INVESTIGACIÓN DEL FRAUDE FISCAL MEDIANTE ANÁLISIS DISCRIMINANTE. APLICACIONES CON LAS MUESTRAS Y PANELES DE IRPF DEL IEF

César Pérez López

Instituto de Estudios Fiscales y Universidad Complutense de Madrid

cesar.perez@ief.minhap.es

ABSTRACT

Este papel de trabajo tiene como finalidad presentar las posibilidades de análisis que se abren ante la disponibilidad de grandes conjuntos de datos con información relativa a impuestos. En el caso que nos atañe se trata de mostrar el uso de la metodología del análisis discriminante aplicada a las muestras y paneles de IRPF del IEF con la finalidad de detectar posible fraude fiscal en este impuesto. A partir de la muestra de IRPF de 2009 y mediante la aplicación de la función discriminante de Fisher se buscará un modelo discriminante que permite asignar una probabilidad de fraude a cualquier declarante de IRPF basándose exclusivamente en la información que declara a la Agencia tributaria en el modelo correspondiente.

1. EL MODELO DE ANÁLISIS DISCRIMINANTE COMO TÉCNICA DE CLASIFICACIÓN Y SEGMENTACIÓN

El análisis discriminante es una técnica que tiene como finalidad construir un modelo predictivo para pronosticar el grupo al que pertenece una observación a partir de determinadas características observadas que delimitan su perfil. Se trata de una técnica estadística que permite asignar o clasificar nuevos individuos u observaciones dentro de grupos o segmentos previamente definidos, razón por la cual es una técnica de clasificación y segmentación ad hoc. El análisis discriminante se conoce en ocasiones como análisis de la clasificación, ya que su objetivo fundamental es producir una regla o un esquema de clasificación que permita a un investigador predecir la población a la que es más probable que tenga que pertenecer una nueva observación o individuo.

El modelo predictivo que pronostica el grupo de pertenencia de una observación en virtud de su perfil define la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. Por tanto, la expresión funcional del análisis discriminante puede escribirse como $y = F(x_1, x_2, \cdots, x_n)$ con la variable dependiente no métrica y las variables independientes métricas. Las categorías de la variable dependiente definen los posibles grupos de pertenencia de las observaciones o individuos y las variables independientes definen el perfil conocido de cada observación. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes medidas sobre un individuo u observación (perfil) para predecir con qué

categoría de la variable dependiente se corresponden para clasificar al individuo en la categoría adecuada.

En la aplicación que aquí se presenta, se utiliza la muestra de IRPF de 2009, tomándose como variables independientes del modelo las declaradas por el individuo en el modelo 100 de IRPF (prácticamente 300 variables) y como variable dependiente una variable dicotómica que toma el valor 1 si el individuo defrauda y toma el valor 0 si el individuo no defrauda. Con el modelo discriminante se buscará predecir la probabilidad que tiene cualquier individuo de defraudar o no, según los valores declarados en las variables del modelo 100. Buscamos por tanto, perfiles de fraude que puedan ayudar en el futuro a la labor inspectora.

Las dos grandes finalidades perseguidas en el uso del análisis discriminante son la descripción de diferencias entre grupos de la variable categórica dependiente y la predicción de pertenencia a los citados grupos. La interpretación de las diferencias entre los grupos responde al objetivo de saber en qué medida un conjunto de características observadas en individuos permite extraer dimensiones que diferencian a los grupos, y cuáles de estas características son las que en mayor medida contribuyen a tales dimensiones, es decir, las de mayor poder de discriminación.

Las características usadas para diferenciar entre los grupos reciben el nombre de *variables discriminantes*. La predicción de pertenencia a los grupos se lleva a cabo determinando una o más ecuaciones matemáticas, denominadas *funciones discriminantes*, que permitan la clasificación de nuevos casos a partir de la información que poseemos sobre ellos. Estas ecuaciones combinan una serie de características o variables de tal modo que su aplicación a un caso nos permite identificar el grupo al que más se parece. En este sentido podremos hablar del carácter predictivo del análisis discriminante

2. HIPÓTESIS EN EL MODELO DISCRIMINANTE

El modelo subyacente en el análisis discriminante requiere de una comprobación de determinados supuestos. Para comenzar, la aplicación del análisis discriminante requiere que contemos con un conjunto de variables discriminantes (características conocidas de los individuos) y una variable nominal que define dos o más grupos (cada modalidad de la variable nominal se corresponde con un grupo diferente). Además, los datos deben corresponder a individuos o casos clasificados en dos o más grupos mutuamente excluyentes. Es decir, cada caso corresponde a un grupo y sólo a uno. Por otra parte, las variables discriminantes han de estar medidas en una escala de intervalo o de razón, lo cual permitiría el cálculo de medias y varianzas y la utilización de éstas en ecuaciones matemáticas. Teóricamente, no existen límites para el número de variables discriminantes, salvo la restricción de que no debe ser nunca superior al número de casos en el grupo más pequeño, pero sí es conveniente contar al menos con 20 sujetos por cada variable discriminante si queremos que las interpretaciones y conclusiones obtenidas sean correctas. ***Todas estas condiciones se cumplen con creces en la aplicación que aquí se trata. En el modelo 100 tenemos más de 400 variables utilizándose en este trabajo prácticamente 300 de ellas. La muestra de IRPF de 2009 tiene cerca de dos millones de declarantes, con lo que el tamaño muestral es suficientemente alto. Las dos categorías de la variable dependiente son mutuamente excluyentes, ya que se trata de***

defraudadores y no defraudadores. Por motivos de confidencialidad legalmente exigidos y escrupulosamente respetados en esta investigación, los datos muestrales de individuos defraudadores y no defraudadores son ficticios, además de utilizar una base de datos totalmente anonimizada. En la práctica, serían defraudadores los individuos de la muestra que la inspección ha determinado fehacientemente como tales defraudadores.

En cuanto a la presencia de datos desaparecidos (*missing*), hay que tener presente que cuando corresponden a la variable de clasificación, los individuos afectados podrían ser excluidos del análisis a la hora de determinar las funciones discriminantes. Si los datos desaparecidos están en variables independientes, hay que asegurarse de que los individuos en los que se registra la ausencia de datos no posean características diferenciales respecto al resto de los individuos, modificando las características de la muestra con la que trabajamos. Si se diera esta circunstancia, sería necesario recurrir a alguno de los procedimientos para tratar los casos desaparecidos (imputación por la media, por regresión, por métodos especiales etc.). ***En nuestro caso, los datos missing se distribuyen aleatoriamente por toda la muestra, situación ideal ante este tipo de problema. Hay contrastes formales, como el contraste de Little, el contaste de las pruebas pareadas y el contraste de la matriz de correlaciones dicotomizadas para constatar este hecho. Por otra parte, la variable dependiente no tiene datos missing****.*

