

REGRESIÓN LOGÍSTICA APLICADA A LOS IMPUESTOS SOBRE LA RENTA

Profesor: César Pérez López

Alumno: José Manuel Mendoza Gómez

Práctica en SPSS

- Introducción

¿Qué es la regresión logística?

La regresión logística es un procedimiento cuantitativo de gran utilidad para problemas donde la variable dependiente toma valores en un conjunto finito. A continuación, desarrollaremos el caso especial en que la variable dependiente o respuesta es dicotómica.

La intención de este proyecto es averiguar, con la ayuda de las variables independientes, si el individuo investigado es potencialmente defraudador (*ya que el nivel de confianza nunca es el 100%. siempre hay que darle el beneficio de la duda*) o de lo contrario, no.

- BBDD

	Nombre	Tipo	Anchura	Decimales	Etiqueta
1	id	Numérico	8	2	Identificador del perceptor
2	f_tmg	Numérico	8	2	
3	f_capinm	Numérico	8	2	
4	f_nhijos	Numérico	8	2	
5	f_aaee	Numérico	8	2	
6	f_planp	Numérico	8	2	
7	f_gastos	Numérico	8	2	
8	marca	Numérico	8	2	
9	factor	Numérico	8	2	Factor de elevación de la muestra
10	cdpost	Numérico	8	2	Código postal
11	estcv	Cadena	3	0	Estado civil de declarante
12	sexo	Cadena	3	0	Sexo del declarante
13	dec	Cadena	3	0	Tipo de declaración
14	prov	Numérico	2	0	Provincia
15	ejnacd	Cadena	12	0	Ejercicio de nacimiento del declarante
16	ejnacc	Cadena	12	0	Ejercicio de nacimiento del cónyuge
17	minusd	Cadena	9	0	Grado de minusvalía del declarante
18	minusc	Cadena	9	0	Grado de minusvalía del cónyuge
19	nmdesc	Numérico	8	2	Número total de descendientes
20	nmdesc0	Numérico	8	2	Número de descendientes <3 años
21	nmdesc3	Numérico	8	2	Número de descendientes >= 3 y < 16 años
22	nmdesc16	Numérico	8	2	Número de descendientes >= 16 y < 18 años
23	nmdesc18	Numérico	8	2	Número de descendientes >= 18 y < 25 años
24	nmdescr	Numérico	8	2	Número de descendientes >=25 años
25	nmdescd	Numérico	8	2	Número de descendientes con edad desconocida
26	nmdesm0	Numérico	8	2	Número de descendientes sin minusvalía
27	nmdesmh6	Numérico	8	2	Número de descendientes con minusvalía >= 33 y < 65 % sin movilidad reducida
28	nmdesmh7	Numérico	8	2	Número de descendientes con minusvalía >= 33 y < 65 % con movilidad reducida
29	nmdesmr	Numérico	8	2	Número de descendientes con minusvalía >= 65 %
30	nmdiscd	Numérico	8	2	Número de descendientes con minusvalía
31	nmasc	Numérico	8	2	Número de ascendientes
32	nmdisca	Numérico	8	2	Número de ascendientes con minusvalía
33	nmm65a	Numérico	8	2	Número de ascendientes > 65 años
34	nmm75a	Numérico	8	2	Número de ascendientes > 75 años
35	nmascm0	Numérico	8	2	Número de ascendientes sin minusvalía
36	nmascmh6	Numérico	8	2	Número de ascendientes con minusvalía >= 33 y < 65 % sin movilidad reducida
37	nmascmh7	Numérico	8	2	Número de ascendientes con minusvalía >= 33 y < 65 % con movilidad reducida
38	nmascmr	Numérico	8	2	Número de descendientes con minusvalía >= 65 %
39	par1	Numérico	8	2	Rdto. del trabajo Dinerarios

La BBDD consta de 223 variables (columnas) y 77109 registros (filas).

La Base de datos es una muestra asociada a la declaración de la renta para distintos individuos. Cada registro es una persona a la que se le asocia un número identificador único. Como son muchos registros esta muestra es representativa de la población. Por lo tanto, una vez validado nuestro modelo, nos servirá para clasificar individuos ajenos al dataset.

Explicación de cada variable

- ID = IDENTIFICADOR DE LAS PERSONAS (número único)
- F_TMG = FRAUDE DE TIPO MARGINAL (creación de sociedades ficticias)
- F_CAPINM = FRAUDE DE CAPITAL INMOBILIARIO
- F_NHIJOS = FRAUDE POR NÚMERO DE HIJOS
- F_AAEE = FRAUDE POR ACTIVIDADES ECONÓMICAS
- F_PLANP = FRAUDE POR PLANES DE PENSIONES
- F_GASTOS = FRAUDE POR GASTOS
- MARCA = Si el individuo ha defraudado (marca=1) o no (marca=0)
- FACTOR = FACTOR DE ELEVACION DE LA MUESTRA
- CDPOST = CÓDIGO POSTAL
- PAR1 = LO QUE GENERAMOS
- PAR2 = RETRIBUCIONES EN ESPECIE
- ...

La variable dependiente marca nos indica si hubo fraude o no fraude

- 1 si hubo fraude
- 0 si no hubo fraude

Por lo tanto, creamos un modelo logit binomial con variable dependiente MARCA que nos indique si ha defraudado o no a partir de las variables económicas que son independientes (variables par1, par2, par279).

$$marca = f(par_1, \dots, par_n)$$

•EDA

Como nuestra BBDD está formada por más de 76000 registros, no podemos usar el procedimiento explorar, salvo que sea muy necesario, ya que el software no está optimizado para datos reales (con muchos registros).

Las variables de ingresos, rentas o económicas nunca se distribuyen como una campana de Gauss, es decir, normalmente. Se distribuyen según la distribución de Pareto (Paretianas).

Sabemos que no hay missing ya que la agencia tributaria ya los ha tratado anteriormente, sin embargo, puede encontrarse atípicos. Estos atípicos no pueden quitarse ya que son observaciones influyentes.

El dataset está formada por 223 variables así que seguramente también haya varias variables independientes correladas entre sí. Es decir, tendríamos una gran correlación.

¿cómo podemos lidiar con el problema de los atípicos y la multicolinealidad?

Haciendo previamente la reducir de la dimensión.

Se puede reducir la dimensión de las variables par_1, \dots, par_n mediante distintas técnicas por el análisis de las componentes principales y obtendremos un menor número de variables fac_1, \dots, fac_n (n componentes principales) que explican aproximadamente la misma variabilidad del problema.

Visualmente, lo que intentan las componentes principales sería:

$$par_1, \dots, par_n \rightarrow \rightarrow \rightarrow Fac_1, \dots, Fac_m$$

Donde $n > m$. Obviamente, Cada factor es combinación lineal de las variables independientes.

