LA INVESTIGACIÓN DEL FRAUDE FISCAL. APLICACIONES CON LAS MUESTRAS Y PANELES DE IRPF DEL IEF Y LOS MODELOS DE REGRESION LOGISTICA

César Pérez López

Instituto de Estudios Fiscales
Universidad Complutense
cesar.perez@ief.minhap.es

ABSTRACT

Este papel de trabajo tiene como finalidad presentar las posibilidades de análisis que se abren en el campo de la investigación del fraude fiscal 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 de la regresión logística aplicada a las muestras y paneles de IRPF del IEF con la finalidad de detectar posible fraude fiscal en este impuesto. A partir de una muestra anual de IRPF, que el Instituto de Estudios Fiscales y la Agencia Tributaria ponen cada año a disposición de todos los investigadores y analistas de modo gratuito, se buscará un modelo logístico 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 REGRESIÓN LOGÍSTICA COMO TÉCNICA DE CLASIFICACIÓN Y SEGMENTACIÓN

La regresión logística puede enfocarse como 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. La regresión logística se puede interpretar como un modelo predictivo de 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 de la regresión logística 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 de la regresión logística 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 logístico 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.

2. EL MODELO DE REGRESIÓN LOGÍSTICA COMO UN MODELO DE ELECCIÓN DISCRETA. IDENTIFICACIÓN, ESTIMACIÓN Y DIAGNOSIS

La expresión funcional del modelo de análisis de la regresión múltiple es $y = F(x_1, x_2, \cdots, x_n)$. La regresión múltiple admite la posibilidad de trabajar con variables dependientes discretas en vez de continuas para permitir la modelización de fenómenos discretos. Cuando la variable dependiente es una variable discreta que refleja decisiones individuales en las que el conjunto de elección está formado por alternativas separadas y mutuamente excluyentes estamos ante los ***modelos de elección discreta***. Cuando la variable dependiente es discreta y toma sólo un número pequeño de valores no tiene sentido tratarla como si fuera una variable continua y suele interesar *caracterizar la probabilidad de que un agente tome una determinada decisión discreta*, condicional a los valores de ciertas variables explicativas. Estas funciones de distribución que caracterizan probabilidades para cada valor de las variables explicativas suelen ser no lineales y no suelen tener solución analítica por lo que suele ser necesario recurrir a métodos numéricos.

Los modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles se llaman *modelos de elección binaria*. Cuando el conjunto de elección tiene varios valores discretos nos encontramos ante los *modelos de elección múltiple o modelos multinomiales*.

Partimos de un modelo de regresión del tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

una de cuyas hipótesis es:

$$E(\varepsilon | X_1, X_2, ..., X_k) = 0$$

lo que nos lleva a escribir el modelo como:

$$E(Y | X_1, ..., X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Pero en el caso de los modelos de elección discreta en los que el conjunto de elección tiene sólo dos alternativas posibles mutuamente excluyentes, *Y* es una variable aleatoria de Bernouilli de parámetro *p*, lo que nos permite escribir:

$$E(Y|X_1,...,X_k) = P(Y=1|X_1,...,X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Estamos ahora ante el ***modelo lineal de probabilidad***, donde, por ejemplo, β_1 mide la variación en la probabilidad de "éxito" (*Y* = 1) ante una variación unitaria en X_1 (con todo lo demás constante).