Por otro lado, la aplicación del análisis discriminante se apoya en una serie de supuestos básicos como la normalidad multivariante, homogeneidad de matrices de varianza-covarianza (homoscedasticidad), linealidad y ausencia de multicolinealidad. El supuesto de *normalidad* exige que cada grupo represente una muestra aleatoria extraída de una población con distribución normal multivariable sobre las variables discriminantes. La normalidad univariante no implica la multivariante, pero como esta última es difícil de comprobar, se contrasta la normalidad univariante mediante pruebas clásicas como la prueba de bondad de ajuste basada en *Chi-cuadrado*, la prueba de Kolmogorov-Smirnov, el test de Shapiro-Wilk o las pruebas de significación basadas en la asimetría y la curtosis. ***En nuestro caso, sabemos que las variables de renta no son normales y que suelen seguir una distribución paretiana truncada. Este problema se solventa utilizando como variables discriminantes los factores resultantes de aplicar un análisis de componentes principales con rotación ortogonal varimax sobre las variables independientes iniciales. Dada la cantidad de variables, la cantidad de factores y el tamaño de la muestra, puede presuponerse la convergencia a la normalidad de los factores por aplicación del teorema central del límite. Además, el uso de factores refuerza la confidencialidad de las variables con más incidencia en el fraude fiscal.***

En cuanto a los casos aislados (*outliers*), es necesario detectar su existencia en cada una de las variables consideradas por separado. Para la detección de casos aislados multivariantes podría recurrirse al cálculo de la distancia de Mahalanobis de cada individuo respecto al centro del grupo o a un método gráfico. ***En nuestro caso, el uso de factores que engloban cada uno de ellos varias variables iniciales, minimiza el efecto de los valores atípicos****.*

El supuesto de homogeneidad de matrices de varianza-covarianza (*homoscedasticidad*) obliga a que las matrices de varianzas-covarianzas para las poblaciones de las que fueron extraídos los grupos sean iguales, hipótesis que suele probarse mediante la prueba de *M* de Box, que no es más que una generalización del test de Barlett para la comprobación de la homogeneidad

de varianzas univariadas y que se basa en los determinantes de las matrices de varianzas-covarianzas para cada grupo. Por otro lado, el supuesto de *linealidad* implica que existen relaciones lineales entre las variables dentro de cada grupo y suele comprobarse a partir de los diagramas de dispersión de las variables o mediante el cálculo de coeficientes de correlación lineal de Pearson. La matriz de correlaciones de las variables también se utiliza para detectar la *multicolinealidad* (variables con correlación muy alta pueden ser redundantes), que puede ser muy nociva en la inversión de matrices requeridas en los algoritmos discriminantes. ***En nuestro caso, estos problemas también desaparecen al utilizar factores en vez de variables iniciales como variables independientes (variables discriminantes) del modelo discriminante. No obstante, se presentarán contrastes formales para estas hipótesis. El problema de la multicolinealidad queda perfectamente resuelto con la utilización de los factores.***

3. COMPONENTES PRINCIPALES

El análisis en componentes principales es una técnica de análisis estadístico multivariante que se clasifica entre los métodos de interdependencia. Se trata de un método multivariante de simplificación o reducción de la dimensión y que se aplica cuando se dispone de un conjunto elevado de variables con datos cuantitativos correlacionadas entre sí persiguiendo obtener un menor número de variables, combinación lineal de las primitivas e incorrelacionadas, que se denominan componentes principales o factores, que resuman lo mejor posible a las variables iniciales con la mínima pérdida de información y cuya posterior interpretación permitirá un análisis más simple del problema estudiado. Esta reducción de muchas variables a pocas componentes puede simplificar la aplicación sobre estas últimas de otras técnicas multivariantes (regresión, clusters, discriminante, etc.).

El elevado número de variables iniciales x_1, x_2,…,x_p se resumen en unas pocas variables C_1, C_2,…,C_k (*componentes principales*) *perfectamente calculables* ($k<<p$) combinanción lineal de las iniciales y que sintetizan la mayor parte de la información contenida en sus datos. Inicialmente se tienen tantas componentes como variables:

$$\begin{aligned} C_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p) \\ &\vdots \\ C_p &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{pp}x_p) \end{aligned}$$

Pero sólo se retienen las k componentes principales que explican un porcentaje alto de la variabilidad de las variables iniciales (C_1, C_2,…, C_k).

La primera componente principal, al igual que las restantes, se expresa como combinación lineal de las variables originales como sigue:

$$C_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \cdots + u_{1p}X_{pi} \quad i=1,..,n$$

Para el conjunto de las n observaciones muestrales y para todas las componentes tenemos:

$$\begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} \cdots & X_{p1} \\ X_{12} & X_{22} \cdots & X_{p2} \\ & \vdots & \\ X_{1n} & X_{2n} \cdots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

En notación abreviada tendremos: $C_1 = X\,u_1$ y:

$$V(C_1) = \frac{\sum_{i=1}^{n} C_{1i}^2}{n} = \frac{1}{n} C_1{}' C_1 = \frac{1}{n} u_1' X'X u_1 = u_1' \left[\frac{1}{n} X'X \right] u_1 = u_1' V u_1$$

La primera componente C_1 se obtiene de forma que su varianza sea máxima sujeta a la restricción de que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, es decir, la variable de los pesos o ponderaciones (u_{11}, u_{12},...,u_{1p})' se toma normalizada. Se trata entonces de hallar C_1 maximizando $V(C_1) = u_1'Vu_1$, sujeta a la restricción:

$$\sum_{j=1}^{p} u_{1i}^2 = u_1' u_1 = 1$$

Se demuestra que, para maximizar $V(C_1)$ se toma el mayor valor propio λ de la matriz V. Sea λ_1 el citado mayor valor propio de V y tomando u_1 como su vector propio asociado normalizado ($u_1'u_1$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la primera componente principal, componente que vendrá definida como:

$$C_1 = u_1 X = u_{11} X_1 + u_{12} X_2 + \cdots + u_{1p} X_p$$

Para maximizar $V(C_2)$ hemos de tomar el segundo mayor valor propio λ de la matriz V (el mayor ya lo había tomado al obtener la primera componente principal) .

Tomando λ_2 como el segundo mayor valor propio de V y tomando u_2 como su vector propio asociado normalizado ($u_2'u_2$=1), ya tenemos definido el vector de ponderaciones que se aplica a las variables iniciales para obtener la segunda componente principal, componente que vendrá definida como:

$$C_2 = u_2 X = u_{21} X_1 + u_{22} X_2 + \cdots + u_{2p} X_p$$

De forma similar, la componente principal h-ésima se define como $C_h = Xu_h$ donde u_h es el vector propio de V asociado a su h-ésimo mayor valor propio. Suele denominarse también a u_h eje factorial h-ésimo.

Se demuestra que la proporción de la variabilidad total recogida por la componente principal h-ésima (porcentaje de inercia explicada por la componente principal h-ésima) vendrá dada por:

$$\frac{\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\lambda_h}{traza(V)}$$

Si las variables están tipificadas, traza(V) = p, con lo que la proporción de la componente h-esima en la variabilidad total será λh/p. También se define el porcentaje de inercia explicada por las k primeras componentes principales (o ejes factoriales) como:

$$\frac{\sum_{h=1}^{k}\lambda_h}{\sum_{h=1}^{p}\lambda_h} = \frac{\sum_{h=1}^{k}\lambda_h}{traza(V)}$$

Cuando las variables originales están muy correlacionadas entre sí, la mayor parte de su variabilidad se puede explicar con muy pocas componentes. Si las variables originales estuvieran completamente incorrelacionadas entre sí, entonces el análisis de componentes principales carecería por completo de interés, ya que en ese caso las componentes principales coincidirían con las variables originales.