Una vez hecho la reducción, empezaríamos a trabajar con las nuevas variables. Por lo tanto, el modelo a estimar quedaría de la forma:

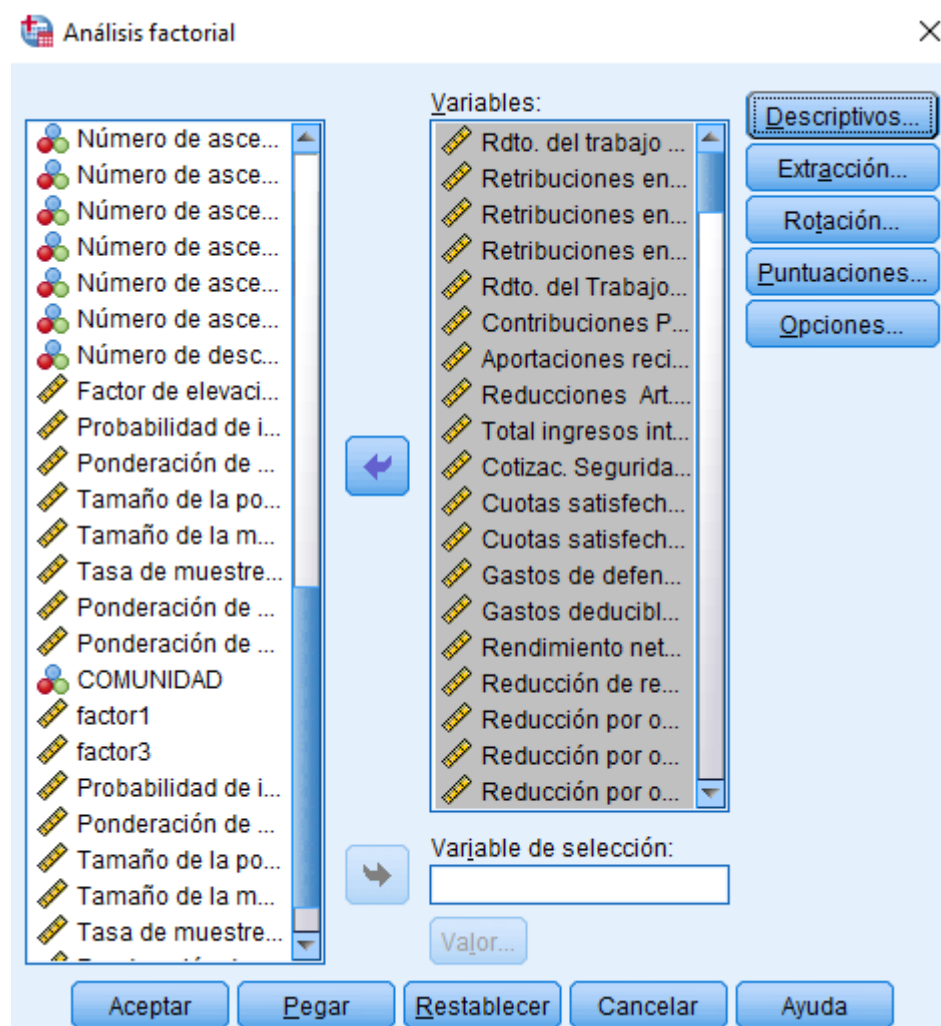
$$marca = f(Fac_1, \dots, Fac_n)$$

Recopilando toda la información anterior tenemos que:

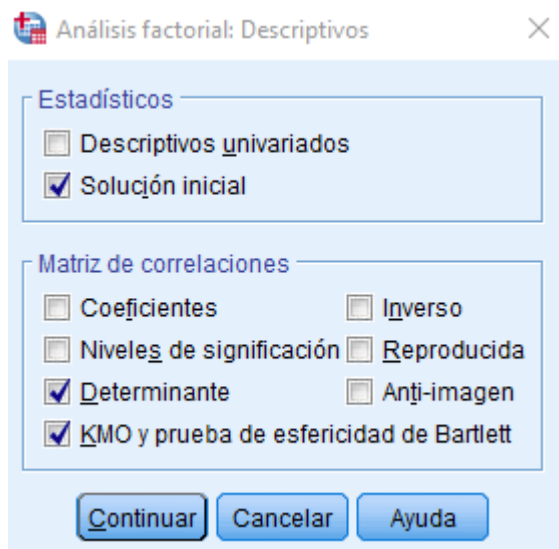
Al reducir la dimensión:

- Ausencia de variables incorreladas.
- Las variables FAC1 FAC63 ya no contienen atípicos.
- Las componentes principales (FAC1,..., FAC63) tienden a la normalidad (En los modelos no se exige que las variables sean normales, en los residuos sí).
- Escala uniforme en las componentes.

- Reducción de la dimensión



Tomamos las variables para reducir desde par_1 hasta par_779



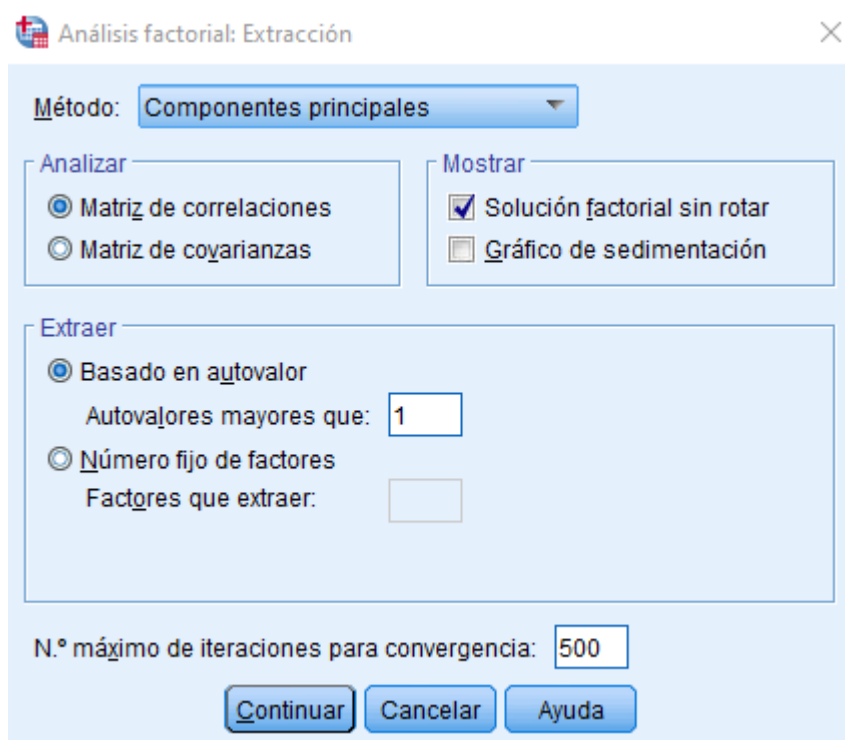
Análisis factorial: Descriptivos

Estadísticos

- ☐ Descriptivos univariados
- ☒ Solución inicial

Matriz de correlaciones

- ☐ Coeficientes
- ☐ Inverso
- ☐ Niveles de significación
- ☐ Reproducida
- ☒ Determinante
- ☐ Anti-imagen
- ☒ KMO y prueba de esfericidad de Bartlett



Análisis factorial: Extracción

Método: Componentes principales

Analizar

- ☒ Matriz de correlaciones
- ☐ Matriz de covarianzas

Mostrar


- ☒ Solución factorial sin rotar
- ☐ Gráfico de sedimentación

Extraer

- ☒ Basado en autovalor
Autovalores mayores que: 1
- ☐ Número fijo de factores
Factores que extraer:

N.º máximo de iteraciones para convergencia: 500

Elegimos el método de reducción por componentes principales. El número máximo de iteraciones para convergencia lo ponemos en 500 ya que estamos trabajando con un problema real.