Realizada la estimación del modelo lineal de probabilidad tenemos que:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k = \hat{P}$$

se puede interpretar como una estimación de la probabilidad de "éxito" (de que *Y* = 1). En algunas aplicaciones tiene sentido interpretar $\hat{\beta}_0$ como la probabilidad de éxito cuando todas las X_j valen 0. Una limitación importante del modelo lineal de probabilidad es que para ciertas combinaciones de las variables explicativas $X_1,...,X_k$, las probabilidades estimadas pueden ser mayores que cero o menores que uno. También es un problema en el modelo lineal de probabilidad la presencia de heteroscedasticidad,

Ante esta circunstancia consideramos:

$$P(Y=1|X_1, X_2,..., X_k) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)$$

que, para evitar los problemas del modelo lineal de probabilidad, se especifican como *Y* = *G*(*Xβ*), donde *G* es una función que toma valores estrictamente entre 0 y 1 (0<*G*(*Z*)<1) para todos los números reales *z*. El problema se soluciona tomando G como una función de distribución de una variable aleatoria, que por ser una probabilidad estará siempre entre cero y uno. Según las diferentes definiciones de *G* tenemos los distintos modelos de elección binaria.

Si $G(z) = \dfrac{e^z}{1+e^z}$ estamos ante el ***modelo Logit***, cuya expresión será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}$$

En el caso del ***modelo Probit*** tenemos:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)\mathbf{d}v$$

donde $\Phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$ es la función de densidad de la *normal* (0,1).

La expresión del modelo Probit será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \int_{-\infty}^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \frac{1}{\sqrt{2\pi}} e^{\frac{-\nu^2}{2}} d\nu$$

Los modelos Probit y Logit, como son modelos no lineales, no podremos estimar por MCO y tendremos que emplear métodos de máxima verosimilitud.

Supongamos que tenemos *n* observaciones idéntica e independientemente distribuidas (muestra aleatoria) que siguen el modelo:

$$P(Y = 1|\mathbf{X}) = G(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

Para obtener el estimador de máxima verosimilitud (MV), condicionado a las variables explicativas, necesitamos la función de verosimilitud:

$$L(\beta) = \prod_{Y_i=1} P_i \prod_{Y_i=0} (1 - P_i) = \prod_{i=1}^{n} G(X_i'\beta)^{Y_i} \left(1 - G(X_i'\beta)\right)^{1-Y_i}$$

con:

$$P_i = P(Y_i = 1|X_{1i}, ..., X_{ki}) = G(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}) = G(X_i'\beta)$$

El estimador de MV de β es el que maximiza el logaritmo de la función de verosimilitud:

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \left[Y_i \ln G(X_i'\beta) + (1 - Y_i) \ln\left(1 - G(X_i'\beta)\right) \right]$$

que será un estimador consistente, asintóticamente normal y asintóticamente eficiente.

Las condiciones de primer orden serán:

$$S(\beta) = \sum_{i=1}^{n} \left[\frac{Y_i}{G(X_i'\beta)} - \frac{(1 - Y_i)}{\left(1 - G(X_i'\beta)\right)} \right] X_i g(X_i'\beta) =$$

$$= \sum_{i=1}^{n} \left[\frac{Y_i - G(X_i'\beta)}{G(X_i'\beta)\left(1 - G(X_i'\beta)\right)} \right] X_i g(X_i'\beta) = 0$$

donde *g*(.) es la función de densidad de la normal o la logística (derivada de la función de distribución).

La no linealidad del problema hace que para obtener el estimador MV de β necesitemos aplicar un algoritmo iterativo y obtener el estimador por métodos numéricos iterativos. Mediante el algoritmo Scoring tenemos:

$$\hat{\beta}^{k+1} = \hat{\beta}^{k} + \left[I(\hat{\beta}^{k}) \right]^{-1} S(\hat{\beta}^{k})$$

La matriz de covarianzas asintótica de $\hat{\beta}$ se estima como:

$$A\hat{\text{v}}\text{ar}(\hat{\beta}) = \left[I(\hat{\beta}) \right]^{-1} = \left(\sum_{i=1}^{n} \frac{\left[g(X_i'\hat{\beta}) \right]^2 X_i X_i'}{G(X_i'\hat{\beta})\left(1 - G(X_i'\hat{\beta})\right)} \right)^{-1}$$

Para realizar ***contrastes de hipótesis (diagnosis) en los modelos Logit y Probit*** tendremos en cuenta que la raíz cuadrada de los elementos de la diagonal principal de la matriz de covarianzas asintótica son los errores estándar (asintóticos) de cada uno de los $\hat{\beta}_j$, que los podemos emplear para construir los estadísticos *t* (que tendrán una distribución asintótica normal) o intervalos de confianza aproximados para cada parámetro. También podemos contrastar varias restricciones simultáneamente. Lo habitual es que lo que nos interese sean restricciones de exclusión por lo que es en lo que nos vamos a centrar.