Como ***criterio general para precisar el número de componentes a retener***, se seleccionan aquellas componentes cuya raíz característica λ_j excede de la media de las raíces características. Recordemos que la raíz característica asociada a una componente es precisamente su varianza. Analíticamente este criterio implica retener todas aquellas componentes en que se verifique que:

$$\lambda_h > \overline{\lambda} = \frac{\sum_{j=1}^{p}\lambda_h}{p}$$

Si se utilizan variables tipificadas, entonces, como ya se ha visto, se verifica que $\sum_{j=1}^{p}\lambda_h = p$,

Con lo que para variables tipificadas se retiene aquellas componentes tales que λ_h >1. La representación gráfica de este criterio se conoce como ***gráfico de sedimentación***.

La dificultad en la interpretación de los componentes estriba en la necesidad de que tengan sentido y midan algo útil en el contexto del fenómeno estudiado. Por tanto, es indispensable considerar el ***peso que cada variable original tiene dentro del componente elegido***, así como las correlaciones existentes entre variables y factores. Un componente es una función lineal de todas las variables, pero puede estar muy bien correlacionado con algunas de ellas, y menos con otras. Ya hemos visto que el coeficiente de correlación entre una componente y una variable se calcula multiplicando el peso de la variable en esa componente por la raíz cuadrada de su valor propio:

$$r_{jh} = u_{hj}\sqrt{\lambda_h}$$

Se demuestra también que estos coeficientes *r* representan la parte de varianza de cada variable que explica cada factor. De este modo, cada variable puede ser representada como una función lineal de los *k* componentes retenidos, donde los pesos o cargas de cada componente o factor (*cargas factoriales*) en la variable coinciden con los coeficientes de correlación.

El cálculo matricial permite obtener de forma inmediata la tabla de coeficientes de correlación variables-componentes (*pxk*), que se denomina ***matriz de cargas factoriales***. Las ecuaciones de las variables en función de las componentes (factores), traspuestas las inicialmente planteadas, son de mayor utilidad en la interpretación de los componentes, y se expresan como sigue:

$$\begin{array}{l} C_1 = r_{11}X_1 + \cdots + r_{1p}X_p \\ C_2 = r_{21}X_1 + \cdots + r_{2p}X_p \\ \vdots \\ C_k = r_{k1}X_1 + \cdots + r_{kp}X_p \end{array} \Rightarrow \begin{array}{l} X_1 = r_{11}C_1 + \cdots + r_{k1}C_k \\ X_2 = r_{12}C_1 + \cdots + r_{k2}C_k \\ \vdots \\ X_p = r_{1p}C_1 + \cdots + r_{kp}C_k \end{array}$$

Es frecuente no encontrar interpretaciones verosímiles a los factores (componentes) obtenidos. Sería deseable, para una más fácil interpretación, que cada componente estuviera relacionada muy bien con pocas variables (coeficientes de correlación *r* próximos a 1 ó -1) y mal con las demás (*r* próximos a 0). Esta optimización se obtiene por una adecuada ***rotación de los ejes*** que definen los componentes principales.

Rotar un conjunto de componentes no cambia la proporción de inercia total explicada, como tampoco cambia las comunalidades de cada variable, que no son sino la proporción de varianza explicada por todos ellos. Las rotaciones más utilizadas son la rotación VARIMAX y la QUARTIMAX (ortogonales) y PROMAX (oblicua).

Sin embargo, los coeficientes, que dependen directamente de la posición de los componentes respecto a las variables originales (cargas factoriales y valores propios), se ven alterados por la rotación.

En nuestro caso hemos obtenido 69 componentes principales Ci (factores) que explican más del 78% de la variabilidad inicial de los datos, resultando así una buena reducción. Las puntuaciones de estos factores serán utilizadas como nuestras 69 variables discriminantes. Evidentemente también conocemos las expresiones de las combinaciones lineales que nos definen cada factor Ci en función de las variables iniciales Xi a partir de los valores de la matriz de cargas factoriales rotadas que se presenta a continuación.

	1	2	3	4	5	6	7	8	9	10	11	12	13
Número total de descendientes	,071	-,009	-,007	,002	-,025	,116	-,013	-,003	-,004	,013	,913	,148	,002
Número de descendientes <3 años	,003	-,004	-,003	,000	-,019	,068	,002	,000	-,012	,005	,295	,044	-,001
Número de descendientes >= 3 y < 16 años	,051	-,008	-,007	,000	-,023	,123	-,009	-,002	-,008	,012	,902	,084	,002
Número de descendientes >= 16 y < 18 años	,026	-,001	-,002	-,001	-,002	,018	-,003	,000	,001	,003	,207	,037	,001
Número de descendientes >= 18 y < 25 años	,055	-,003	-,002	,004	-,001	-,011	-,012	-,002	,014	,002	,152	,132	,002
Número de descendientes >=25 años	-,003	,000	,000	,000	,005	-,013	,000	,001	,000	,001	-,005	,025	,001
Número de descendientes con edad desconocida	-,001	,000	,001	,000	,000	-,006	,004	-,001	,000	-,002	-,004	,008	,000
Número de descendientes sin minusvalía	,071	-,009	-,007	,002	-,025	,117	-,013	-,003	-,004	,012	,914	,143	,002

C1 =0,071X1+0,03X2+0,051X3+0,026x4+…….

C2=-0,009X1-0,04X2-0,08X3.0,001 X4 +……..

C3=-0,007X1-0,003 X2-0,007X3-0,002X4+……..

……

……

3. ESTIMACIÓN DEL MODELO DISCRIMINANTE

En el análisis discriminante, una vez comprobado el cumplimiento de los supuestos subyacentes al modelo matemático, se persigue obtener una serie de funciones lineales a partir de las variables independientes que permitan interpretar las diferencias entre los grupos y clasificar a los individuos en alguna de las subpoblaciones definidas por la variable dependiente. Estas funciones lineales se denominan funciones discriminantes y son combinaciones lineales de las variables discriminantes. En el caso general de análisis discriminante con *G* grupos (*G* > 2) llamado ***análisis discriminante múltiple***, el número máximo de funciones o ejes discriminantes que se pueden obtener viene dado por *min(G*-1*,k).* Por tanto pueden obtenerse hasta *G*-1 ejes discriminantes, si el número de variables explicativas *k* es mayor o igual que *G*-1, hecho que suele ser siempre cierto, ya que en las aplicaciones prácticas el número de variables explicativas suele ser grande.