 **Análisis factorial: Rotación** ✕

Método

☐ Ninguno ☐ Quartimax
☒ Varimax ☐ Equamax
☐ Oblimin directo ☐ Promax


Delta: Kappa:

Mostrar

☒ Solución rotada ☐ Gráficos de cargas

N.º máximo de iteraciones para convergencia:

En problemas de componentes principales siempre hay que rotar para facilitar la interpretación de cada factor.

 **Análisis factorial: Puntuaciones factoriales** ✕

☒ Guardar como variables

Método

☒ Regresión
☐ Bartlett
☐ Anderson-Rubin

☐ Mostrar matriz de coeficientes de las puntuaciones factoriales

Las puntuaciones son las coordenadas de las componentes principales.

- DIAGNOSIS

Matriz de correlaciones^{a,b}

a. Determinante
= ,000

b. Esta matriz no
es cierta
positiva.

Como no podemos ver la matriz de correlaciones dado a que nuestra base de datos es enorme, un método alternativo para ver si podemos reducir el modelo es mirando su determinante. Así, si el determinante de la matriz de correlaciones es próximo a cero, podemos seguir con la reducción. Es recomendable que el determinante no supere la milésima.

Comunalidades

	Inicial	Extracción
Rdto. del trabajo Dinerarios	1,000	,988
Retribuciones en especie (valoración)	1,000	,995
Retribuciones en especie (ingresos a cuenta)	1,000	,997
Retribuciones en especie (ingresos a cuenta repercutidos)	1,000	,982
Rdto. del Trabajo En especie.	1,000	,990
Contribuciones Planes Pensiones.	1,000	,599
Aportaciones recibidas al patrimonio protegido de las personas con discapacidad del que es titular el contribuyente	1,000	,846
Reducciones Art. 18 apartados 2 y 3, y dispos. trans. 11ª y 12ª Ley del Impuesto	1,000	,752
Total ingresos integros computables [(01)+(05)+ (06)+(07)-(08)]	1,000	,994
Cotizac. Seguridad Social, Mutuality Funcionarios, detracciones derechos pasivos y Coleg. Huérfanos.	1,000	,938

Las comunalidades extraídas (no teóricas) deben ser lo más cercanas a la unidad posibles. Esto quiere decir que estamos reduciendo satisfactoriamente.

61	1,001	,535	79,910	1,001	,535	79,910	1,004	,537	79,908
62	1,001	,535	80,445	1,001	,535	80,445	1,003	,537	80,444
63	1,000	,535	80,980	1,000	,535	80,980	1,001	,535	80,980

Hay 4 criterios para retener el número de factores necesarios del modelo. El criterio que vamos a usar será el de la variabilidad mínima exigida y el de la variabilidad total explicada.

Por el método de la variabilidad mínima exigida, al estar trabajando con la matriz de correlaciones, este criterio defiende que nos quedemos con los factores que superan o igualen la unidad en los autovalores. Por lo tanto, nos recomienda quedarnos con 63.

Es decir, se producirá una reducción de dimensiones de 728 variables a 63.

Además, la columna de la variabilidad total explicada por el modelo se encuentra en 80.980%. Es un valor muy bueno ya que solo se pierde el 19% de la variabilidad de los datos.

La matriz factorial nos asocia cada variable de la base de datos con cada componente principal.

	Componente														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cuota íntegra estatal	,986	,067	-,103	-,003	,006	-,075	-,045	-,041	,012	,004	-,009	,020	,017	-,011	-,008
Cuota líquida estatal incrementada	,985	,070	-,111	-,003	,000	-,073	-,055	-,034	-,007	,012	-,015	,013	,023	-,009	-,010
Cuota líquida estatal	,985	,070	-,111	-,003	-,002	-,073	-,054	-,034	-,007	,012	-,015	,013	,024	-,009	-,010
Parte de las cuotas íntegras del ejercicio 2009 que corresponde a la Comunidad Autónoma [50% de (698) + (699)]	,984	,076	-,108	-,005	,006	-,076	-,045	-,042	,010	,004	-,010	,023	,014	-,011	-,007
Cuota líquida incrementada total	,983	,079	-,114	-,004	,004	-,075	-,058	-,034	-,008	,012	-,013	,014	,023	-,009	-,010
Cuota íntegra autonómica o complementaria	,983	,091	-,111	-,006	,006	-,077	-,045	-,044	,010	,006	-,010	,023	,015	-,011	-,008
Cuota resultante de la autoliquidación	,982	,076	-,119	-,002	,004	-,081	-,058	-,036	-,013	,013	-,017	,012	,006	-,002	,005

Por ejemplo, la primera componente factorial predominan las variables cuotas.

Una vez reducido nuestro modelo ya podemos continuar con la estimación del logit.

- ESTIMACIÓN

Regresión logística

Dependientes:

Bloque 1 de 1

Anterior

Covariables:

Método:

Variable de selección:

Regresión logística: Guardar

Valores pronosticados

☒ Probabilidades
☒ Grupo de pertenencia

Influencia

☒ De Cook
☐ Valores de influencia
☐ DfBetas

Residuos

☒ No estandarizados
☐ Logit
☐ Método de Student
☐ Estandarizados
☐ Desviación

Exportar información del modelo a un archivo XML

☐ Incluir la matriz de covarianzas

Regresión logística: Opciones

Estadísticos y gráficos

☐ Gráficos de clasificación

☒ Bondad de ajuste de Hosmer-Lemeshow

☐ Listado de residuos por caso

☒ Valores atípicos fuera de 2 Desviación estándar

☐ Todos los casos

☒ Correlaciones de estimaciones

☐ Historial de iteraciones

☒ CI para exp(B): 95 %

Visualización

☐ En cada paso ☒ En el último paso

Probabilidad para el método por pasos

Entrada: 0,05 Eliminación: 0,10

Punto de corte para la clasificación: 0,5

Iteraciones máximas: 500

☐ Conservar memoria para análisis complejos o conjuntos de datos grandes

☒ Incluir constante en modelo

Continuar Cancelar Ayuda

Avisos

La estimación ha fallado debido a un problema numérico. Las razones posibles son: (1) el valor de EPS es demasiado pequeño (si no se especifica, el valor predeterminado utilizado puede ser demasiado pequeño para este conjunto de datos) o (2) el valor de EPS es demasiado pequeño (si no se especifica, el valor predeterminado utilizado puede ser demasiado pequeño para este conjunto de datos).


Bloque 1: Método = Entrar

Variables en la ecuación

Contrastado un error en la estimación

El logit ha fallado ya que el método de máxima verosimilitud no tiene suficiente potencia para converger. Intentaremos resolver el logit mediante el modelo lineal generalizado.