Para contrastar la hipótesis nula de que un conjunto de parámetros es igual a cero podemos emplear varios procedimientos:

- *Estadístico de Wald*: se distribuye asintóticamente como una *Chi-cuadrado* con *q* (nº de restricciones) grados de libertad y lo proporcionan la mayoría de los programas.

- *Contraste de razón de verosimilitudes* (*Likelihood Ratio* (*LR*) *test*): Se basa en la diferencia entre el logaritmo de la función de verosimilitud en el modelo sin restringir y en el restringido:

$$LR = 2\left(l(\hat{\beta}_{NR}) - l(\hat{\beta}_{R}) \right)$$

que se distribuye asintóticamente como una *Chi-cuadrado* con *q* grados de libertad.

En cuanto a las ***medidas de la bondad de ajuste en los modelos Logit y Probit*** tenemos:

- *Porcentaje de predicciones correctas*: Para cada *i* calculamos la probabilidad estimada de que $Y_i = 1$:

$$\hat{P}_i = \hat{P}(Y_i = 1 | X_{1i}, \ldots, X_{ki}) = G(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki})$$

- Si $\hat{P}_i > 0{,}5$ nuestra predicción será que Y_i es 1 y si $\hat{P}_i \leq 0{,}5$ nuestra predicción será que Y_i es 0. El % de veces en que el valor de Y_i observado coincida con la predicción es el % de predicciones correctas. Lo interesante es calcular por separado el % de predicciones correctas de ceros y de unos.

- *Pseudo – R^2* (*de McFadden*)*:* Está basado en el logaritmo de la función de verosimilitud:

$$Pseudo - R^2 = 1 - \frac{l(\hat{\beta})}{l(\hat{\beta}_0)}$$

donde $l(\hat{\beta})$ es el logaritmo de la función de verosimilitud para el modelo estimado y $l(\hat{\beta}_0)$ el de un modelo sólo con término constante. Como $|l(\hat{\beta})| < |l(\hat{\beta}_0)|$, el valor *Pseudo – R^2* está entre 0 y 1.

- *Criterios de Información*: Son medidas que tratan de buscar un equilibrio entre la bondad del ajuste, medida en base al valor del logaritmo de la función de verosimilitud, y una especificación parsimoniosa del modelo (Ejemplos: *Akaike* (AIC), *Schwarz* (SC) y *Hannan-Quinn* (HQ)). Se escoge el modelo con menor valor del criterio de información.

A la hora de ***interpretar las estimaciones en los modelos Probit y Logit***, generalmente lo que nos interesa es conocer el efecto de variaciones en una variable X_j sobre la probabilidad de respuesta, que si la variable es continua será:

$$\Delta\hat{P}(Y = 1|\mathbf{X}) \approx \left[g(\mathbf{X}\hat{\beta})\hat{\beta}_j \right] \Delta X_j$$

Como $g(X\hat{\beta})$ depende de *X* habrá que calcular los efectos parciales para valores interesantes de *X* (las medias muestrales, valores máximos y mínimos de las variables de interés, etc.). También se puede calcular el efecto parcial para cada individuo y después calcular su media.

3. ANÁLISIS EXPLORATORIO DE LOS DATOS PREVIO A LA APLICACIÓN DEL MODELO

Para comenzar, la aplicación del análisis logístico requiere que contemos con un conjunto de regresores (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, los regresores 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 regresores, 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 regresor 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 la función logística. 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, el modelo logístico se optimiza par regresores normales y poco correlados entre sí para garantizar la ausencia de multicolinealidad. 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 regresoras 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***.

