovector asociado a la primera función discriminante lo es de la matriz $T^{-1}E$ (que no es simétrica en general).

Como $Ea = \lambda Ta$,

$$a'Ea = \lambda a'Ta = \lambda$$

Luego si tomo el vector asociado al máximo autovalor, se obtendrá la función que recoge el máximo *poder discriminante*.

El autovalor asociado a la función discriminante indica la proporción de varianza total explicada por las m funciones discriminantes que recoge la variable y_i.

Para obtener más funciones discriminantes, se siguen sacando los autovectores de la matriz $(T^{-1}E)$ asociados a los autovalores elegidos en orden decreciente:

$$\begin{array}{lcl} a_2' & \Rightarrow & a_2'x = y_2 \\ \vdots & \dots & \vdots \\ a_m' & \Rightarrow & a_m'x = y_m \end{array}$$

donde $m = \text{mín}(q-1, p)$

Estos vectores son linealmente independientes y dan lugar a funciones incorreladas entre sí.

La suma de todos los autovalores, $\sum_{i=1}^{m} \lambda_i$, es la proporción de varianza total que queda explicada, o se conserva, al considerar sólo los ejes o funciones discriminantes. Como consecuencia, el porcentaje explicado por y_i del total de varianza explicada por $y_1, \dots, y_m$ es

$$\frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} \cdot 100\,\%$$

Análisis Discriminante con SPSS

Cuando se utiliza SPSS se suelen considerar varias fases en el análisis discriminante.

Comprobación de los supuestos paramétricos del análisis discriminante

En sentido estricto, la función discriminante minimiza la probabilidad de equivocarse al clasificar los individuos en cada grupo. Para ello, las variables originales se deben distribuir como una normal multivariante y las matrices de covarianzas deben ser iguales en todos los grupos. En la práctica es una técnica robusta y funciona bien aunque las dos restricciones anteriores no se cumplan.

Si un conjunto de variables se distribuye como una normal multivariante, entonces cualquier combinación lineal de ellas se distribuye como una normal univariante. Por ello, si alguna de las variables originales no se distribuye como una normal, entonces es seguro que todas las variables conjuntamente no se distribuirán como una normal multivariante.

La segunda restricción se refiere a la igualdad entre las matrices de covarianzas de los grupos. Para comprobar esto, se puede usar la prueba *M de Box*, que está incluida en el SPSS. Dicha prueba tiene como hipótesis nula que las matrices de covarianzas son iguales. Se basa en el cálculo de los determinantes de las matrices de covarianzas de cada grupo. El valor obtenido se aproxima por una F de Snedecor. Si el p-valor es menor que 0,05 se rechaza la igualdad entre las matrices de covarianzas.

El test *M de Box* es sensible a la falta de normalidad multivariante, es decir, matrices iguales pueden aparecer como significativamente diferentes si no existe normalidad. Por otra parte, si las muestras son grandes, pierde efectividad (es más fácil rechazar la hipótesis nula).

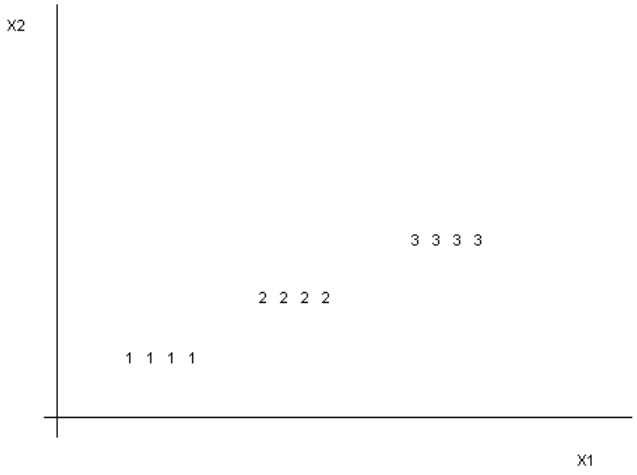
Selección de las variables discriminantes

Primero se puede realizar un análisis descriptivo univariante calculando las medias y las desviaciones estándar de las variables originales para cada uno de los grupos por

separado. Si para alguna variable las medias de los grupos son diferentes y la variabilidad es pequeña, se considera que dicha variable será importante a la hora de discriminar a los grupos.

A continuación, se observan las relaciones entre las variables. Se calculan matrices de correlaciones en lugar de matrices de covarianzas por ser más fácilmente interpretables. Además de analizar la correlación entre pares de variables sin distinguir grupos, se debe analizar las correlaciones dentro de cada grupo y luego considerar la media de las mismas

Se calcula también la matriz *Pooled within-groups correlation matrix*. Dicha matriz se calcula como una matriz media de correlaciones calculadas por separado en cada grupo. A menudo no se parece a la matriz de correlaciones *total*. Veamos, por ejemplo, el siguiente gráfico de dos variables y tres grupos:



Si se considera cada grupo por separado (1, 2 y 3), el coeficiente de correlación entre x_1 y x_2 es 0 (el hecho de variar x_1 no influye en x_2: la pendiente de la recta de regresión es 0). Si hallamos la media de esos coeficientes, su valor es también 0; sin embargo, el coeficiente de correlación calculado para todos los datos sin tener en cuenta a los grupos está próximo a 1, porque cuando aumenta el valor de x_1 también lo hace el valor de x_2.

Estadísticos usados

***F* de Snedecor** Se compara para cada variable las desviaciones de las medias de cada uno de los grupos a la media total, entre las desviaciones a la media *dentro* de cada grupo. Si F es grande para cada variable, entonces las medias de cada grupo están muy separadas y la variable discrimina bien. Si F es pequeña, la variable discriminará poco, ya que habrá poca homogeneidad en los grupos y éstos estarán muy próximos.

λ de Wilks También se la denomina U-estadístico. Cuando se considera a las variables de modo individual, la λ es igual al cociente entre la suma de cuadrados *dentro* de los grupos y la suma de cuadrados *total* (sin distinguir grupos). Es decir, equivale a las desviaciones a la media dentro de cada grupo, entre las desviaciones a la media total sin distinguir grupos. Si su valor es pequeño, la variable discrimina mucho: la variabilidad total se debe a las diferencias entre grupos, no a las diferencias dentro de grupos.

Variables originales que se consideran

La idea del Análisis discriminante es construir funciones lineales de las variables originales que discriminen entre los distintos grupos. Sin embargo, no todas las variables discriminan de la misma forma o tienen los mismos valores de la F de Snedecor o de la λ de Wilks. Por ello, a la hora de construir las funciones lineales, no es necesario incluir a todas las variables iniciales en la función.