• ESTIMACIÓN POR MODELO LINEAL GENERALIZADO (MLG)

 Personalizado

☒ Personalizado

Distribución: Binomial Función de enlace: Logit


Parámetro

☒ Especificar valor

Valor:

☐ Estimar valor

Potencia:

 Modelos lineales generalizados ✕


Tipo de modelo Respuesta Predictores Modelo Estimación Estadísticos Medias marginales estimadas Guardar Exportar

Variables:

Variable dependiente

Variable dependiente: marca

Orden de las categorías (sólo distribución multinomial): Ascendente

 Modelos lineales generalizados ✕

Tipo de modelo Respuesta Predictores Modelo Estimación Estadísticos Medias marginales estimadas Guardar Exportar

Variables:

Factores:

Opciones...

Covariables:

REGR factor score 1 for analysis 1 [FAC1_1]
 REGR factor score 2 for analysis 1 [FAC2_1]
 REGR factor score 3 for analysis 1 [FAC3_1]
 REGR factor score 4 for analysis 1 [FAC4_1]
 REGR factor score 5 for analysis 1 [FAC5_1]
 REGR factor score 6 for analysis 1 [FAC6_1]
 REGR factor score 7 for analysis 1 [FAC7_1]
 REGR factor score 8 for analysis 1 [FAC8_1]

Modelos lineales generalizados

Tipo de modelo Respuesta Predictores **Modelo** Estimación Estadísticos Medias marginales estimadas Guardar Exportar

Especificar efectos del modelo

Factores y covariables:

- ☒ FAC1_1
- ☒ FAC2_1
- ☒ FAC3_1
- ☒ FAC4_1
- ☒ FAC5_1
- ☒ FAC6_1
- ☒ FAC7_1
- ☒ FAC8_1
- ☒ FAC9_1
- ☒ FAC10_1
- ☒ FAC11_1
- ☒ FAC12_1
- ☒ FAC13_1
- ☒ FAC14_1

Construir términos

Tipo: Efectos principales

Modelo:

- FAC1_1
- FAC2_1
- FAC3_1
- FAC4_1
- FAC5_1
- FAC6_1
- FAC7_1
- FAC8_1
- FAC9_1
- FAC10_1
- FAC11_1
- FAC12_1
- FAC13_1
- FAC14_1
- FAC15_1
- FAC16_1

Número de efectos en el modelo: 63

Construir término anidado

Término:

Por * (Dentro) Añadir al modelo Borrar

☒ Incluir la interacción en el modelo

Modelos lineales generalizados

Tipo de modelo Respuesta Predictores Modelo **Estimación** Estadísticos Medias marginales estimadas Guardar Exportar

Estimación de parámetros

Método: Híbrido

Número máximo de iteraciones de puntuación de Fisher: 1

Método de parámetro de escala: Valor fijo

Valor: 1

Matriz de covarianzas

☒ Estimador basado en el modelo

☐ Estimador robusto

☐ Obtener valores iniciales para las estimaciones de parámetros a partir de un conjunto de datos

Valores iniciales...

Iteraciones

Iteraciones máximas: 500

Máxima subdivisión por pasos: 5

☐ Comprobar separación de los puntos de los datos

Iteración de inicio: 20

Criterios de convergencia

Debe especificarse al menos un criterio de convergencia con un mínimo mayor que 0.

☒ Cambio en las estimaciones de los parámetros

Mínimo: 1E-006

Tipo: Absoluta

☐ Cambio en la log-verosimilitud

Absoluta

Imprimir

☒ Resumen de procesamiento de casos
☐ Estadísticos descriptivos
☒ Información de modelo
☒ Estadísticos de bondad de ajuste
☒ Estadísticos de resumen del modelo
☒ Estimaciones de los parámetros

☒ Incluir estimaciones de los parámetros exponenciales
☐ Matriz de covarianzas de las estimaciones de los parámetros
☐ Matriz de correlaciones de las estimaciones de los parámetros

☐ Matrices (L) de los coeficientes de contraste
☐ Funciones estimables generales
☐ Historial de iteraciones
 Intervalo de impresión:
☐ Contraste de multiplicadores de Lagrange de parámetro de escala o parámetro auxiliar para binomial negativa

Bondad de ajuste^a

	Valor	gl	Valor/gl
Desviación	35567,878	76597	,464
Desviación escalada	35567,878	76597	
Chi-cuadrado de Pearson	1,243E+133	76597	1,622E+128
Chi-cuadrado de Pearson escalado	1,243E+133	76597	
Logaritmo de verosimilitud ^b	-17783,939		
Criterio de información Akaike (AIC)	35695,878		
AIC corregido para muestras finitas (AICC)	35695,986		
Criterio de información bayesiana (BIC)	36288,069		
AIC coherente (CAIC)	36352,069		

Variable dependiente: marca

En la tabla de bondad de ajuste podemos ver los indicadores de capacidad predictiva.

Estos indicadores miden la cantidad de información. Por sí solos no tienen sentido. Solo lo tienen si se comparan con otros modelos, a menudo con el probit. El que menor valor tenga de estos indicadores es el modelo más adecuado.

Estimaciones de parámetro										
Parámetro	B	Desv. Error	95% de intervalo de confianza de Wald		Contraste de hipótesis			Exp(B)	95% de intervalo de confianza de Wald para Exp(B)	
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.		Inferior	Superior
(Intersección)	-19,043	,2728	-19,577	-18,508	4873,070	1	,000	5,368E-9	3,145E-9	9,163E-9
REGR factor score 1 for analysis 1	-25,815	,6514	-27,092	-24,539	1570,672	1	,000	6,145E-12	1,714E-12	2,203E-11
REGR factor score 2 for analysis 1	-2,475	,0966	-2,665	-2,286	656,074	1	,000	,084	,070	,102
REGR factor score 3 for analysis 1	-2,165	,0628	-2,288	-2,042	1186,860	1	,000	,115	,101	,130
REGR factor score 4 for analysis 1	-19,658	,4361	-20,513	-18,804	2032,331	1	,000	2,900E-9	1,234E-9	6,818E-9
REGR factor score 5 for analysis 1	-1,596	,0494	-1,692	-1,499	1043,845	1	,000	,203	,184	,223
REGR factor score 6 for analysis 1	-3,627	,1180	-3,858	-3,395	944,074	1	,000	,027	,021	,034
REGR factor score 7 for analysis 1	-25,007	1,9888	-28,905	-21,109	158,091	1	,000	1,380E-11	2,798E-13	6,802E-10
REGR factor score 8 for analysis 1	-7,794	,3242	-8,430	-7,159	577,976	1	,000	,000	,000	,001
REGR factor score 9 for analysis 1	1,122	,1627	,803	1,440	47,537	1	,000	3,070	2,232	4,223
REGR factor score 10 for analysis 1	-5,190	,0687	-5,324	-5,055	5702,363	1	,000	,006	,005	,006

Estimaciones de parámetros para las 10 componentes principales. La mayoría son muy significativos. Para este tipo de modelos, su p-valor debe ser inferior a 0.5 para que sea significativo ya que el modelo logístico es muy no lineal.