La matriz de correlaciones de las variables suele utilizarse para detectar la *multicolinealidad* (variables con correlación muy alta pueden ser redundantes), que puede ser muy nociva en la estimación del modelo logístico. ***En nuestro caso, estos problemas también desaparecen al utilizar factores en vez de variables iniciales como variables independientes del modelo logísticoe. 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, logística, 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:

$$C_1 = u_1X = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p$$

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:

$$C_2 = u_2X = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p$$

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 $\lambda h/p$. 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 caracterísica 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:

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

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 64 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 64 variables independientes. 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 LOGÍSTICO

Realizada la estimación del modelo logístico con las componentes como regresores se obtienen resultados óptimos. Se observa una alta significatividad conjunta de los parámetros estimados. Asimismo la significatividad individual es también muy aceptable ya que solo los factores 21, 22, 50 y 56 presentan valores demasiado altos. Es de resaltar que en una regresión logística con tantos regresores presenta una diagnosis óptima. La razón se apoya en las propiedades óptimas de las componentes principales.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Ratio de verosim	1533955.52	64	<.0001
Puntuación	673538.330	64	<.0001
Wald	257073.397	64	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.1812	0.0284	104456.410	<.0001
FAC1_1	1	-38.4853	0.2520	23316.7343	<.0001
FAC2_1	1	2.1462	0.0288	5562.8002	<.0001
FAC3_1	1	1.9205	0.0243	6251.7755	<.0001
FAC4_1	1	-0.5515	0.0360	234.2139	<.0001
FAC5_1	1	-8.7398	0.0566	23851.6826	<.0001
FAC6_1	1	0.8963	0.0157	3260.9900	<.0001
FAC7_1	1	-52.4940	0.4374	14405.5666	<.0001
FAC8_1	1	0.1511	0.0230	43.0982	<.0001
FAC9_1	1	-0.2207	0.0364	36.8578	<.0001
FAC10_1	1	-0.3283	0.0194	285.3028	<.0001
FAC11_1	1	-7.2508	0.2738	701.3429	<.0001
FAC12_1	1	-0.1542	0.00423	1326.0459	<.0001
FAC13_1	1	-1.7378	0.0406	1832.1360	<.0001
FAC14_1	1	-0.5946	0.00364	26619.1787	<.0001
FAC15_1	1	-1.6134	0.2526	40.7836	<.0001
FAC16_1	1	-0.5208	0.0155	1123.3355	<.0001
FAC17_1	1	-0.1458	0.0146	100.1208	<.0001
FAC18_1	1	0.0483	0.00949	25.9369	<.0001
FAC19_1	1	-9.0904	0.0420	46880.3818	<.0001
FAC20_1	1	0.0236	0.00742	10.1609	0.0014
FAC21_1	1	0.00844	0.00852	0.9827	0.3215
FAC22_1	1	-0.00456	0.00583	0.6133	0.4335
FAC23_1	1	0.0659	0.00233	801.1594	<.0001
FAC24_1	1	1.6560	0.0340	2369.5671	<.0001
FAC25_1	1	-13.5757	0.0445	93260.0842	<.0001
FAC26_1	1	-0.1859	0.0212	77.0530	<.0001
FAC27_1	1	0.6515	0.0203	1032.0140	<.0001