Cada una de las funciones discriminantes D_i se obtiene como función lineal de las *k* variables explicativas *X*, es decir:

$$D_i = u_{i1}X_1 + u_{i2}X_2 + \cdots + u_{ik}X_k \quad i=1,2,...,G-1$$

Los *G*-1 ejes discriminantes vienen definidos respectivamente por los vectores u_1, u_2,...,u_{G-1} definidos mediante las siguientes expresiones:

$$u_1 = \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1k} \end{bmatrix} \quad u_2 = \begin{bmatrix} u_{21} \\ u_{22} \\ \vdots \\ u_{2k} \end{bmatrix} \quad \ldots \quad u_{G-1} = \begin{bmatrix} u_{G-11} \\ u_{G-12} \\ \vdots \\ u_{G-1k} \end{bmatrix}$$

Consideremos las matrices *F, T* y *W* del análisis de la varianza múltiple correspondiente. Para la obtención del primer eje discriminante, se maximiza λ_1, donde:

$$\lambda_1 = \frac{u_1' F u_1}{u_1' W u_1}$$

La solución a este problema se obtiene derivando λ_1 respecto de *u* e igualando a cero, es decir:

$$\frac{\partial \lambda_1}{\partial u_1} = \frac{2Fu_1(u_1'Wu_1) - 2Wu_1(u_1'Fu_1)}{(u_1'Wu_1)^2} = 0 \Rightarrow 2Fu_1(u_1'Wu_1) - 2Wu_1(u_1'Fu_1) = 0$$

De donde:

$$\frac{2Fu_1}{2Wu_1} = \frac{u_1'Fu_1}{u_1'Wu_1} = \lambda_1 \Rightarrow Fu_1 = Wu_1\lambda_1 \Rightarrow W^{-1}Fu_1 = \lambda_1 u_1$$

Por tanto, la ecuación para la obtención del primer eje discriminante $W^{-1}Fu_1 = \lambda_1 u_1$ se traduce en la obtención de un vector propio u_1 asociado a la matriz no simétrica $W^{-1}F$.

De los valores propios λ_i que se obtienen al resolver la ecuación $W^{-1}Fu_1 = \lambda_1 u_1$ se retiene el mayor, ya que precisamente λ_1 es la ratio que queremos maximizar y u_1 es el vector propio asociado al mayor valor propio de la matriz $W^{-1}F$.

Dado que λ_1 es la ratio a maximizar, nos medirá, una vez calculado, el poder discriminante del primer eje discriminante. Como estamos en un caso general de análisis discriminante con *G* grupos (*G* > 2), el número máximo de ejes discriminantes que se pueden obtener viene dado por *min*(*G* -1,*k*). Por tanto, pueden obtenerse hasta *G*-1 ejes discriminantes, si el número de variables explicativas *k* es mayor o igual que *G*-1, hecho que suele ser siempre cierto, ya que en las aplicaciones prácticas el número de variables explicativas suele ser grande.

El resto de los ejes discriminantes vendrá dado por los vectores propios asociados a los valores propios de la matriz $W^{-1}F$ ordenados de mayor a menor. Así, el segundo eje discriminante tendrá menos poder discriminatorio que el primero, pero más que cualquiera de los restantes.

Como la matriz $W^{-1}F$ no es simétrica, los ejes discriminantes no serán en general ortogonales (perpendiculares entre sí).

Podemos concluir que los ejes discriminantes son las componentes de los vectores propios normalizados asociados a los valores propios de la matriz $W^{-1}F$ ordenados en sentido decreciente (a mayor valor propio mejor eje discriminante).

4. CONTRASTES DE SIGNIFICACIÓN EN EL MODELO DISCRIMINANTE

En cuanto a los ***contrastes de significación***, en el análisis discriminante múltiple se plantean contrastes específicos para determinar si cada uno de los valores λ_i que se obtienen al resolver la ecuación $W^{-1}Fu = \lambda u$ es estadísticamente significativo (o lo que es lo mismo, se trata de contrastar si las funciones discriminantes correspondientes son significativas), es decir, para determinar si contribuye o no a la discriminación entre los diferentes grupos.

Este tipo de contrastes se realiza a partir del estadístico *V* de Bartlett, que es una función de la Λ de Wilks y que se aproxima a una *Chi-cuadrado*. Su expresión es la siguiente:

$$V = -\left\{n - 1 - \frac{k+G}{2}\right\} Ln(\Lambda) \rightarrow \chi^2_{k(G-1)} \qquad \Lambda = \frac{|W|}{|T|}$$

La hipótesis nula de este contraste es $H_0 : \mu_1 = \mu_2 = \ldots \mu_G$, y ha de ser rechazada para que se pueda continuar con el análisis discriminante, porque en caso contrario las variables clasificadoras utilizadas no tendrían poder discriminante alguno.

No olvidemos que *W* era la matriz suma de cuadrados y productos cruzados intragrupos en el análisis de la varianza múltiple y *T* era la matriz suma de cuadrados y productos cruzados total.

También existe un ***estadístico de Bartlett para contrastación secuencial***, que se elabora como sigue:

$$\frac{1}{\Lambda} = \frac{|T|}{|W|} = |W|^{-1}|T| = |W^{-1}T| = |W^{-1}(W+F)| = |I + W^{-1}F|$$

Pero como el determinante de una matriz es igual al producto de sus valores propios, se tiene que:

$$\frac{1}{\Lambda} = (1+\lambda_1)(1+\lambda_2)\cdots(1+\lambda_{G-1})$$

Esta expresión puede sustituirse en la expresión del estadístico *V* vista anteriormente, para obtener la expresión alternativa siguiente para el estadístico de Bartlett:

$$V = -\left\{n - 1 - \frac{k+G}{2}\right\} Ln(\Lambda) = -\left\{n - 1 - \frac{k+G}{2}\right\} \sum_{g=1}^{G-1} Ln(1+\lambda_g) \rightarrow \chi^2_{k(G-1)}$$

Si se rechaza la hipótesis nula de igualdad de medias, al menos uno de los ejes discriminantes es estadísticamente significativo, y será el primero, porque es el que más poder discriminante tiene.

Una vez visto que el primer eje discriminante es significativo, se pasa a analizar la significatividad del segundo eje discriminante a partir del estadístico:

$$V = -\left\{ n - 1 - \frac{k+G}{2} \right\} \sum_{g=2}^{G-1} Ln(1+\lambda_g) \rightarrow \chi^2_{(k-1)(G-2)}$$

De la misma forma se analiza la significatividad de sucesivos ejes discriminantes, pudiendo establecerse el estadístico *V* de Bartlett genérico para contrastación secuencial de la significatividad del eje discriminate j-ésimo como:

$$V = -\left\{ n - 1 - \frac{k+G}{2} \right\} \sum_{g=j+1}^{G-1} Ln(1+\lambda_g) \rightarrow \chi^2_{(k-j)(G-j-1)} \qquad j = 0,1,2,\cdots,G-2$$

En este proceso secuencial se van eliminando del estadístico *V* las raíces características que van resultando significativas, deteniendo el proceso cuando se acepte la hipótesis nula de no significatividad de los ejes discriminantes que queden por contrastar.

Como una medida descriptiva complementaria de este contraste se suele calcular el porcentaje acumulativo de la varianza después de la incorporación de cada nueva función discriminante.