Una vez validado nuestro modelo vemos las variables creadas por SPSS

MeanPr...	Predict...	Residual	CooksD...
,899	,00	,101	,000
,269	1,00	-,269	,000
,000	1,00	,000	,000
,709	,00	,291	,000
,178	1,00	-,178	,000
,877	,00	-,877	,000
,000	1,00	,000	,000
,000	1,00	,000	,000
,660	,00	,340	,000
,391	1,00	-,391	,000
,043	1,00	-,043	,000

En primer lugar, se encuentran la variable “*MeanPredicted*”. Esta variable nos muestra que probabilidad hay de que el individuo no sea un defraudador

La siguiente variable es “*PredictedValue*” y su función es clasificar a cada individuo en un grupo, siendo 0 el grupo de los no defraudadores y 1 el de los defraudadores.

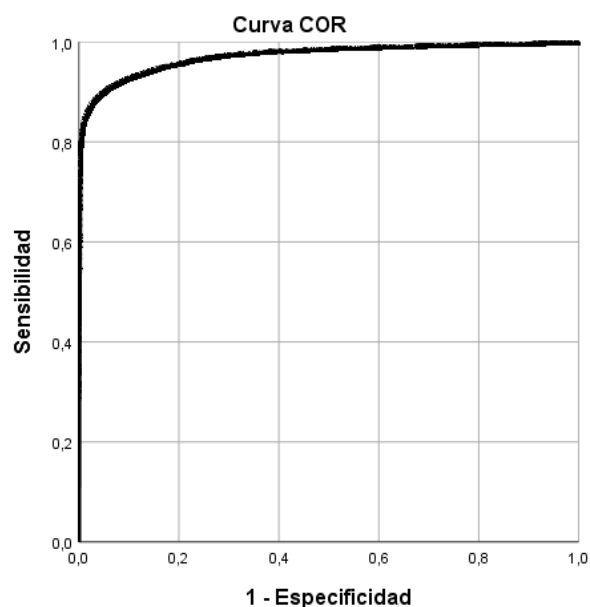
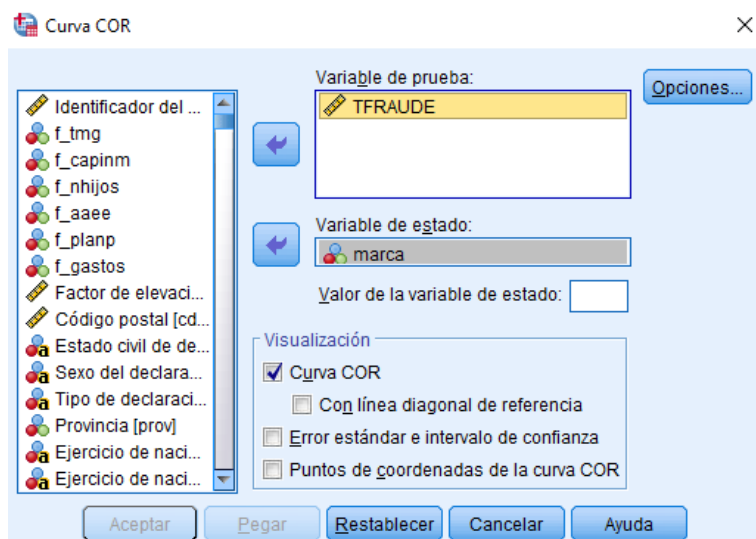
Las dos siguientes columnas (“*Residual*”, “*CooksDistance*”) son los residuos y las distancias de Cook respectivamente para cada observación.

Vamos a realizar un análisis más profundo de las personas defraudadoras graficando la curva ROC. Para ello creamos una nueva variable que valga el complementario de MeanPredict:

$$TFraude = 1 - MeanPredict$$

Tfraude indica la probabilidad de fraude del individuo

- Curva ROC



Los segmentos de diagonal se generan mediante empates.

Área bajo la curva

de resultado de prueba: TFRAUDE

Área

,971

les de resultado de prueba: TFRAUDE tienen, como mínimo
tre el grupo de estado real positivo y el grupo de estado real
sticas podrían estar sesgadas.

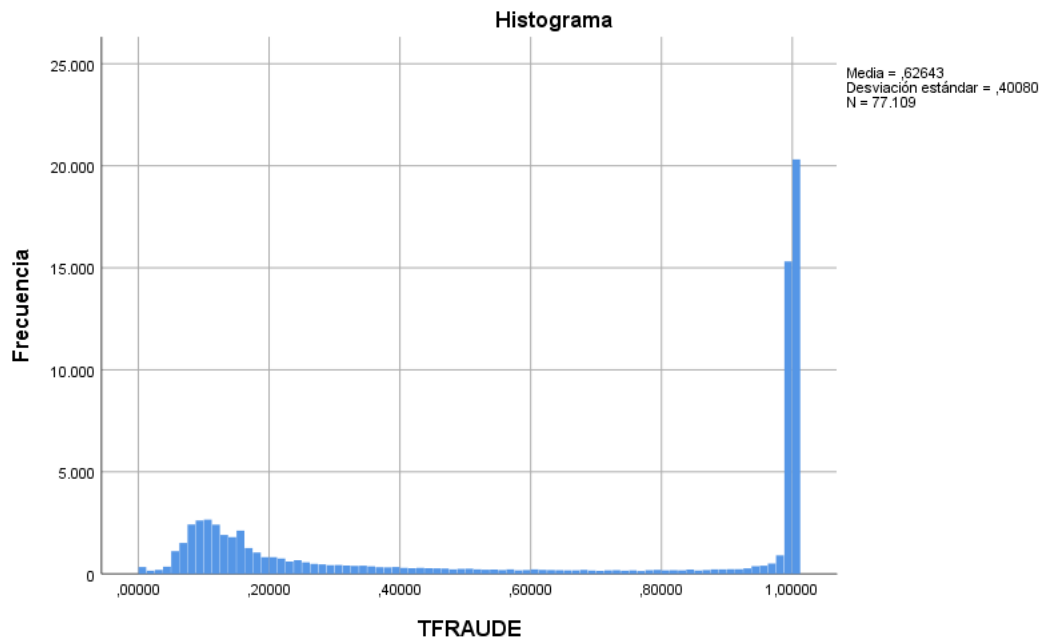
El área bajo la curva es de 0.971, por lo tanto, es un valor muy bueno para nuestro modelo.

El área bajo la curva nos indica que clasificará a los individuos correctamente en un 97,1%

Vamos a ver a partir de que punto de corte tendríamos que investigar a los individuos para evitar problemas fraudulentos. Decidir cual es el punto de corte optimo se ve en el histograma de la variable Tfraude.

- EDA de Tfraude





A partir del punto de corte 0.8 hay personas potencialmente fraudulentas.

Si hay personas que tenían un 0 en la variable marca y en nuestra columna de valores predichos un 1 también serían personas potencialmente fraudulentas. Esta discordancia de valores nos informa que los investigadores han cometido fallos en la estimación del modelo.