Como criterio general para seleccionar una variable se emplea la selección del valor de la λ de Wilks o, de modo equivalente, del valor de su F asociada.

Se usan fundamentalmente dos métodos de selección de variables: el método *directo* y el método *stepwise*.

En el método *directo* se consideran todas las variables originales que verifiquen un criterio de selección.

El método *stepwise* es un método que funciona con varios pasos:

(*i*) Se incluye en el análisis la variable que tenga el mayor valor aceptable para el criterio de selección o de *entrada*.

(*ii*) Se evalúa el criterio de selección para las variables no seleccionadas. La variable que presenta el valor más alto para el criterio se selecciona (siempre que esté dentro de un límite).

(*iii*) Se examinan las variables seleccionadas según un criterio de *salida* y se examinan también las variables no seleccionadas, para ver si cumplen el criterio de entrada. Se excluyen o se incluyen variables según cumplan los criterios de entrada y de salida.

(*iv*) Se repite el paso (*iii*) hasta que ninguna variable más pueda ser seleccionada o eliminada.

Además de todo lo anterior, en el SPSS se considera un número máximo de pasos, dado que una variable puede ser incluida y eliminada en más de una ocasión. Se toma el doble del número de variables originales como número máximo de pasos del método stepwise.

En el SPSS se considera también para cada variable la tolerancia asociada.

Tolerancia

Se define para un conjunto de p variables, R_i, el coeficiente de correlación múltiple que expresa el porcentaje de variabilidad de la variable x_i $(i = 1, \ldots, p)$ recogida por el resto de $(p-1)$ variables. Si se eleva al cuadrado R_i^2 se obtiene el coeficiente de determinación. Entonces, la tolerancia se define como $1 - R_i^2$. Así, cuanto mayor sea la tolerancia de una variable, más información independiente del resto de variables recogerá.

De este modo, si en una iteración dada del procedimiento *stepwise* la variable seleccionada verifica que su tolerancia con respecto a las variables ya incluidas en la función discriminante es muy pequeña entonces la variable no se incluye en dicha etapa. Así, se evita la redundancia de información.

Cálculo de la F y de la λ de Wilks multivariantes para fijar los criterios de entrada y salida

Para un conjunto de variables se define la F como

$$F = \frac{|B|}{|W|}$$

donde

$|B|$ = determinante de la matriz de covarianzas *entre grupos*.

$|W|$ = determinante de la suma de las matrices de covarianzas *dentro de los grupos*.

En general, el determinante de una matriz de covarianzas da una medida de la variabilidad total de un conjunto de variables.

A partir de este valor de F, se puede calcular la correspondiente λ de Wilks, ya que

$$F = \frac{n-k-p-1}{k-1}\left(\frac{1}{\lambda}-1\right)$$

donde

n = número de observaciones

k = número de grupos

p = número de variables

La F y la λ de Wilks se interpretan del mismo modo que en el caso univariante. Cuando se comparan covarianzas entre grupos, se hace en base a los centroides de los grupos, es decir, a los vectores de medias de las variables en cada grupo.

Estadísticos que se calculan en el procedimiento stepwise

F de entrada (*F* to enter):

Expresa la disminución en la λ de Wilks que se produce si se incluye una variable dada entre las que no están dentro de la función discriminante. Si el valor es pequeño, la disminución de la λ de Wilks será inapreciable y la variable no entrará en la función.

F de salida (*F* to remove):

Expresa el incremento que se produce en la λ de Wilks, si se elimina de la función discriminante una variable dada. Si el valor de la F de salida es pequeño, el incremento no será significativo y la variable se eliminará del análisis.

Correlación Canónica

Da una medida del grado de asociación entre las puntuaciones discriminantes de cada uno de los objetos y el grupo concreto de pertenencia:

$$\eta^2 = \frac{SC_{ENTRE}}{SC_{TOTAL}},$$

es decir, es la proporción de la variabilidad total debida a la diferencia entre grupos para las funciones discriminantes.

Cuando sólo se tienen dos grupos, la correlación canónica es igual al coeficiente de correlación entre la puntuación discriminante y el grupo de pertenencia, que se representa por una variable codificada en 0–1 (en SPSS).

Significación y coeficientes de las funciones discriminantes

Cuando no existen diferencias entre los grupos, las funciones discriminantes sólo indican variabilidad aleatoria (*ruido*). Se puede usar la λ de Wilks para realizar un test en el cual la hipótesis nula es que las medias de las funciones discriminantes en cada grupo son iguales.

Cuando se tienen varios grupos y varias funciones, se calcula una λ de Wilks total mediante el producto de las λ de Wilks de cada función. Ésta se puede aproximar por una χ^2, usando la siguiente transformación:

$$V = -\left(n - 1 - \frac{p+k}{2}\right)\ln(\lambda)$$

de modo que $V \sim \chi^2_{p(k-1)}$ aproximadamente. De este modo, si λ es pequeño V es grande y se rechaza la hipótesis nula.

Si la significación asociada al valor de la χ^2 es menor que 0,05 (o bien otro valor prefijado) se rechaza la hipótesis nula (a dicho nivel de confianza).

Interpretación de los coeficientes de la función discriminante

Si usamos variables originales tipificadas, se obtienen los coeficientes a_{ij} que relacionan las variables con las funciones discriminantes:

	y_1	$\cdots$	y_m
x_1	a_{11}	$\cdots$	a_{m1}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_p	a_{1p}	$\cdots$	a_{mp}

Se pueden interpretar las magnitudes de los coeficientes como indicadores de la importancia relativa de las variables en cada función discriminante. Así, si a_{ij} es grande en valor absoluto, entonces hay una fuerte asociación entre la variable x_j y la función y_i, en relación al resto de variables. Aún así, al existir en general correlaciones significativas entre las variables originales, se debe tener cuidado al hacer interpretaciones precipitadas.

Matriz de estructura

Otra forma de calcular la contribución de cada variable a una función discriminante es examinar las correlaciones entre los valores de la función y los valores de las variables. Se calculan, dentro de cada grupo, las correlaciones entre las variables y las puntuaciones; luego se combinan en una matriz *pooled within-groups correlation matrix*. Los valores obtenidos dan una medida de las contribuciones.

Clasificación de los objetos

Una vez calculadas las funciones discriminantes, es decir, las combinaciones lineales de las variables originales, a cada objeto se le puede asignar una puntuación o valor dado en la función discriminante.