Un modo de valorar la importancia discriminante de cada una de las funciones consiste en compararlas entre sí, de modo que conozcamos cuáles destacan en relación a las demás. Bastaría sumar todos los autovalores y dividir por esta cantidad cada uno de ellos. Este cálculo nos conduciría a los porcentajes relativos, los cuales indican el porcentaje que una función posee sobre el poder discriminante total acumulado por el conjunto de funciones. Los porcentajes relativos nos dan información de la importancia de una función en relación a las restantes, pero no aportan un criterio definitivo para decidir si una función discriminante ha de ser retenida.

No es posible fijar un porcentaje mínimo a partir del cual pudiéramos afirmar que la función discriminante resulta de interés para nuestros propósitos de discriminación. Podría suceder que aunque el porcentaje que representa un autovalor sea muy superior al de otras funciones, todas ellas resulten igualmente poco significativas de cara a establecer diferencias entre los grupos.

Otro modo de juzgar la importancia de las funciones discriminantes se basa en el cálculo del coeficiente de correlación canónica que, al igual que el autovalor, mide las desviaciones de las puntuaciones discriminantes entre los grupos respecto a las desviaciones dentro de los grupos. El ***coeficiente de correlación canónica*** $\left(r^*\right)$ está relacionado con el autovalor mediante la siguiente expresión, referida a una función discriminante *i*:

$$r_i^* = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

Este coeficiente proviene del análisis de correlaciones canónicas, que ya sabemos que estudia el grado de asociación entre dos conjuntos de variables medidas en escala de intervalo. Se desarrolla creando *q* pares de combinaciones lineales, siendo *q* el número de variables en el conjunto más pequeño. Las combinaciones lineales en cada par se generan maximizando la correlación entre ambas. Para el primer par tendremos el mayor grado de asociación; para el segundo, se determinan las combinaciones lineales de modo que presenten el mayor grado de asociación entre sí, pero con la condición adicional de que no esté correlacionada con las del primer par; y así sucesivamente hasta el q-ésimo par. El coeficiente de correlación canónica es una medida idéntica a la correlación de Pearson entre las combinaciones lineales de un par. En el caso que nos ocupa, las variables discriminantes constituyen uno de los conjuntos de variables, mientras que el otro conjunto surge de representar los grupos mediante *G*-1 variables *dummy*. Si para estos dos grupos generamos *q* pares de combinaciones lineales, la función discriminante constituirá una parte del par y la combinación dada por los grupos la otra parte. El coeficiente de correlación canónica se interpreta como una medida de la asociación entre los dos conjuntos de variables.

Otra interpretación posible del coeficiente de correlación canónica se basa en el análisis de varianza, en cuyo contexto recibe la denominación de *coeficiente eta*. En el caso del análisis que nos ocupa, tomaríamos como variable dependiente a la función discriminante y como variable independiente a los grupos. Mediante el *coeficiente eta* puede ser medido el grado en que difieren las medias alcanzadas por la función discriminante en los grupos. El ***coeficiente eta al cuadrado*** representaría el porcentaje de varianza de la función discriminante explicada por la diferencia entre grupos. En términos del análisis de varianza, tendríamos:

$$eta^2 = \frac{SC_{\text{int}\,ergrupos}}{SC_{total}}$$

Recurriendo al coeficiente de correlación canónica, puede ser evaluada la relevancia de las funciones discriminantes. Un valor alto para este coeficiente indicaría que existe una relación entre el grupo de pertenencia y los valores de la función discriminante. Es decir, la función adopta diferentes valores en los grupos considerados y responde satisfactoriamente al propósito de discriminar entre los grupos. Mientras que el porcentaje relativo nos indicaba cuál era la función más potente, la correlación canónica nos indica en qué grado ésta resulta relevante. En una situación en la que los grupos no sean suficientemente diferentes respecto a las variables analizadas, podremos determinar una función discriminante que presenta el mayor porcentaje relativo. Sin embargo, el criterio de la correlación canónica nos permitirá rechazar esta función, dado que el valor del mismo habrá de ser bajo.

Para nuestro modelo observamos los resultados del contraste de la Lambda de Wilks cuyo p-valor (Sig.) pequeño valida la significatividad del modelo discriminante en su conjunto. También observamos los resultados de la prueba M de Box, cuyo p-valor pequeño muestra la ausencia de heteroscedasticidad en el modelo. La ausencia de multicolinealidad viene determinada por el uso de los factores que son incorrelados y la hipótesis de normalidad la asegura el uso de factores rotados y el teorema central del límite, como ya se ha indicado

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,647	839585,157	63	,000

Resultados de la prueba

M de Box		125753695,7
F	Aprox.	62375,585
	gl1	2016
	gl2	7,112E+12
	Sig.	,000

Contrasta la hipótesis nula de que las matrices de covarianzas poblacionales son iguales.

5. SELECCIÓN DE VARIABLES DISCRIMINANTES

A veces el análisis discriminante es utilizado sin que tengamos la certeza de que nuestras variables poseen una suficiente capacidad de discriminación. En ese caso, el investigador partiría de una lista de variables, sin que pueda precisar cuáles van a ser las variables discriminantes. En principio, contaríamos con una serie de variables, sin que conozcamos las que resultarán más relevantes de cara a diferenciar entre los grupos, y precisamente uno de los resultados que podemos esperar del análisis discriminante es descubrir cuáles son las variables útiles para lograr ese fin. Determinadas variables habrían de ser eliminadas, dada su baja contribución a la discriminación de los grupos. Habrá otras variables que, aun siendo buenos discriminadores, aportan la misma información y resultan redundantes.

Uno de los algoritmos para seleccionar las variables útiles comúnmente usado es el denominado *método stepwise*, o *método paso a paso*, que puede considerarse desde el punto de vista de la selección hacia adelante o hacia atrás. En el *Método de selección paso a paso hacia delante* (*forward*), la primera variable que entra a formar parte del análisis es la que maximiza la separación entre grupos.

A continuación, se forman parejas entre esta variable y las restantes, de modo que encontremos la pareja que produce la mayor discriminación. La variable que contribuye a la mejor pareja es seleccionada en segundo lugar. Con ambas variables, podrían formarse triadas de variables para determinar cuál de éstas resulta más discriminante. De este modo quedaría seleccionada la tercera variable. El proceso continuaría hasta que todas las variables hayan sido seleccionadas o las variables restantes no supongan un suficiente incremento en la capacidad de discriminación.

En el *Método de selección paso a paso hacia atrás* (*backward*), todas las variables son consideradas inicialmente, y van siendo excluidas una a una en cada etapa, eliminando del modelo aquéllas cuya supresión produce el menor descenso en la discriminación entre los grupos. Incluso a veces las direcciones hacia delante y hacia atrás se combinan en la aplicación del método *stepwise*. Se partiría de una selección hacia adelante de variables, aunque revisando tras cada paso el

conjunto de variables resultantes, por si pudiera excluirse alguna de ellas. Esto puede ocurrir cuando la incorporación de una variable supone que alguna de las anteriormente consideradas resulta redundante.