Práctica en SAS

Introducción

Ya hemos visto como se hace la regresión logística en SPSS. Ahora vamos a comparar los resultados de ambos softwares y elegiremos con cual salida nos quedamos.

Veremos también, las diferencias notorias en cuanto a la precisión de la estimación de los autovalores.

Resumen

- Reducción de dimensión

Nfactors es el número de factores que se enviaran a la salida.

```
proc factor data=datos rotate=varimax out=salida;  
var par1--par779;  
run;
```

Con la submuestra 1ª reducimos por componentes principales a través de PROC FACTOR. Rotamos para que los factores sean fácilmente interpretables.

La opción VAR sirve para seleccionar las variables que se van a reducir son de par1 a par 779. Se pone dos guiones por precaución ya que puede ser que una variable entre ese intervalo que no exista.

Autovalores de la matriz de correlación: Total = 187 Promedio = 1				
	Autovalor	Diferencia	Proporción	Acumulada
1	25.8935497	15.8915244	0.1385	0.1385
2	10.0020254	3.6031508	0.0535	0.1920
3	6.3988746	1.1687600	0.0342	0.2262
4	5.2301146	0.6990374	0.0280	0.2541
5	4.5310771	0.0647157	0.0242	0.2784

53	1.0223443	0.0065970	0.0055	0.7561
54	1.0157473	0.0032157	0.0054	0.7615
55	1.0125316	0.0058809	0.0054	0.7669
56	1.0066508	0.0014934	0.0054	0.7723
57	1.0051574	0.0006590	0.0054	0.7777
58	1.0044985	0.0014327	0.0054	0.7830
59	1.0030657	0.0016901	0.0054	0.7884
60	1.0013756	0.0006817	0.0054	0.7937
61	1.0006939	0.0001869	0.0054	0.7991
62	1.0005070	0.0004918	0.0054	0.8044
63	1.0000152	0.0000872	0.0053	0.8098
64	0.9999280	0.0016879	0.0053	0.8151

La salida nos muestra los autovalores de la matriz de correlaciones. A diferencia de SPSS, SAS es más preciso con las estimaciones. Es decir, ofrece más decimales.

Por el método de la variabilidad mínima exigida, al estar trabajando con la matriz de correlaciones, este criterio defiende que nos quedemos con los factores que superan o igualen la unidad en los autovalores. Por lo tanto, nos recomienda quedarnos con 63.

Es decir, se producirá una reducción de dimensiones de 728 variables a 63.

Además, la columna de la variabilidad total explicada por el modelo se encuentra en 80.980%. Es un valor muy bueno ya que solo se pierde el 19% de la variabilidad de los datos.

En nuestro modelo, retendremos 63 factores. Por lo tanto, volvemos a hacer el PROC FACTOR, pero esta vez con la opción NFACTORS=63

```
proc factor data=datos rotate=varimax nfactors=63 out=salida;
var par1--par779;
run;
```

Mirando el fichero salida vemos que ha enviado allí las 63 componentes principales:

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9
-0.232233149	-0.015234422	-0.046452412	-0.151189571	-0.462012238	-0.002163417	0.0285622402	-0.026932527	-0.0268352
-0.235501621	-0.032398861	-0.000517361	-0.054920271	-0.281066188	0.0036839439	0.0184606802	0.0247852482	0.0872783726
-0.224070273	-0.053100992	-0.050354251	-0.167791558	-0.236149959	-0.011435509	0.0155525471	-0.038378718	0.0239877001
-0.226676981	-0.010880595	-0.020260573	-0.086276554	-0.453958396	0.003326286	0.0206458583	-0.013365135	0.0007089323
-0.229795578	-0.01203948	0.0115813657	-0.030213889	-0.419589977	0.0185222592	0.0295596795	0.0502796398	0.1517719486
-0.233796781	-0.024620901	-0.029155161	-0.132925606	-0.416649104	0.0000493424	0.0303774871	-0.01124618	0.0014457553
-0.198474377	-0.108531688	-0.059745403	-0.641507948	-1.3622272	-0.028817854	0.0312734372	-0.282297079	0.0145289895

Una vez reducido el número de variables podemos realizar el modelo logístico.

```
proc logistic data=salida plots=ALL;  
  model marca=factor1-factor63;  
  output out=salida1 predprobs=individual reschi=residuos resdev=influencia;  
run;
```

variable independiente marca en función de los factores principales (factor1, ..., factor53)

la opción PREDPROBS sirve para guardar probabilidades predichas de fraude de los individuos

la opción RESCHI sirve para guardar los residuos y la opción RESDEV para ver los valores de influencia.

Estado de convergencia del modelo

Límite de iteración alcanzado sin convergencia.

Warning: Convergence was not attained in 25 iterations. You may want to increase the maximum number of iterations (MAXITER= option) or change the convergence criteria (ABSFCNV=, FCONV=, GCONV=, XCONV= options) in the MODEL statement.

Warning: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

El modelo ha alcanzado el límite de iteración y no ha convergido. Esto es debido a que por defecto SAS solo hace 25 iteraciones. Como estamos en un problema real, vamos a poner 500 iteraciones con la opción MAXITER.

```
proc logistic data=salida plots=ALL;  
  model marca=factor1-factor63/ maxiter=500 ;  
  output out=salida1 predprobs=individual reschi=residuos resdev=influencia;  
run;
```

El problema de convergencia se ha resuelto. A diferencia de SPSS que nos mandaba a los modelos lineales generalizados, SAS lo consigue hacer cambiando el número de iteraciones máximas.

- DIAGNOSIS

Información del modelo		
Conjunto de datos	WORK.SALIDA	
Variable de respuesta	marca	marca
Número de niveles de respuesta	2	
Modelo	logit binario	
Técnica de optimización	Puntuación de Fisher	

Al contrario de SPSS que usaba por defecto el método por máxima verosimilitud, SAS ha usado el test de puntuaciones de Fisher. Esto es debido a que SAS automáticamente si detecta algún problema usa por defecto uno con mayor potencia.

Perfil de respuesta		
Valor ordenado	marca	Frecuencia total
1	FRAUDE GRAL	48362
2	LEGAL GRAL	28747

El numero de personas que han cometido fraude a priori es de 48362 personas. El otro grupo es el que no ha cometido acciones fraudulentas.

Estadístico de ajuste del modelo		
Criterio	Sólo término independiente	Término independiente y covariables
AIC	101852.85	41733.261
SC	101862.10	42232.922
-2 LOG L	101850.85	41625.261

Los indicadores de capacidad predictiva tienen valores muy similares entre sí.

Estos indicadores miden la cantidad de información. Por sí solos no tienen sentido. Solo lo tienen si se comparan con otros modelos, a menudo con el probit. El que menor valor tenga de estos indicadores es el modelo más adecuado.

Probar hipótesis nula global: BETA=0			
Test	Chi-cuadrado	DF	Pr > ChiSq
Ratio de verosim	60225.5876	53	<.0001
Puntuación	28443.5534	53	<.0001
Wald	10060.9776	53	<.0001

Test de significatividad conjunta. Por todos los test se comprueba que podemos rechazar la hipótesis de que los parámetros sean iguales a cero. Es decir, conjuntamente tienen significatividad los parámetros del modelo.