Esto equivale al valor que se recoge en una ecuación de regresión. Así, si x_{ij} es el valor que alcanza el objeto i-ésimo en la variable j-ésima, entonces la *puntuación* o valor alcanzado en la función discriminante k será:

$$y_{ik} = a_{k1}x_{i1} + \cdots + a_{kp}x_{ip} + a_{k0}$$

Regla de Bayes

Se pueden usar las puntuaciones discriminantes para obtener una regla para clasificar los casos en los grupos. En el SPSS se usa la regla de Bayes.

Así, la probabilidad de que un objeto j, con una puntuación discriminante $D = (y_{j1}, ..., y_{jm})$, pertenezca al grupo i-ésimo se puede estimar mediante la regla de Bayes:

$$P\left(G_i|D\right) = \frac{P\left(D|G_i\right)P(G_i)}{\sum_{i=1}^{k} P\left(D|G_i\right)P(G_i)}$$

$P(G_i)$ es la probabilidad *a priori* y es una estima de la *confianza* de que un objeto pertenezca a un grupo si no se tiene información previa. Por ejemplo, si 70 de 113 personas sobreviven en la muestra, la probabilidad de sobrevivir se aproxima por 70/113.

Las probabilidades a priori se pueden determinar de distintos modos. Si la muestra se considera representativa de la población, se pueden usar las proporciones de los casos en cada grupo como estimadores de dichas probabilidades. Cuando todos los grupos tienen el mismo número de objetos y no se tiene ningún tipo de información previa, se asignan probabilidades a priori iguales para todos los grupos.

$P(D|G_i)$ es la probabilidad de obtener la puntuación D estando en el grupo i-ésimo. Como las puntuaciones discriminantes se calculan a partir de combinaciones lineales de p variables, distribuidas según una normal, se distribuyen a su vez como una normal, cuya media y varianza se estiman a partir de todas las puntuaciones que se recogen en el grupo i-ésimo.

$P(G_i|D)$ es la probabilidad a posteriori que se estima a través de $P(G_i)$ y de $P(D|G_i)$. En realidad, mide lo mismo que la $P(G_i)$, pero refina la medida de incertidumbre al tener

en cuenta la información que recogen las puntuaciones discriminantes D. Es decir, lo que interesa es calcular la probabilidad de que un objeto pertenezca al grupo G_i, dado que presenta la puntuación D.

Se asignará un objeto al grupo G_i cuya probabilidad a posteriori sea máxima, es decir, dado que presenta la puntuación D.

Matriz de confusión

Da una idea de la tasa de clasificaciones incorrectas. Como se sabe el grupo al que pertenece cada objeto, se puede comprobar la efectividad del método de clasificación usando la máxima probabilidad a posteriori, cuando se observa el porcentaje de casos bien clasificados. No obstante, se tiene que tener en cuenta también la tasa de clasificaciones incorrectas esperadas según las probabilidades a priori.

Ejemplos

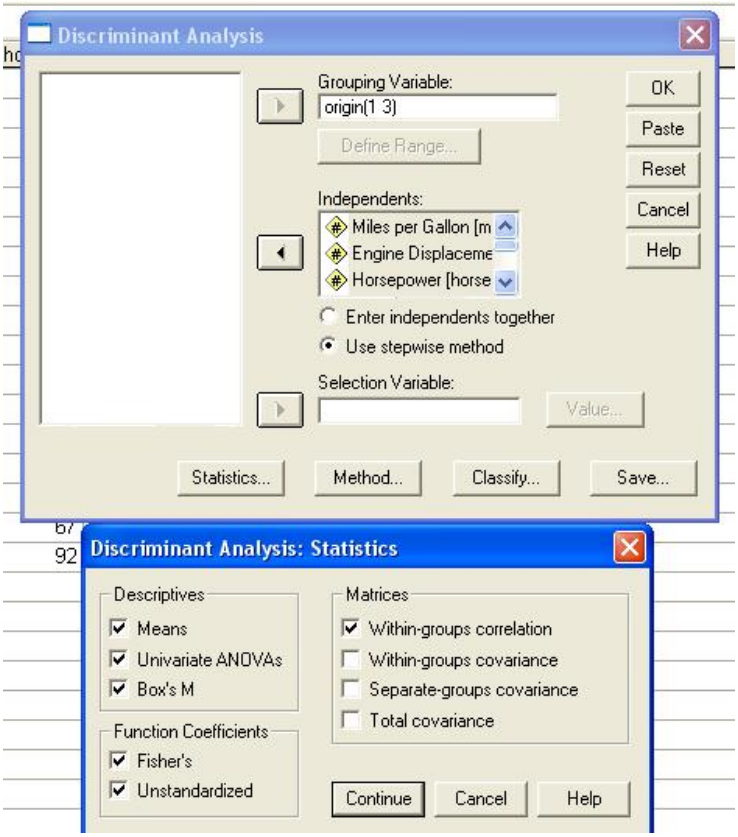
Se consideran los datos recogidos sobre 32 cráneos en el Tibet.