FAC28_1	1	0.1183	0.00259	2088.3052	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
FAC29_1	1	-1.6895	0.0328	2661.0973	<.0001
FAC30_1	1	-59.7531	1.2764	2191.4384	<.0001
FAC31_1	1	-0.4683	0.0155	915.7232	<.0001
FAC32_1	1	-1.9949	0.0124	25768.7471	<.0001
FAC33_1	1	-1.0171	0.0277	1344.0732	<.0001
FAC34_1	1	-0.1232	0.00906	184.8458	<.0001
FAC35_1	1	-5.4618	0.2295	566.2158	<.0001
FAC36_1	1	-0.0803	0.0168	22.8948	<.0001
FAC37_1	1	-0.0175	0.0134	1.7067	0.1914
FAC38_1	1	0.0246	0.00374	43.2868	<.0001
FAC39_1	1	0.0830	0.0254	10.6532	0.0011
FAC40_1	1	-5.6587	0.0668	7184.0836	<.0001
FAC41_1	1	-0.0841	0.0109	59.8219	<.0001
FAC42_1	1	0.0291	0.0177	2.7195	0.0991
FAC43_1	1	0.1033	0.0229	20.3119	<.0001
FAC44_1	1	-0.1752	0.0141	155.2633	<.0001
FAC45_1	1	0.2148	0.0120	318.1266	<.0001
FAC46_1	1	-1.1791	0.00502	55270.7252	<.0001
FAC47_1	1	-0.5285	0.0212	620.7570	<.0001
FAC48_1	1	-0.0377	0.0153	6.0653	0.0138
FAC49_1	1	-0.0252	0.0141	3.1626	0.0753
FAC50_1	1	0.0401	0.0365	1.2079	0.2717
FAC51_1	1	-0.0547	0.00352	241.0795	<.0001
FAC52_1	1	-0.0218	0.0120	3.2918	0.0696
FAC53_1	1	0.2008	0.0131	233.5738	<.0001
FAC54_1	1	0.0293	0.00820	12.7963	0.0003
FAC55_1	1	0.1685	0.00282	3576.2727	<.0001
FAC56_1	1	0.0382	0.0426	0.8043	0.3698
FAC57_1	1	-0.2356	0.0175	180.1805	<.0001
FAC58_1	1	0.0720	0.0143	25.3027	<.0001
FAC59_1	1	0.5584	0.0308	327.7969	<.0001
FAC60_1	1	-0.0996	0.0374	7.0892	0.0078
FAC61_1	1	-1.0455	0.0227	2112.1564	<.0001
FAC62_1	1	-0.0425	0.0101	17.5808	<.0001
FAC63_1	1	-0.1050	0.00851	152.2248	<.0001
FAC64_1	1	-1.1932	0.0782	232.6527	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
FAC1_1	<0.001	<0.001	<0.001
FAC2_1	8.552	8.083	9.048
FAC3_1	6.824	6.507	7.157
FAC4_1	0.576	0.537	0.618
FAC5_1	<0.001	<0.001	<0.001
FAC6_1	2.451	2.376	2.527
FAC7_1	<0.001	<0.001	<0.001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
FAC8_1	1.163	1.112	1.217
FAC9_1	0.802	0.747	0.861
FAC10_1	0.720	0.693	0.748
FAC11_1	<0.001	<0.001	0.001
FAC12_1	0.857	0.850	0.864
FAC13_1	0.176	0.162	0.190
FAC14_1	0.552	0.548	0.556
FAC15_1	0.199	0.121	0.327
FAC16_1	0.594	0.576	0.612
FAC17_1	0.864	0.840	0.889
FAC18_1	1.050	1.030	1.069
FAC19_1	<0.001	<0.001	<0.001
FAC20_1	1.024	1.009	1.039
FAC21_1	1.008	0.992	1.025
FAC22_1	0.995	0.984	1.007
FAC23_1	1.068	1.063	1.073
FAC24_1	5.239	4.901	5.600
FAC25_1	<0.001	<0.001	<0.001
FAC26_1	0.830	0.797	0.866
FAC27_1	1.918	1.844	1.996
FAC28_1	1.126	1.120	1.131
FAC29_1	0.185	0.173	0.197
FAC30_1	<0.001	<0.001	<0.001
FAC31_1	0.626	0.607	0.645
FAC32_1	0.136	0.133	0.139
FAC33_1	0.362	0.342	0.382
FAC34_1	0.884	0.869	0.900
FAC35_1	0.004	0.003	0.007
FAC36_1	0.923	0.893	0.954
FAC37_1	0.983	0.957	1.009
FAC38_1	1.025	1.017	1.032
FAC39_1	1.087	1.034	1.142
FAC40_1	0.003	0.003	0.004
FAC41_1	0.919	0.900	0.939
FAC42_1	1.030	0.995	1.066
FAC43_1	1.109	1.060	1.160
FAC44_1	0.839	0.816	0.863
FAC45_1	1.240	1.211	1.269
FAC46_1	0.308	0.305	0.311
FAC47_1	0.590	0.565	0.615
FAC48_1	0.963	0.935	0.992
FAC49_1	0.975	0.948	1.003
FAC50_1	1.041	0.969	1.118
FAC51_1	0.947	0.940	0.953
FAC52_1	0.978	0.956	1.002
FAC53_1	1.222	1.191	1.254