Análisis del estimador de máxima verosimilitud					
Parámetro	DF	Estimador	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	12.2441	0.2002	3742.0015	<.0001
Factor1	1	21.1426	0.6474	1066.6340	<.0001
Factor2	1	0.6461	0.0710	82.7068	<.0001
Factor3	1	1.6433	0.0632	675.2087	<.0001
Factor4	1	13.3557	0.3179	1764.9869	<.0001
Factor5	1	1.0233	0.0418	598.3915	<.0001
Factor6	1	1.0790	0.0875	151.9745	<.0001
Factor7	1	23.2939	1.9313	145.4788	<.0001
Factor8	1	4.9397	0.3212	236.5550	<.0001
Factor9	1	1.4570	0.1035	198.0601	<.0001
Factor10	1	3.3869	0.0431	6162.0174	<.0001
Factor11	1	28.4974	1.0511	735.1101	<.0001
Factor12	1	5.5455	0.2002	767.3329	<.0001
Factor13	1	1.7844	0.0878	413.3600	<.0001
Factor14	1	6.2816	0.1986	1000.7434	<.0001
Factor15	1	0.7528	0.0167	2020.3358	<.0001
Factor16	1	0.1365	0.0262	27.1350	<.0001
Factor17	1	0.6009	0.0640	88.2791	<.0001
Factor18	1	-0.0278	0.0372	0.5596	0.4544
Factor19	1	18.5527	2.1729	72.9007	<.0001
Factor20	1	3.2570	0.6005	29.4164	<.0001
Factor21	1	0.0415	0.0672	0.3817	0.5367
Factor22	1	-0.2380	0.0116	418.4258	<.0001
Factor23	1	0.0558	0.0340	2.6906	0.1009
Factor24	1	0.0336	0.0476	0.4986	0.4801
Factor25	1	11.6359	0.8119	205.3876	<.0001
Factor26	1	-0.00341	0.0203	0.0280	0.8670
Factor27	1	15.5454	0.2347	4386.3722	<.0001
Factor28	1	-0.3014	0.0142	449.0437	<.0001
Factor29	1	2.3538	0.0467	2535.3575	<.0001
Factor30	1	0.8246	0.0719	131.5438	<.0001

Factor31	1	1.8406	0.1852	98.7854	<.0001
Factor32	1	-0.1260	0.0153	67.5974	<.0001
Factor33	1	0.2262	0.1715	1.7393	0.1872
Factor34	1	0.0377	0.1110	0.1154	0.7341
Factor35	1	-0.0394	0.0203	3.7874	0.0516
Factor36	1	-0.3845	0.0833	21.3072	<.0001
Factor37	1	-0.0841	0.0103	66.4907	<.0001
Factor38	1	0.7452	0.0289	666.6655	<.0001
Factor39	1	0.6457	0.0243	703.4953	<.0001
Factor40	1	4.2665	0.2654	258.3288	<.0001
Factor41	1	1.9808	0.0701	797.3010	<.0001
Factor42	1	0.7275	0.0180	1626.2620	<.0001
Factor43	1	0.0787	0.0347	5.1541	0.0232
Factor44	1	0.0762	0.0237	10.3077	0.0013
Factor45	1	-0.2688	0.0629	18.2684	<.0001
Factor46	1	0.6859	0.0515	177.4685	<.0001
Factor47	1	-0.1850	0.0280	43.7431	<.0001
Factor48	1	0.7567	0.0177	1835.1052	<.0001
Factor49	1	-0.00060	0.0389	0.0002	0.9877
Factor50	1	0.3495	0.0733	22.7539	<.0001
Factor51	1	0.2863	0.0397	52.0562	<.0001
Factor52	1	-1.0341	0.0414	622.8568	<.0001
Factor53	1	1.3651	0.1213	126.7021	<.0001
Factor54	1	-0.4588	0.0166	761.4131	<.0001
Factor55	1	1.3975	0.0692	407.5810	<.0001
Factor56	1	-0.1634	0.0188	75.1994	<.0001
Factor57	1	-0.0278	0.0300	0.8579	0.3543
Factor58	1	2.7696	0.1058	685.0386	<.0001
Factor59	1	0.0760	0.0533	2.0363	0.1536
Factor60	1	0.1850	0.1056	3.0696	0.0798
Factor61	1	-0.1112	0.0198	31.5712	<.0001
Factor62	1	-0.0961	0.0119	64.9957	<.0001
Factor63	1	-0.0204	0.0117	3.0485	0.0808

Prueba de Wald significatividad individual para modelos no lineales.

Actúa de la misma forma que la t de student para modelos lineales. La mayoría de los factores tienen P-valor bajo, por lo tanto, son muy significativos. Algunos p-valores son altos, pero para modelos no lineales las situaciones de aceptación/rechazo de significatividad individual son más laxas. Cuando tenemos muchas variables no es necesario quitarlo. En todo caso podríamos quitarlo con la opción LSENTRY y LSTAY.

Estimadores de cocientes de disparidad;			
Efecto	Estimador del punto	Límites de confianza al 95% de Wald	
Factor1	>999.999	>999.999	>999.999
Factor2	1.908	1.660	2.193
Factor3	5.172	4.569	5.855
Factor4	>999.999	>999.999	>999.999
Factor5	2.782	2.563	3.020
Factor6	2.942	2.478	3.492
Factor7	>999.999	>999.999	>999.999
Factor8	139.722	74.454	262.205
Factor9	4.293	3.505	5.259
Factor10	29.575	27.177	32.185
Factor11	>999.999	>999.999	>999.999
Factor12	256.091	172.977	379.140
Factor13	5.956	5.015	7.074
Factor14	534.653	362.285	789.031
Factor15	2.123	2.054	2.194
Factor16	1.146	1.089	1.207
Factor17	1.824	1.609	2.067
Factor18	0.973	0.904	1.046
Factor19	>999.999	>999.999	>999.999
Factor20	25.971	8.004	84.263

El odds ratio nos indica que variables son las que más inciden en la variable fraude. Es decir, mientras más grande sea el odds ratio, más incidencia tiene en la probabilidad de fraude, y mientras más bajo sea este valor (menor a 1) influye negativamente a la hora de detectar fraude.

Hemos encontrado varios factores que incrementan la detección de fraude entre ellos están el factor 1, 4, 7, 11, 14, 19 y muchos más.