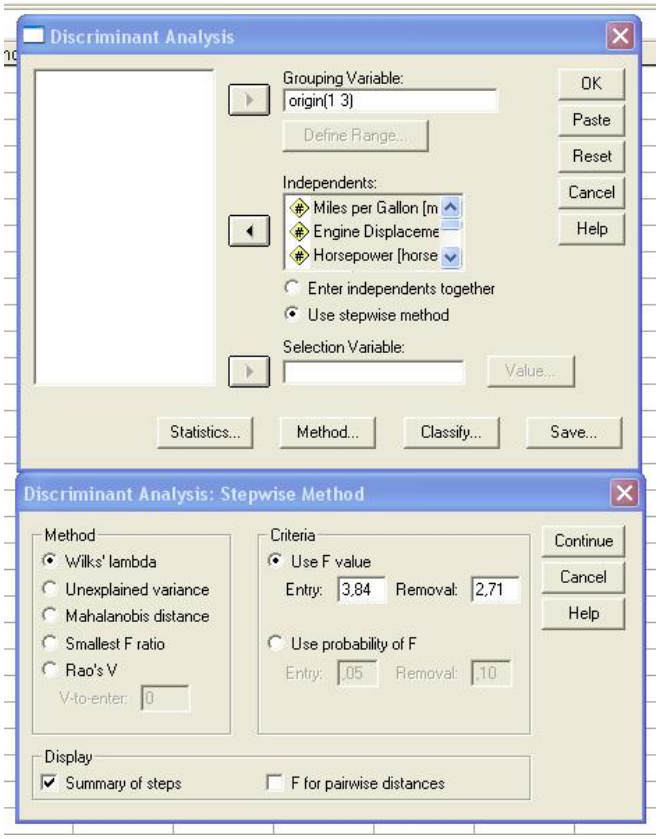
	Longitud	Anchura	Altura	Altura.Cara	.Anchura.Cara	Tipo
1	190.50	152.50	145.00	73.50	136.50	1
2	172.50	132.00	125.50	63.00	121.00	1
3	167.00	130.00	125.50	69.50	119.50	1
4	169.50	150.50	133.50	64.50	128.00	1
5	175.00	138.50	126.00	77.50	135.50	1
6	177.50	142.50	142.50	71.50	131.00	1
7	179.50	142.50	127.50	70.50	134.50	1
8	179.50	138.00	133.50	73.50	132.50	1
9	173.50	135.50	130.50	70.00	133.50	1
10	162.50	139.00	131.00	62.00	126.00	1
11	178.50	135.00	136.00	71.00	124.00	1
12	171.50	148.50	132.50	65.00	146.50	1
13	180.50	139.00	132.00	74.50	134.50	1
14	183.00	149.00	121.50	76.50	142.00	1
15	169.50	130.00	131.00	68.00	119.00	1
16	172.00	140.00	136.00	70.50	133.50	1
17	170.00	126.50	134.50	66.00	118.50	1
18	182.50	136.00	138.50	76.00	134.00	2
19	179.50	135.00	128.50	74.00	132.00	2
20	191.00	140.50	140.50	72.50	131.50	2
21	184.50	141.50	134.50	76.50	141.50	2
22	181.00	142.00	132.50	79.00	136.50	2
23	173.50	136.50	126.00	71.50	136.50	2
24	188.50	130.00	143.00	79.50	136.00	2
25	175.00	153.00	130.00	76.50	142.00	2
26	196.00	142.50	123.50	76.00	134.00	2
27	200.00	139.50	143.50	82.50	146.00	2
28	185.00	134.50	140.00	81.50	137.00	2
29	174.50	143.50	132.50	74.00	136.50	2
30	195.50	144.00	138.50	78.50	144.00	2
31	197.00	131.50	135.00	80.50	139.00	2
32	182.50	131.00	135.00	68.50	136.00	2

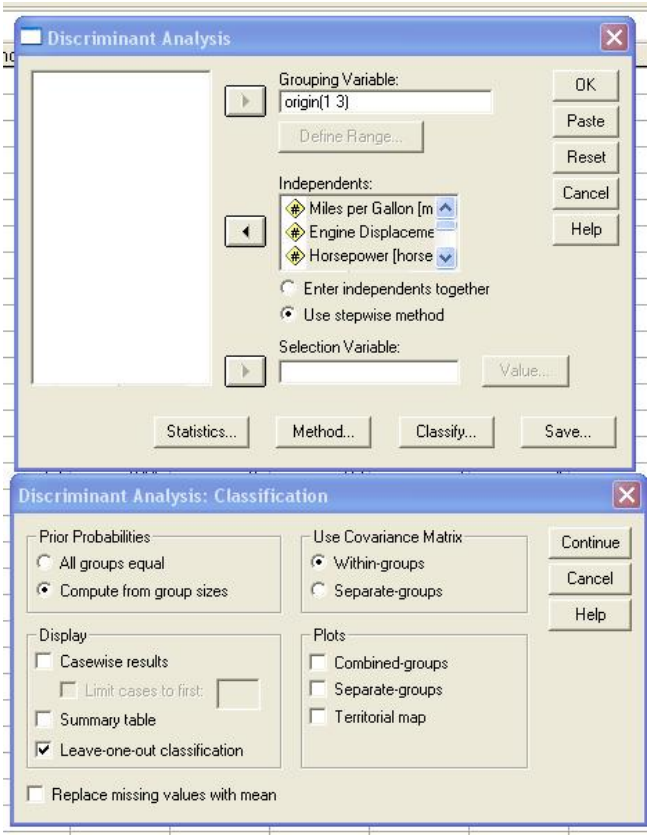
Los datos corresponden a dos tipos raciales diferentes en los que se practicaron diferentes medidas antropométricas de longitudes, anchuras de cráneo y de cara. Se trata de hacer un análisis discriminante sobre los dos tipos raciales.

Se toma una muestra de 50 vehículos producidos en EE.UU, Japón y Europa. Se consideran las siguientes variables: *Consumo*, *Cilindrada*, *Potencia*, *Peso*, *Aceleración*, *Año del modelo* y *Número de cilindros*. Se trata de hacer un análisis discriminante sobre los tres tipos de vehículos, en función de su origen.

Análisis Discriminante (con SPSS)






Discriminant Analysis
Grouping Variable:
origin(1 3)
Define Range...
OK
Paste
Reset
Cancel
Help
Independents:
Miles per Gallon [m
Engine Displaceme
Horsepower [horse
Enter independents together
Use stepwise method
Selection Variable:
Value...
Statistics...
Method...
Classify...
Save...
Discriminant Analysis: Classification
Prior Probabilities
All groups equal
Compute from group sizes
Use Covariance Matrix
Within-groups
Separate-groups
Continue
Cancel
Help
Display
Casewise results
Limit cases to first:
Summary table
Leave-one-out classification
Plots
Combined-groups
Separate-groups
Territorial map
Replace missing values with mean

Media, desviación típica, número de casos válidos (ponderado y no ponderado) para cada uno de los grupos y para la muestra total:

Group Statistics					
Country of Origin		**Mean**	**Std. Deviation**	**Valid N (listwise)**	
				Unweighted	**Weighted**
American	**Miles per Gallon**	19,92	7,236	25	25,000
	Engine Displacement (cu. inches)	245,44	94,885	25	25,000
	Horsepower	123,56	44,563	25	25,000
	Vehicle Weight (lbs.)	3368,28	799,303	25	25,000
	Time to Accelerate from 0 to 60 mph (sec)	14,85	2,311	25	25,000
	Model Year (modulo 100)	75,16	3,496	25	25,000
	Number of Cylinders	6,24	1,763	25	25,000
European	**Miles per Gallon**	28,92	6,345	9	9,000
	Engine Displacement (cu. inches)	105,56	21,190	9	9,000
	Horsepower	76,56	18,882	9	9,000
	Vehicle Weight (lbs.)	2341,44	395,406	9	9,000
	Time to Accelerate from 0 to 60 mph (sec)	16,78	3,081	9	9,000
	Model Year (modulo 100)	74,67	3,464	9	9,000
	Number of Cylinders	4,00	,000	9	9,000
Japanese	**Miles per Gallon**	30,64	6,966	16	16,000
	Engine Displacement (cu. inches)	106,50	30,124	16	16,000
	Horsepower	83,81	22,489	16	16,000
	Vehicle Weight (lbs.)	2288,94	388,479	16	16,000
	Time to Accelerate from 0 to 60 mph (sec)	15,23	2,058	16	16,000
	Model Year (modulo 100)	78,38	2,941	16	16,000
	Number of Cylinders	4,13	,806	16	16,000
Total	**Miles per Gallon**	24,97	8,572	50	50,000
	Engine Displacement (cu. inches)	175,80	98,537	50	50,000
	Horsepower	102,38	40,616	50	50,000
	Vehicle Weight (lbs.)	2838,06	819,660	50	50,000
	Time to Accelerate from 0 to 60 mph (sec)	15,32	2,443	50	50,000
	Model Year (modulo 100)	76,10	3,621	50	50,000
	Number of Cylinders	5,16	1,707	50	50,000

Tabla de ANOVA con estadísticos F que permiten contrastar la hipótesis de igualdad de medias entre los grupos en cada variable independiente. La tabla de ANOVA incluye también el estadístico lambda de Wilks univariante. La información de esta tabla suele utilizarse como prueba preliminar para detectar si los grupos difieren en las variables de clasificación seleccionadas; sin embargo, debe tenerse en cuenta que una variable no significativa a nivel univariante podría aportar información discriminativa a nivel multivariante.

Tests of Equality of Group Means

	Wilks' Lambda	F	Df1	df2	Sig.
Miles per Gallon	,641	13,186	2	47	,000
Engine Displacement (cu. inches)	,490	24,428	2	47	,000
Horsepower	,719	9,195	2	47	,000
Vehicle Weight (lbs.)	,573	17,546	2	47	,000
Time to Accelerate from 0 to 60 mph (sec)	,915	2,180	2	47	,124
Model Year (modulo 100)	,808	5,586	2	47	,007
Number of Cylinders	,591	16,281	2	47	,000

Correlación intra-grupos. Muestra la matriz de correlaciones intra-grupo combinada, es decir la matriz de correlaciones entre las variables independientes estimada a partir de las correlaciones obtenidas dentro de cada grupo.

Pooled Within-Groups Matrices

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)	Model Year (modulo 100)	Number of Cylinders
Correlation	**Miles per Gallon**	1,000	-,664	-,693	-,719	,421	,722	-,571
	Engine Displacement (cu. inches)	-,664	1,000	,851	,788	-,520	-,442	,914
	Horsepower	-,693	,851	1,000	,725	-,660	-,546	,740
	Vehicle Weight (lbs.)	-,719	,788	,725	1,000	-,302	-,363	,766
	Time to Accelerate from 0 to 60 mph (sec)	,421	-,520	-,660	-,302	1,000	,354	-,484
	Model Year (modulo 100)	,722	-,442	-,546	-,363	,354	1,000	-,357
	Number of Cylinders	-,571	,914	,740	,766	-,484	-,357	1,000

Box's Test of Equality of Covariance Matrices

Log Determinants		
Country of Origin	**Rank**	**Log Determinant**
American	3	16,939
European	3	13,649
Japanese	3	14,181
Pooled within-groups	3	16,386
The ranks and natural logarithms of determinants printed are those of the group covariance matrices.		

Test Results		
Box's M		41,689
F	**Approx.**	3,061
	df1	12
	df2	3043,281
	Sig.	,000
Tests null hypothesis of equal population covariance matrices.		

Stepwise Statistics

Variables Entered/Removed(a,b,c,d)									
Step	**Entered**	**Wilks' Lambda**							
		Statistic	**df1**	**df2**	**Df3**	**Exact F**			
						Statistic	**df1**	**df2**	**Sig.**
1	Engine Displacement (cu. inches)	,490	1	2	47,000	24,428	2	47,000	,000
2	Model Year (modulo 100)	,406	2	2	47,000	13,083	4	92,000	,000
3	Horsepower	,344	3	2	47,000	10,569	6	90,000	,000
At each step, the variable that minimizes the overall Wilks' Lambda is entered.									
a Maximum number of steps is 14.									
b Minimum partial F to enter is 3.84.									
c Maximum partial F to remove is 2.71.									
d F level, tolerance, or VIN insufficient for further computation.									

Variables in the Analysis				
Step		**Tolerance**	**F to Remove**	**Wilks' Lambda**
1	**Engine Displacement (cu. inches)**	1,000	24,428	
2	**Engine Displacement (cu. inches)**	,804	22,737	,808
	Model Year (modulo 100)	,804	4,756	,490
3	**Engine Displacement (cu. inches)**	,275	14,713	,569
	Model Year (modulo 100)	,701	5,981	,436
	Horsepower	,240	4,063	,406

Variables Not in the Analysis					
Step		**Tolerance**	**Min. Tolerance**	**F to Enter**	**Wilks' Lambda**
0	**Miles per Gallon**	1,000	1,000	13,186	,641
	Engine Displacement (cu. inches)	1,000	1,000	24,428	,490
	Horsepower	1,000	1,000	9,195	,719
	Vehicle Weight (lbs.)	1,000	1,000	17,546	,573
	Time to Accelerate from 0 to 60 mph (sec)	1,000	1,000	2,180	,915
	Model Year (modulo 100)	1,000	1,000	5,586	,808
	Number of Cylinders	1,000	1,000	16,281	,591
1	**Miles per Gallon**	,559	,559	,419	,482
	Horsepower	,275	,275	2,887	,436
	Vehicle Weight (lbs.)	,379	,379	,174	,487
	Time to Accelerate from 0 to 60 mph (sec)	,730	,730	3,246	,430
	Model Year (modulo 100)	,804	,804	4,756	,406
	Number of Cylinders	,165	,165	,796	,474
2	**Miles per Gallon**	,331	,331	1,496	,381
	Horsepower	,240	,240	4,063	,344
	Vehicle Weight (lbs.)	,379	,351	,154	,404
	Time to Accelerate from 0 to 60 mph (sec)	,711	,654	3,746	,348
	Number of Cylinders	,162	,150	,810	,392
3	**Miles per Gallon**	,325	,235	1,557	,321
	Vehicle Weight (lbs.)	,368	,214	,457	,337
	Time to Accelerate from 0 to 60 mph (sec)	,557	,188	1,101	,328
	Number of Cylinders	,159	,097	1,142	,327