Antes de ser sometidas a cualquier criterio de selección, las variables que van a ser consideradas en un análisis discriminante deben ser revisadas para determinar si satisfacen ciertas condiciones mínimas, sin cuyo cumplimiento habrían de ser descartadas. Del mismo modo, tras la selección de variables, podríamos revisar las que han quedado incluidas para decidir si alguna de ellas debería ser eliminada. Estas condiciones se basan en la *tolerancia de las variables discriminantes* y en los *estadísticos multivariantes parciales F* (*F de entrada y F de salida*), utilizados para garantizar que el incremento de discriminación debido a la variable supera un nivel fijado. Una variable deberá superar las condiciones impuestas en relación a la tolerancia y a *F* de entrada antes de que apliquemos los criterios de selección. Después de ser introducida una variable, habremos de comprobar que todas las seleccionadas hasta ese momento satisfacen la condición fijada para el estadístico *F* de salida. Una variable que inicialmente fue seleccionada, puede ser ahora inadecuada debido a que otras variables introducidas posteriormente aporten la misma contribución a la separación de grupos.

La ***Tolerancia*** es una medida del grado de asociación lineal entre las variables independientes. La tolerancia para una variable no seleccionada es 1- R^2, donde *R* es la correlación múltiple entre esta variable y todas las variables ya incluidas, cuando han sido obtenidas a partir de la matriz de correlaciones intragrupos. Interesan valores altos de la tolerancia.

El ***Estadístico F de entrada*** representa el incremento producido en la discriminación tras la incorporación de una variable respecto al total de discriminación alcanzado por las variables ya introducidas. Una *F* pequeña aconsejaría no seleccionar la variable, pues su aporte a la discriminación de los grupos no sería importante. El estadístico *F* puede ser utilizado para realizar una prueba estadística, que permita determinar la significación del incremento producido en la discriminación. El estadístico se distribuye según *F* con $(g-1)$ y $(n-s-g+1)$ grados de libertad, donde *n* es el número de individuos, *g* el de grupos y *s* el de variables discriminantes.

El ***Estadístico F de salida*** es un estadístico multivariante parcial, que permite valorar el descenso en la discriminación si una variable fuera extraída del conjunto de las ya seleccionadas. Aquellas variables para las cuales el valor de *F* es bajo, podrían ser descartadas antes de proceder a un nuevo paso en el método de selección de variables. El estadístico *F* permitiría llevar a cabo una prueba de significación. Los grados de libertad con que se distribuye *F* son en este caso de $(g-1)$ y $(n-s-g)$. Tras el último paso en la aplicación del método *stepwise*, el estadístico *F* de salida puede ser usado para ordenar las variables seleccionadas de acuerdo con su contribución a la separación de los grupos. Las variables a las que corresponda el valor más alto de *F* serían las que mayor aportación hacen a la discriminación.

Una vez que sabemos que las variables discriminantes cumplen unas condiciones mínimas para ser seleccionadas como tales, aplicaremos ya ***criterios formales de selección***

paso a paso sobre ellas. Hay varios criterios para la selección de variables discriminantes paso a paso. Destacan los siguientes:

Criterio basado en la minimización de la lambda de Wilks. Se selecciona en cada paso la variable que, una vez incorporada a la función discriminante, produce el valor de lambda más pequeño para el conjunto de variables incluidas en la función.

Criterio basado en la V de Rao. Criterio basado en la medida de Rao de la distancia que separa a los grupos. La *V* de Rao también se conoce como traza de Lawley-Hotelling, y para cada paso viene definida por la expresión:

$$V = (n-g)\sum_{i=1}^{p'}\sum_{j=1}^{p'} w_{ij}\sum_{k=1}^{g} n_k \left(\overline{X}_{ik} - \overline{X}_{i}\right)\left(\overline{X}_{jk} - \overline{X}_{j}\right)$$

donde *p*' es el número de variables presentes en el modelo (incluyendo la añadida o suprimida en esa etapa), n_k el tamaño de la muestra en el grupo *k*, el valor w_{ij} corresponde a los elementos de la matriz inversa de covarianzas intragrupos, y las medias presentes en cada uno de los factores del producto representan los valores medios de una variable dentro del grupo *k* y en el grupo global. Los valores *n* y *g* corresponden, como en casos anteriores, al tamaño de la muestra total y al número de grupos. Cuanto mayores sean las diferencias entre los grupos mayor será el valor de *V*. La contribución de una variable al modelo puede evaluarse a partir del incremento que se produce en *V* al ser ésta añadida al modelo. Contando con un suficiente número de grados, *V* se distribuye según *Chi-cuadrado* con $p'(g-1)$ grados de libertad. El cambio producido en *V* tras la adición o supresión de una variable sigue el mismo modelo de distribución, con un número de grados de libertad coincidente con $(g-1)$ veces el número de variables añadidas o suprimidas en cada paso.

Por tanto, tras añadir una variable, podemos contrastar la significación estadística del cambio de un modelo que maximiza las diferencias entre los grupos, pero sin atender a la cohesión interna de los mismos, la cual no se tiene en cuenta en el cálculo de *V*.

Criterio basado en la distancia de Mahalanobis. La distancia de Mahalanobis es una medida de la separación entre dos grupos. De acuerdo con este criterio, mediríamos la distancia de Mahalanobis al cuadrado D^2 entre todos los grupos respecto a las variables incluidas en el modelo, y determinaríamos qué pareja de grupos se encuentran más cercanos (poseen el valor más pequeño para D^2). De las variables que permanecen fuera del modelo, seleccionaríamos para ser incluida aquélla que maximiza D^2 para la pareja de grupos inicialmente más próximos. La expresión de D^2 para el caso de dos grupos *a* y *b* puede escribirse como:

$$D_{ab}^2 = (n-g)\sum_{i=1}^{p'}\sum_{j=1}^{p'} w_{ij}\left(\overline{X}_{ia} - \overline{X}_{ib}\right)\left(\overline{X}_{ja} - X_{jb}\right)$$

donde los elementos incluidos en la expresión analítica tienen el mismo significado que les atribuíamos al hablar de la *V* de Rao, y los factores del producto son las diferencias entre las medias de las variables del modelo para ambos grupos.

Criterio basado en la F intergrupos. A partir de la distancia de Malahanobis es posible calcular un estadístico *F* para medir la diferencia entre dos grupos y contrastar la hipótesis nula de igualdad de medias para ambos. La expresión de este estadístico, en el caso de dos grupos *a* y *b*, es la siguiente:

$$F = \frac{(n. - 1 - p')n_a n_b}{p'(n. - 2)(n_a + n_b)} D_{ab}^2$$

y podría ser usado también como criterio para la selección de variables. En cada paso, seleccionaríamos aquella variable que conduce al mayor valor de *F* en la pareja de grupos que inicialmente resultaban más próximos entre sí. La diferencia con respecto al criterio basado en la distancia de Mahalanobis al cuadrado, radica en que aquí se tienen en cuenta los tamaños de los grupos.