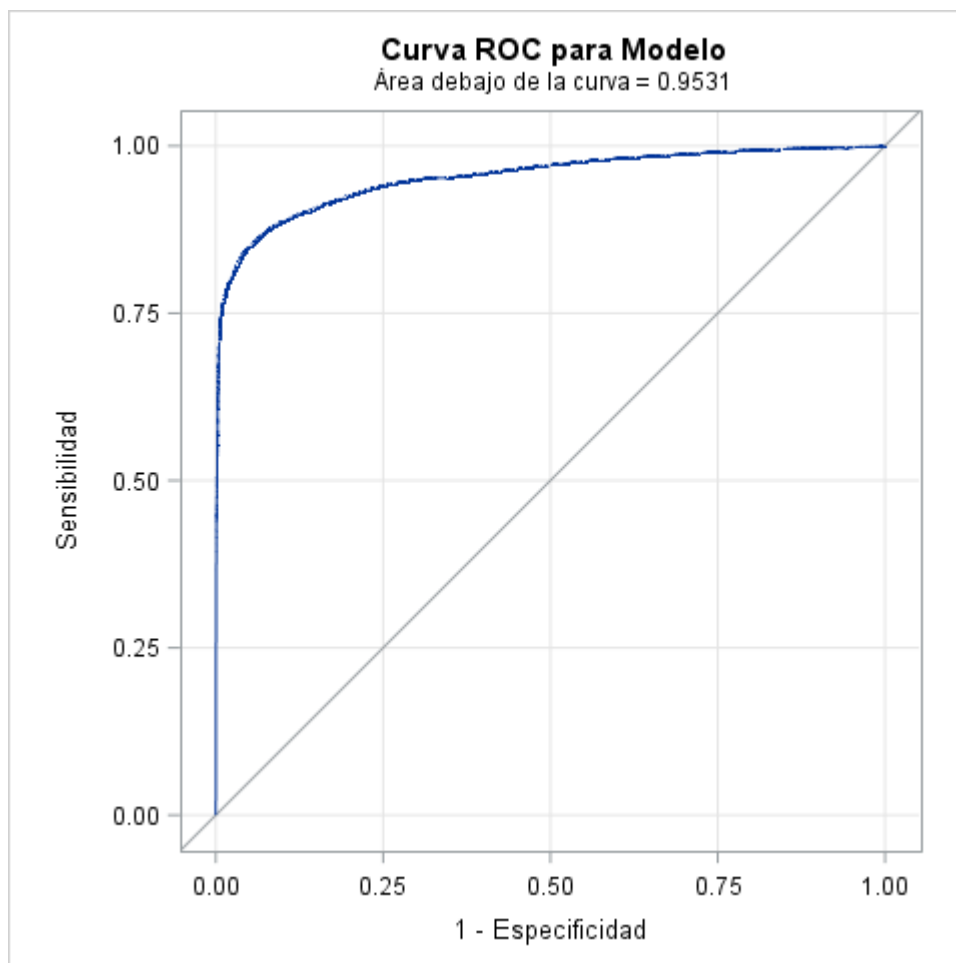
Una alternativa a la matriz de confusión son la asociación de probabilidades predichas y respuestas observadas.

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	95.3	D de Somers	0.906
Discordancia de porcentaje	4.7	Gamma	0.906
Porcentaje ligado	0.0	Tau-a	0.424
Pares	1390262414	c	0.953

Cuando el grado de asociación entre las probabilidades predichas y respuestas observadas es alto, el valor de las predicciones es bueno.

La concordancia de porcentaje es mayor del 95%, luego el modelo es bueno.

Como dijimos anteriormente, SAS no nos muestra por defecto la matriz de confusión. Así que lo que hace es mostrarnos la manera gráfica de la matriz de confusión, es decir, la CURVA ROC.



El área debajo de la curva es de 0.9556. un valor muy alto. Es decir, que si introducimos personas ajenas a la base de datos, el modelo podrá clasificar correctamente a una persona fraudulenta con un 95,56% de probabilidad.

Como hicimos anteriormente con SPSS, nos interesa trabajar con las personas que son potencialmente defraudadoras. Por lo tanto, tenemos que calcular a partir de que punto de corte tendríamos que investigar a los individuos para evitar problemas fraudulentos.

Decidir cuál es el punto de corte optimo se ve en el histograma de la variable IP_FRAUDE.

- EDA de IP_FRAUDE

```
proc univariate data=salida1 all;
VAR IP_FRAUDE_GRAL;
run;
```

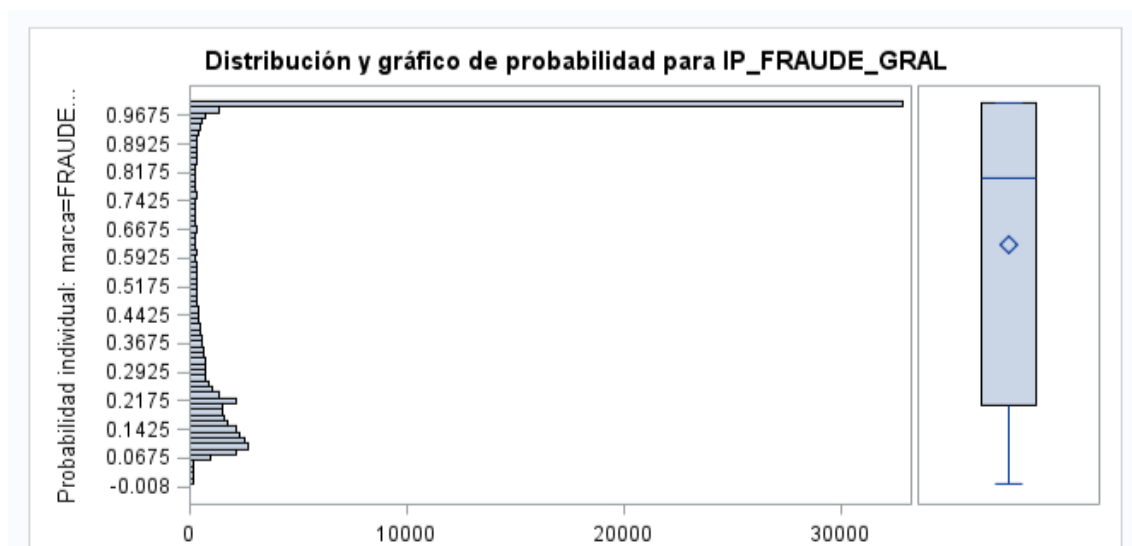
Test para normalidad				
Test	Estadístico		p valor	
Kolmogorov-Smirnov	D	0.257991	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1210.996	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7481.506	Pr > A-Sq	<0.0050

La variable IP_FRAUDE no se distribuye normalmente

Medias recortadas								
Porcentaje recortado en la cola	Número recortado en la cola	Media recortada	Media recortada de error estándar	95% Limites de confianza		DF	t para H0: Mu0=0.00	Pr > t
25.00	19278	0.690197	0.002568	0.685165	0.695230	38552	268.7998	<.0001

Medias winsorizadas								
Porcentaje winsorizado en la cola	Número winsorizado en la cola	Media winsorizada	Media winsorizada de error estándar	95% Limites de confianza		DF	t para H0: Mu0=0.00	Pr > t
25.00	19278	0.648116	0.002568	0.643083	0.653149	38552	252.4095	<.0001

Nos indica a ver si puede haber un atípico. No hay ninguno mayor que 1. No hay atípicos



Para probabilidades pequeñas, hay muchas personas que no defraudan. Sin embargo, para probabilidades altas hay muchas personas defraudadoras.

Sería interesante hallar a partir de que probabilidad investigar a alguna persona que cometa fraudes. En nuestro caso, a partir de 0.8925 hay personas potencialmente defraudadoras.