Wilks' Lambda									
Step	**Number of Variables**	**Lambda**	**df1**	**df2**	**df3**	**Exact F**			
						Statistic	**df1**	**df2**	**Sig.**
1	1	,490	1	2	47	24,428	2	47,000	,000
2	2	,406	2	2	47	13,083	4	92,000	,000
3	3	,344	3	2	47	10,569	6	90,000	,000

Summary of Canonical Discriminant Functions

Eigenvalues				
Function	**Eigenvalue**	**% of Variance**	**Cumulative %**	**Canonical Correlation**
1	1,263(a)	81,6	81,6	,747
2	,284(a)	18,4	100,0	,470
a First 2 canonical discriminant functions were used in the analysis.				

Wilks' Lambda				
Test of Function(s)	**Wilks' Lambda**	**Chi-square**	**df**	**Sig.**
1 through 2	,344	49,067	6	,000
2	,779	11,495	2	,003

Standardized Canonical Discriminant Function Coefficients		
	Function	
	1	**2**
Engine Displacement (cu. inches)	1,595	-,304
Horsepower	-,819	1,091
Model Year (modulo 100)	-,019	1,164

Structure Matrix		
	Function	
	1	**2**
Engine Displacement (cu. inches)	,906(*)	,110
Number of Cylinders(a)	,858(*)	,114
Vehicle Weight (lbs.)(a)	,669(*)	,129
Horsepower	,549(*)	,197
Miles per Gallon(a)	-,505(*)	,286
Time to Accelerate from 0 to 60 mph (sec)(a)	-,294(*)	-,150
Model Year (modulo 100)	-,278	,703(*)
Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.		
* Largest absolute correlation between each variable and any discriminant function		
a This variable not used in the analysis.		

Coeficientes de clasificación de Fisher. Pueden utilizarse directamente para la clasificación. Se obtiene un conjunto de coeficientes para cada grupo, y se asigna un caso al grupo para el que tiene una mayor puntuación discriminante.

Canonical Discriminant Function Coefficients		
	Function	
	1	**2**
Engine Displacement (cu. inches)	,023	-,004
Horsepower	-,023	,031
Model Year (modulo 100)	-,006	,350
(Constant)	-1,150	-29,070
Unstandardized coefficients		

Functions at Group Centroids		
Country of Origin	**Function**	
	1	**2**
American	1,088	,027
European	-,980	-1,000
Japanese	-1,149	,520
Unstandardized canonical discriminant functions evaluated at group means		

Classification Statistics

Prior Probabilities for Groups			
Country of Origin	**Prior**	**Cases Used in Analysis**	
		Unweighted	**Weighted**
American	,500	25	25,000
European	,180	9	9,000
Japanese	,320	16	16,000
Total	1,000	50	50,000

Classification Function Coefficients			
	Country of Origin		
	American	**European**	**Japanese**
Engine Displacement (cu. Inches)	-,015	-,057	-,067
Horsepower	,668	,684	,735
Model Year (modulo 100)	10,521	10,173	10,707
(Constant)	-435,516	-404,685	-447,914
Fisher's linear discriminant functions			

Validación cruzada: para comprobar la capacidad predictiva de la función discriminante,. para ello el SPSS genera tantas funciones discriminantes como casos válidos tiene el análisis; cada una de esas funciones se obtiene eliminando un caso; después, cada caso es clasificado utilizando la función discriminante en la que no ha intervenido.

Classification Results(b,c)

		Country of Origin	Predicted Group Membership			Total
			American	European	Japanese	
Original	Count	American	17	3	5	25
		European	1	6	2	9
		Japanese	0	2	14	16
	%	American	68,0	12,0	20,0	100,0
		European	11,1	66,7	22,2	100,0
		Japanese	,0	12,5	87,5	100,0
Cross-validated(a)	Count	American	17	3	5	25
		European	1	6	2	9
		Japanese	0	2	14	16
	%	American	68,0	12,0	20,0	100,0
		European	11,1	66,7	22,2	100,0
		Japanese	,0	12,5	87,5	100,0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b 74,0% of original grouped cases correctly classified.

c 74,0% of cross-validated grouped cases correctly classified.

Análisis Discriminante (con R)

```
# Se carga la librería MASS
library(MASS)

# Se hace un análisis discriminante lineal
dis <- lda(Tipo ~ Longitud + Anchura + Altura + Altura.Cara +
Anchura.Cara, data=Tibet, prior=c(0.5,0.5))

dis
Call:
lda(Tipo ~ Longitud + Anchura + Altura + Altura.Cara + Anchura.Cara,
    data = Tibet, prior = c(0.5, 0.5))

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
  Longitud  Anchura   Altura Altura.Cara Anchura.Cara
1 174.8235 139.3529 132.0000    69.82353     130.3529
2 185.7333 138.7333 134.7667    76.46667     137.5000

Coefficients of linear discriminants:
                      LD1
Longitud      0.047726591
Anchura      -0.083247929
Altura       -0.002795841
Altura.Cara   0.094695000
Anchura.Cara  0.094809401


# Se consideran las medidas de dos nuevos craneos
nuevosdatos <-
rbind(c(171,140.5,127.0,69.5,137.0),c(179.0,132.0,140.0,72.0,138.5))

# Asigno a los dos nuevos datos los nombres de las variables
colnames(nuevosdatos) <- colnames(Tibet[,-6])

nuevosdatos <- data.frame(nuevosdatos)

# Se predice el grupo de pertenencia de los nuevos datos
predict(dis,newdata=nuevosdatos)$class

[1] 1 2
Levels: 1 2

$posterior
          1         2
1 0.7545066 0.2454934
2 0.1741016 0.8258984

$x
         LD1
1 -0.6000350
2  0.8319908
```

```
# Se predicen los datos originales en los grupos segun
# la function discriminante
grupo <- predict(dis,method="plug-in")$class

# Se observa el numero de datos originales bien y mal clasificados
table(grupo,Type)
```

```
     Type
grupo  1  2
    1 14  3
    2  3 12
```