Criterio basado en la varianza residual. Sumando para cada pareja de grupos la varianza residual no explicada por la función discriminante, tendremos una varianza residual total expresada por:

$$R = \sum_{i=1}^{g-1} \sum_{j=i+1}^{g} \frac{4}{4 + D_{a_i b_j}^2}$$

La variable seleccionada en cada paso será aquella que minimiza el total de la varianza no explicada por la función discriminante.

Una vez estimado el modelo discriminante tomando como variables independientes los 69 factores (ninguno resulta expulsado del modelo por los criterios de selección de variabes discriminantes), se obtienen los coeficientes de las dos funciones discriminantes de Fisher.

D0 = -1,151 - 0,381C_1 +-------- - 0,057C_{67} + 0,003C_{68} - 0,011C_{69}

D1 = -0,856 + 0,227C_1 +-------- + 0,034C_{67} - 0,002C_{68} + 0,007C_{69}

REGR factor score 60 for analysis 1	-,063	,038
REGR factor score 61 for analysis 1	-,005	,003
REGR factor score 62 for analysis 1	-,004	,003
REGR factor score 63 for analysis 1	,008	-,005
REGR factor score 64 for analysis 1	-,036	,021
REGR factor score 65 for analysis 1	-,019	,011
REGR factor score 66 for analysis 1	-,045	,027
REGR factor score 67 for analysis 1	-,057	,034
REGR factor score 68 for analysis 1	,003	-,002
REGR factor score 69 for analysis 1	-,011	,007
(Constante)	-1,151	-,856

Funciones discriminantes lineales de Fisher

6. INTERPRETACIÓN DE LAS FUNCIONES DISCRIMINANTES. CLASIFICACIÓN DE LOS INDIVIDUOS

Halladas las funciones discriminantes, y fijado el número de ellas que se retiene, es necesario interpretar el significado de las mismas.

El análisis discriminante, decíamos en las primeras páginas, puede ser utilizado con dos finalidades básicas: interpretar las diferencias existentes entre varios grupos o pronosticar la clasificación de los sujetos. Para el investigador interesado en obtener una regla de decisión que permita clasificar nuevos casos, el número de dimensiones consideradas en el espacio discriminante y su significado posiblemente no atraigan su atención. Puede ser más interesante la utilización de las funciones discriminantes para pronosticar el grupo al que quedará adscrito un nuevo caso no contemplado al extraer las funciones. ***Un primer criterio sería clasificar al individuo en el grupo para el que su función discriminante, aplicada en los valores de las variables independientes del individuo concreto (puntuación discriminante), tiene un valor mayor (no olvidemos que hay tantas funciones discriminantes como grupos en la variable dependiente, en nuestro caso 2)***. Este procedimiento de clasificación resulta muy sensible a la violación del supuesto de igualdad de matrices de varianzas-covarianzas. Cuando no se verifica dicho supuesto, los casos tienden a ser clasificados en el grupo en el que se registra la mayor dispersión.

En realidad, la clasificación de un sujeto podría hacerse a partir de sus valores en las variables discriminantes o en las funciones discriminantes. En el primer caso, no podríamos hablar propiamente de un análisis discriminante, pues no es necesario el cálculo de las funciones discriminantes, sino la utilización de funciones de clasificación. Uno y otro tipo de funciones sirven al mismo objetivo, pero la clasificación a partir de las funciones discriminantes

es más cómoda y suele llevar a mejores resultados en la mayoría de los casos. Los diferentes procedimientos usados para la clasificación se basan en la comparación de un caso con los centroides de grupo, a fin de ver a cuál de ellos resulta más próximo.

Un procedimiento alternativo para la clasificación de un caso se basa en el cálculo de su distancia a los centroides de cada uno de los grupos o ***funciones de distancia generalizada***. El caso sería adscrito a aquel grupo con cuyo centroide existe una menor distancia. La distancia de Mahalanobis es una medida adecuada para valorar la proximidad entre casos y centroides. Un caso será clasificado en el grupo respecto al cual presenta la distancia más pequeña. Ello significaría que a ese grupo corresponde el centroide cuyo perfil sobre las variables discriminantes resulta más parecido al perfil del caso.

Otro de los procedimientos seguidos para asignar un caso a uno de los grupos es utilizar las ***probabilidades de pertenencia al grupo***. Un caso se clasifica en el grupo al que su pertenencia resulta más probable. El cálculo de probabilidad de pertenencia a un grupo asume que todos los grupos tienen un tamaño similar. No se tiene en cuenta que a priori es posible anticipar una mayor probabilidad de pertenencia a un determinado grupo cuando en la población el porcentaje de sujetos que pertenece a cada grupo es muy diferente. En tal situación, conviene incorporar al cálculo las ***probabilidades a priori***, con lo que se consigue mejorar la predicción final y reducir lo errores de clasificación. De acuerdo con este planteamiento, la regla de Bayes sería útil para calcular la probabilidad posterior de pertenencia del caso a un grupo (***probabilidad a posteriori***), conocida la probabilidad a priori para el mismo. Un caso será clasificado en el grupo en el que su pertenencia cuenta con una mayor probabilidad a posteriori. Podría ocurrir que dos casos que son clasificados en el mismo grupo tengan probabilidades bastante diferentes, o que las probabilidades de que un sujeto pertenezca a dos grupos distintos no sean muy diferentes entre sí, en cuyo caso, aun asignándolo a la clase en la que cuenta con mayor probabilidad, su clasificación no sería tan clara. Por ese motivo, resulta interesante conocer para cada individuo no sólo la ***máxima probabilidad***, sino también las probabilidades de pertenecer a otros grupos.

La probabilidad de pertenecía de un individuo a un grupo *i* de la variable dependiente se evalúa mediante:

$$P_i = \frac{e^{F_i}}{\sum_i e^{F_i}}$$

F_i son las puntuaciones de las funciones discriminantes en el grupo *i*.

Si se utilizan propiedades a priori π_i diferentes de pertenecía a los grupos, la probabilidad anterior tiene la siguiente expresión:

$$P_i = \frac{\pi_i e^{F_i}}{\sum_i \pi_i e^{F_i}}$$

Un procedimiento muy útil para la representación gráfica de la clasificación de casos es el ***mapa territorial***, que consiste en situar en el eje horizontal y en el vertical dos funciones discriminantes (o variables discriminantes) y separar en el plano resultante, por medio de

líneas, las zonas o territorios que ocuparían los sujetos clasificados en cada grupo. Lógicamente, cuando el número de funciones es mayor que dos, el plano no es suficiente para representar todas las dimensiones del espacio discriminante. En ese caso suelen representarse únicamente las dos primeras, que son las que en mayor medida contribuyen a la separación de los grupos. El problema del número de dimensiones en la representación se agrava cuando en la clasificación trabajamos con las variables y no con las funciones discriminantes. Es una razón más para preferir procedimientos de clasificación basados en estas últimas. No obstante, cuando sólo contamos con una función discriminante, la representación del mapa territorial se hará sobre una línea, y no en un plano. Cuando los casos o individuos están bien clasificados, su representación sobre el plano formado por las dos funciones les situaría en el territorio correspondiente al grupo. En cambio, cuando la discriminación es débil, puede haber un cierto número de sujetos que caen fuera del territorio que serían casos mal clasificados. Las líneas que constituyen las fronteras entre el territorio ocupado por los diferentes grupos se determinan a partir de la posición de los centroides. Para el caso de dos grupos, la línea divisoria sería la mediatriz del segmento que une a los dos respectivos centroides, siempre y cuando las matrices de covarianza de los grupos sean idénticas. Si no fuera así, la línea estaría más próxima al centroide correspondiente al grupo con menor varianza. Si existen más de dos grupos, el trazado de las líneas se complica.