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
FAC54_1	1.030	1.013	1.046
FAC55_1	1.184	1.177	1.190
FAC56_1	1.039	0.956	1.130
FAC57_1	0.790	0.763	0.818
FAC58_1	1.075	1.045	1.105
FAC59_1	1.748	1.645	1.857
FAC60_1	0.905	0.841	0.974
FAC61_1	0.352	0.336	0.368
FAC62_1	0.958	0.939	0.978
FAC63_1	0.900	0.885	0.915
FAC64_1	0.303	0.260	0.353

Association of Predicted Probabilities and Observed Responses			
Concordancia de porcentaje	95.1	**D de Somers**	0.904
Discordancia de porcentaje	4.8	**Gamma**	0.905
Porcentaje ligado	0.1	**Tau-a**	0.423
Pares	870296773633	**c**	0.952

En la última tabla de la salida se observa la matriz de confusión que presenta un porcentaje de aciertos del 95,1%

Una vez estimado el modelo logístico tomando como variables independientes los 64 factores, se obtienen los coeficientes de la función logística.

6. INTERPRETACIÓN DE LAS FUNCION LOGISTICA. CLASIFICACIÓN DE LOS INDIVIDUOS

Hallada la función logística estimada, podemos calcular la probabilidad de fraude de un nuevo individuo sustituyendo los valores de sus variables regresoras en la función logística estimada.

Tenemos que tener presente que la función logística se calcula 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 calcula ya la función logística para ese individuo

6. CONCLUSIONES

A la vista de los resultados de esta investigación, se puede concluir que el uso de modelos predictivos planteados adecuadamente puede aportar herramientas muy eficientes para la investigación del fraude fiscal.

La disponibilidad actual de datos sobre impuestos accesibles a los investigadores, fiables y bien diseñados abre nuevas líneas de investigación en el campo del fraude que pueden tener resultados óptimos.

De esta forma se pueden complementar las herramientas actuales de análisis del fraude basadas esencialmente en filtros (Zújar y otras) con otras herramientas basadas en modelos predictivos cuya diagnosis las hacen muy fiables.

El desarrollo actual de la tecnología facilita la aplicación de estos tipos de modelización al análisis del fraude porque además de implementar los algoritmos matemáticos eliminan los problemas de capacidad de movimiento de grandes volúmenes de información. Las actuales herramientas de Minería de Datos y Big Data son un gran avance para este tipo de investigaciones.

BIBLIOGRAFÍA

Ferro Veiga J.M. Delito fiscal y blanqueo de capital – Alcalá Grupo editorial - 2013

Hinojosa Torralbo J.J. Medidas y procedimientos contra el fraude fiscal – Atelier - 2013

Peláez Martos J.M. Fraude fiscal, blanqueo de capitales y corrupción en el sector inmobiliario – Cisspraxis. 2009

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.