Análisis Discriminante (con SAS)

```
/* Analisis Discriminante de los datos de coches con 3 grupos */
options ls=80 nodate nonumber;
title 'Analisis Discriminante con 3 grupos de coches';
data coches;
infile 'C:\...\ADSAS.txt';
input mpg engine horse weight accel year origin cylinder;
run;

/* Analisis Discriminante con todas las variables */
proc discrim data=coches
pool=test simple manova wcov crossvalidate;
class origin;
var mpg engine horse weight accel year cylinder;
run;

/* Analisis Discriminante Stepwise con todas las variables */
proc stepdisc data=coches
sle=0.05 sls=0.05;
class origin;
var mpg engine horse weight accel year cylinder;
run;
```

Analisis Discriminante con 3 grupos de coches

The DISCRIM Procedure

Observations	50	DF Total	49
Variables	7	DF Within Classes	47
Classes	3	DF Between Classes	2

Class Level Information

origin	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	25	25.0000	0.500000	0.333333
2	_2	9	9.0000	0.180000	0.333333
3	_3	16	16.0000	0.320000	0.333333

Within-Class Covariance Matrices

origin = 1, DF = 24

Variable	mpg	engine	horse	weight
mpg	52.3639	-603.5698	-245.8552	-4912.5463
engine	-603.5698	9003.0900	3797.7017	60922.7050
horse	-245.8552	3797.7017	1985.8400	25744.2533
weight	-4912.5463	60922.7050	25744.2533	638885.6267
accel	6.9017	-153.7553	-73.8030	-707.9432
year	19.2432	-220.3650	-102.8433	-1525.9217
cylinder	-11.0957	155.9733	63.1933	1188.7633

origin = 1, DF = 24

Variable	accel	year	cylinder
mpg	6.9017	19.2432	-11.0957
engine	-153.7553	-220.3650	155.9733
horse	-73.8030	-102.8433	63.1933
weight	-707.9432	-1525.9217	1188.7633
accel	5.3401	2.3962	-2.6370
year	2.3962	12.2233	-3.7900
cylinder	-2.6370	-3.7900	3.1067

---

origin = 2, DF = 8

Variable	mpg	engine	horse	weight
mpg	40.2544	-43.1389	-82.3764	-468.6861
engine	-43.1389	449.0278	117.5278	7327.8472
horse	-82.3764	117.5278	356.5278	2054.7222
weight	-468.6861	7327.8472	2054.7222	156345.7778
accel	12.8193	12.1764	-35.5861	459.0486
year	13.5333	24.7083	-33.2917	625.2917
cylinder	0.0000	0.0000	0.0000	0.0000

origin = 2, DF = 8

Variable	accel	year	cylinder
mpg	12.8193	13.5333	0.0000
engine	12.1764	24.7083	0.0000
horse	-35.5861	-33.2917	0.0000
weight	459.0486	625.2917	0.0000
accel	9.4919	7.6667	0.0000
year	7.6667	12.0000	0.0000
cylinder	0.0000	0.0000	0.0000

origin = 3, DF = 15

Variable	mpg	engine	horse	weight
mpg	48.5200	-38.4300	-97.5446	-1883.6371
engine	-38.4300	907.4667	468.1667	8756.3667
horse	-97.5446	468.1667	505.7625	8304.1208
weight	-1883.6371	8756.3667	8304.1208	150915.7958
accel	4.1705	-34.1833	-36.5471	-542.7379
year	14.6558	14.9333	-17.4583	-284.5083
cylinder	0.9608	20.6000	8.0917	132.8083

origin = 3, DF = 15

Variable	accel	year	cylinder
mpg	4.1705	14.6558	0.9608
engine	-34.1833	14.9333	20.6000
horse	-36.5471	-17.4583	8.0917
weight	-542.7379	-284.5083	132.8083
accel	4.2343	0.8742	-0.6308
year	0.8742	8.6500	1.0833
cylinder	-0.6308	1.0833	0.6500

Simple Statistics

Total-Sample

Variable	N	Sum	Mean	Variance	Standard Deviation
mpg	50	1249	24.97000	73.48541	8.5724
engine	50	8790	175.80000	9710	98.5373
horse	50	5119	102.38000	1650	40.6156
weight	50	141903	2838	671843	819.6602
accel	50	765.90000	15.31800	5.96804	2.4430
year	50	3805	76.10000	13.11224	3.6211
cylinder	50	258.00000	5.16000	2.91265	1.7066

origin = 1

Variable	N	Sum	Mean	Variance	Standard Deviation
mpg	25	497.90000	19.91600	52.36390	7.2363
engine	25	6136	245.44000	9003	94.8846
horse	25	3089	123.56000	1986	44.5628
weight	25	84207	3368	638886	799.3032
accel	25	371.20000	14.84800	5.34010	2.3109
year	25	1879	75.16000	12.22333	3.4962
cylinder	25	156.00000	6.24000	3.10667	1.7626

---

origin = 2

Variable	N	Sum	Mean	Variance	Standard Deviation
mpg	9	260.30000	28.92222	40.25444	6.3446
engine	9	950.00000	105.55556	449.02778	21.1903
horse	9	689.00000	76.55556	356.52778	18.8819
weight	9	21073	2341	156346	395.4058
accel	9	151.00000	16.77778	9.49194	3.0809
year	9	672.00000	74.66667	12.00000	3.4641
cylinder	9	36.00000	4.00000	0	0

---

origin = 3

Variable	N	Sum	Mean	Variance	Standard Deviation
mpg	16	490.30000	30.64375	48.51996	6.9656
engine	16	1704	106.50000	907.46667	30.1242
horse	16	1341	83.81250	505.76250	22.4892
weight	16	36623	2289	150916	388.4788
accel	16	243.70000	15.23125	4.23429	2.0577
year	16	1254	78.37500	8.65000	2.9411
cylinder	16	66.00000	4.12500	0.65000	0.8062

Within Covariance Matrix Information

origin	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	7	30.05306
2	6	8.92659
3	7	22.37342
Pooled	7	30.35552

Test of Homogeneity of Within Covariance Matrices

Notation: K = Number of Groups
P = Number of Variables
N = Total Number of Observations - Number of Groups
N(i) = Number of Observations in the i'th Group - 1

$$V = \frac{\prod |\text{Within SS Matrix}(i)|^{N(i)/2}}{|\text{Pooled SS Matrix}|^{N/2}}$$

$$RHO = 1.0 - \left(\sum \frac{1}{N(i)} - \frac{1}{N} \right) \frac{2P^2 + 3P - 1}{6(P+1)(K-1)}$$

$$DF = .5(K-1)P(P+1)$$

Under the null hypothesis:

$$-2\ RHO \ln \frac{N^{PN/2}}{\prod N(i)^{PN(i)/2}} V$$

is distributed approximately as Chi-Square(DF).