En nuestro caso calculamos para cada individuo de la muestra el grupo de la variable dependiente al que pertenece (Dis_1 = 0 significa no defraudador y Dis_1 = 1 significa defraudador). También calculamos las probabilidades de pertenecía de cada individuo al grupo de la variable dependiente en que se clasifica (Dist1_2 da la probabilidad de pertenencia al grupo de no defraudadores y Dist2_2 da la probabilidad de pertenencia al grupo de defraudadores).

Dis_1	Dis1_1	Dis1_2	Dis2_2
,00	-1,03551	,78351	,21649
,00	-1,03624	,78370	,21630
,00	-1,35129	,85426	,14574
,00	-,88620	,74236	,25764
,00	-,65831	,67047	,32953
1,00	-,17233	,49208	,50792
,00	-,24188	,51862	,48138
,00	-,96414	,76446	,23554
,00	-1,28954	,84212	,15788
,00	-,80430	,71773	,28227
,00	-,95725	,76256	,23744
,00	-,40250	,57926	,42074
,00	-1,04950	,78711	,21289
,00	-,62165	,65799	,34201
,00	-,58082	,64383	,35617
,00	-1,08126	,79512	,20488
,00	-,87810	,73999	,26001
,00	-,90870	,74888	,25112
,00	-,47127	,60462	,39538
,00	-1,35286	,85456	,14544
,00	-,82113	,72291	,27709
,00	-,93036	,75504	,24496
1,00	,82297	,17489	,82511
,00	-1,43980	,87029	,12971
,00	-1,04035	,78476	,21524

7. EVALUACIÓN DEL MODELO DISCRIMINANTE. BONDAD DEL AJUSTE

Una forma de valorar la bondad de la clasificación de los individuos realizada es aplicar el procedimiento a los casos para los que conocemos su grupo de adscripción, y comprobar si coinciden el grupo predicho y el grupo observado. El porcentaje de casos correctamente clasificados indicaría la corrección del procedimiento. La ***matriz de clasificación***, también denominada ***matriz de confusión***, permite presentar para los casos observados en un grupo, cuántos de ellos se esperaban en ese grupo y cuántos en los restantes. De esta forma, resulta fácil constatar qué tipo de errores de clasificación se producen. La estructura de la matriz de clasificación sería la mostrada en la Figura 5-17, donde cada valor n_{ij} representa el número de casos del grupo *i* que tras aplicar las reglas de clasificación son adscritos al grupo *j*. Los valores situados en la diagonal descendente constituyen, por tanto, el número de casos que han sido correctamente clasificados.

En la matriz de clasificación, es frecuente encontrar estos valores en forma de porcentajes. Si el porcentaje de casos correctamente clasificados es alto, cabe esperar que las funciones discriminantes también proporcionen buenos resultados a la hora de predecir el grupo al que se adscribirá cualquier nuevo sujeto perteneciente a la misma población de donde fue extraída la muestra. Este porcentaje puede ser tomado como una medida no sólo

de la bondad de la clasificación, sino también de las diferencias existentes entre los grupos; si la clasificación es buena se deberá a que las variables discriminantes permiten diferenciar entre los grupos. En nuestro caso se observa que se clasifican bien el 79,4% de los individuos muestrales.

Resultados de la clasificación[a]

			Grupo de pertenencia pronosticado		
		marca	,00	1,00	Total
Original	Recuento	,00	655317	65054	720371
		1,00	332977	875146	1208123
	%	,00	91,0	9,0	100,0
		1,00	27,6	72,4	100,0

a. Clasificados correctamente el 79,4% de los casos agrupados originales.

7. ASIGNACIÓN DE PROBABILIODAD DE FRAUDE Y GRUPO DE PERTENECIA PARA UN INDIVIDUO CUALQUIERA DE LA POBLACIÓN DE DECLARANTES

Obtenidas las funciones discriminantes, se puede asignar para un individuo futuro el grupo de riesgo de fraude al que pertenece. Bastará evaluar las funciones discriminantes para los valores de sus variables independientes y observar qué función discriminante tiene un mayor valor. El individuo se clasificará en el grupo de riesgo para el que su función discriminante tiene un mayor valor.

Tenemos que tener presente que las funciones discriminantes se calculan a partir de factores y que los valores de los factores se calculan a partir de sus combinaciones lineales en función de las variables iniciales.

Por lo tanto, primero habrá que calcular los factores Ci para la vriables Xi conocidas para el individuo:

C1 =0,071X1+0,03X2+0,051X3+0,026x4+…….

C2=-0,009X1-0,04X2-0,08X3.0,001 X4 +……..

C3=-0,007X1-0,003 X2-0,007X3-0,002X4+……..

……

……

Conocidos los factores, se calculan ya las funciones discriminantes para ese individuo

D0 = -1,151 - 0,381C_1 +-------- - 0,057C_{67} + 0,003C_{68} - 0,011C_{69}

D1 = -0,856 + 0,227C_1 +-------- + 0,034C_{67} - 0,002C_{68} + 0,007C_{69}

Finalmente se puede calcular la probabilidad de que le individuo no defraude y defraude mediante las expresiones:

$$P_0 = \frac{e^{D_0}}{e^{D_0} + e^{D_1}}$$

$$P_1 = \frac{e^{D_1}}{e^{D_0} + e^{D_1}}$$

BIBLIOGRAFÍA

Onrubia J., Picos F. y Pérez C. Panel de declarantes de IRPF 1999-2007: diseño, metodología y guía de utilización. Instituto de Estudios Fiscales – 2011

Pérez C. - Técnicas de Análisis multivariante de datos - Pearson Prentice Hall – 2004

Pérez C. – Técnicas de análisis multivariante con SPSS- Garceta Editorial – 2009

Pérez C. – Técnicas de muestreo estadístico – Garceta Editorial – 2010

Pérez C. – El sistema Estadístico SAS – Garceta Editorial – 2011

Pérez C. – Técnicas de segmentación. Conceptos, herramienta y aplicaciones – Garceta Editorial 2011

Pérez C. y Santín D. - Minería de datos. Técnicas y herramientas- Thomson – 2007

Pérez C. y Santín D. –Data Mining. Soluciones con Enterprise Miner- RA-MA – 2006

Pérez C., Burgos J., Huete S. y Gallego C. – La muestra de declarantes de IRPF 2009. Documento de trabajo número 11 de 2012 del Instituto de Estudios Fiscales.

Valdés T. Los métodos del análisis discriminante como herramienta la servicio de la inspección pública. Monografía nº 21 de Instituto de Estudios Fiscales