Chi-Square	DF	Pr > ChiSq
220.637339	56	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV_j^{-1} \, (\bar{X}_i - \bar{X}_j) + \ln |COV_j|$$

Generalized Squared Distance to origin

From origin	1	2	3
1	30.05306	291281292	67.84795
2	35.53867	8.92659	38.59525
3	34.59605	907076	22.37342

Multivariate Statistics and F Approximations

S=2 M=2 N=19.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.28802184	5.06	14	82	<.0001
Pillai's Trace	0.88078861	4.72	14	84	<.0001
Hotelling-Lawley Trace	1.88585602	5.43	14	62.325	<.0001
Roy's Greatest Root	1.49339170	8.96	7	42	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

Classification Summary for Calibration Data: WORK.COCHES
Resubstitution Summary using Quadratic Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)' \, COV_j^{-1} \, (X-\bar{X}_j) + \ln |COV_j|$$

Posterior Probability of Membership in Each origin

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Number of Observations and Percent Classified into origin

From origin	1	2	3	Total
1	18	7	0	25
	72.00	28.00	0.00	100.00
2	0	9	0	9
	0.00	100.00	0.00	100.00
3	0	9	7	16
	0.00	56.25	43.75	100.00
Total	18	25	7	50
	36.00	50.00	14.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for origin

	1	2	3	Total
Rate	0.2800	0.0000	0.5625	0.2808
Priors	0.3333	0.3333	0.3333	

Classification Summary for Calibration Data: WORK.COCHES
Cross-validation Summary using Quadratic Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_{(X)j})' \, COV_{(X)j}^{-1} \, (X-\bar{X}_{(X)j}) + \ln |COV_{(X)j}|$$

Posterior Probability of Membership in Each origin

$$Pr(j|X) = \exp(-.5\ D_j^2(X)) \Big/ \sum_k \exp(-.5\ D_k^2(X))$$

Number of Observations and Percent Classified into origin

From origin	1	2	3	Total
1	17	8	0	25
	68.00	32.00	0.00	100.00
2	3	5	1	9
	33.33	55.56	11.11	100.00
3	2	11	3	16
	12.50	68.75	18.75	100.00
Total	22	24	4	50
	44.00	48.00	8.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for origin

	1	2	3	Total
Rate	0.3200	0.4444	0.8125	0.5256
Priors	0.3333	0.3333	0.3333	

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE

Observations	50	Variable(s) in the Analysis	7
Class Levels	3	Variable(s) will be Included	0
		Significance Level to Enter	0.05
		Significance Level to Stay	0.05

Class Level Information

origin	Variable Name	Frequency	Weight	Proportion
1	_1	25	25.0000	0.500000
2	_2	9	9.0000	0.180000
3	_3	16	16.0000	0.320000

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 2, 47

Variable	R-Square	F Value	Pr > F	Tolerance
mpg	0.3594	13.19	<.0001	1.0000
engine	0.5097	24.43	<.0001	1.0000
horse	0.2812	9.20	0.0004	1.0000
weight	0.4275	17.55	<.0001	1.0000
accel	0.0849	2.18	0.1244	1.0000
year	0.1920	5.59	0.0067	1.0000
cylinder	0.4093	16.28	<.0001	1.0000

Variable engine will be entered.

Variable(s) that have been Entered

engine

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.490318	24.43	2	47	<.0001
Pillai's Trace	0.509682	24.43	2	47	<.0001
Average Squared Canonical Correlation	0.254841				

The STEPDISC Procedure
Stepwise Selection: Step 2

Statistics for Removal, DF = 2, 47

Variable	R-Square	F Value	Pr > F
engine	0.5097	24.43	<.0001

No variables can be removed.

Statistics for Entry, DF = 2, 46

Variable	Partial R-Square	F Value	Pr > F	Tolerance
mpg	0.0179	0.42	0.6604	0.3645
horse	0.1115	2.89	0.0659	0.2227
weight	0.0075	0.17	0.8412	0.2187
accel	0.1237	3.25	0.0480	0.7623
year	0.1714	4.76	0.0133	0.7843
cylinder	0.0335	0.80	0.4571	0.1009

Variable year will be entered.

Variable(s) that have been Entered

engine year

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.406296	13.08	4	92	<.0001
Pillai's Trace	0.674371	11.95	4	94	<.0001
Average Squared Canonical Correlation	0.337185				

The STEPDISC Procedure
Stepwise Selection: Step 3

Statistics for Removal, DF = 2, 46

Variable	Partial R-Square	F Value	Pr > F
engine	0.4971	22.74	<.0001
year	0.1714	4.76	0.0133

No variables can be removed.

Statistics for Entry, DF = 2, 45

Variable	Partial R-Square	F Value	Pr > F	Tolerance
mpg	0.0624	1.50	0.2349	0.2261
horse	0.1530	4.06	0.0239	0.2034
weight	0.0068	0.15	0.8576	0.2093
accel	0.1427	3.75	0.0313	0.6454
cylinder	0.0347	0.81	0.4514	0.0932

Variable horse will be entered.

Variable(s) that have been Entered

engine horse year

Multivariate Statistics

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.344148	10.57	6	90	<.0001
Pillai's Trace	0.779271	9.79	6	92	<.0001
Average Squared Canonical Correlation	0.389636				

The STEPDISC Procedure
Stepwise Selection: Step 4

Statistics for Removal, DF = 2, 45

Variable	Partial R-Square	F Value	Pr > F
engine	0.3954	14.71	<.0001
horse	0.1530	4.06	0.0239
year	0.2100	5.98	0.0050

No variables can be removed.

Statistics for Entry, DF = 2, 44

Variable	Partial R-Square	F Value	Pr > F	Tolerance
mpg	0.0661	1.56	0.2222	0.1779
weight	0.0204	0.46	0.6360	0.1380
accel	0.0477	1.10	0.3416	0.1435
cylinder	0.0493	1.14	0.3285	0.0676

No variables can be entered.

No further steps are possible.

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	engine		0.5097	24.43	<.0001	0.49031755	<.0001
2	2	year		0.1714	4.76	0.0133	0.40629584	<.0001
3	3	horse		0.1530	4.06	0.0239	0.34414795	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	engine		0.25484122	<.0001
2	2	year		0.33718537	<.0001
3	3	horse		0.38963551	<.0001