Tesis para el título de
Licenciado en Ciencias de la Computación

Propuesta para la comparación de la Regresión Logística y el Análisis Discriminante de Fisher en presencia de variables continuas y categóricas

Tesista
Andrés Leonardo Carolo
Ciencias de la Computación
L.U. 2144/86

Directora
Dra. Ana Silvia Haedo
Profesora Adjunta
Departamento de Computación

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

2007

Abstract

When multivariate datasets that have a categorical response variable are analyzed, it is important to classify the cases in classes. To perform the analysis several methods have been proposed, two of them are Logistic Regression and the Fisher Discriminant Analysis.

In several situations a wrong classification of cases belonging to one class can have a larger cost than a wrong classification of cases corresponding to the other class. This happens in studies on illness in which a classification of sick or healthy people, or cases that survived a treatment or died is performed. To classify into the healthy class a case belonging to the sick class can produce serious consequences.

The Logistic Regression and the Fisher Discriminant Analysis will be compared in six datasets, to classify healthy or sick people, or survivor or dead people, and in each datasets the quantity of cases belonging to illness is lower than the cases belonging to wealth and have categorical and continuous prediction variables. In these situations it is frequent the use of the Logistic Regression that permits the usage of categorical prediction variables transformed to dummy, which is a way to express the same information using binary variables. In the assumptions, the Fisher Discriminant Analysis does not permit the use of categorical prediction variables, but in this work they are also used transformed to dummy, violating these assumptions.

A model with the smaller possible size and having a good adjustment is prefered. The logistic adjustment implies to evaluate the coefficient signification to decide what variables stay in the model, having statistics to base it. It will be proposed a criterion to adjust the model in the Fisher Discriminant Analysis and both methods will be compared observing the classification of cases belonging to illness.

Resumen

Cuando se analizan conjuntos de datos multivariados que tienen una variable de respuesta categórica interesa clasificar los casos en clases. Para esto se han propuesto varios métodos, dos de ellos son Regresión Logística y Análisis Discriminante de Fisher.

En varias situaciones puede tener un costo mayor clasificar mal los casos pertenecientes a una clase que clasificar mal los casos pertenecientes a otra. Esto ocurre en los estudios sobre enfermedades, donde se busca clasificar entre sanos o enfermos, o casos que sobrevivieron a un tratamiento o que murieron. Clasificar dentro de la clase de los sanos a un caso perteneciente a la clase de los enfermos puede traer serias consecuencias.

Se compararán la Regresión Logística y el Análisis Discriminante de Fisher en seis conjuntos de datos donde se busca clasificar entre sanos y enfermos, o que sobrevivieron o murieron, y en cada conjunto de datos la cantidad de casos pertenecientes a las enfermedades es menor que la cantidad de casos pertenecientes a la clase de los sanos y tienen variables de predicción categóricas y continuas. En estas situaciones es frecuente el uso de la Regresión Logística que permite el uso de variables de predicción categóricas transformándolas a dummy, que es una forma de representar la misma información por medio de variables binarias. En los supuestos, el Análisis Discriminante de Fisher, no permite el uso de variables de predicción categóricas pero en este trabajo se utilizan y también transformadas a dummy, violando los supuestos.

Es preferible un modelo con la menor dimensión posible y que tenga un buen ajuste, el ajuste logístico implica evaluar la significación de los coeficientes para así decidir que variables quedan en el modelo, para ello se cuentan con estadísticos para fundamentarlos. Se propondrá un criterio para ajustar el modelo en el Análisis Discriminante de Fisher. Se compararán los dos métodos observando las clasificaciones de los casos correspondientes a las enfermedades.

1 Objetivos del trabajo

Encontrar criterios para Comparar Regresión Logística y Análisis Discriminante de Fisher a fin de decidir qué método es aconsejable usar en conjuntos de datos referentes a distintos estudios.

Definir un criterio para ajustar el modelo en Análisis Discriminante de Fisher.

Analizar qué ocurre cuando se violan los supuestos del uso de variables continuas e igualdad de matrices de varianzas-covarianzas en el Análisis Discriminante de Fisher y de distribución normal multivariada en el criterio que se presenta al aplicarlo a conjuntos de datos usados en la Regresión Logística.

2 Introducción

Cuando se analizan conjuntos de datos multivariados que tienen una variable de respuesta categórica interesa clasificar los casos en clases. Se han propuesto varios métodos para clasificar casos en clases, dos de ellos son Regresión Logística y Análisis Discriminante de Fisher. Dado que los conjuntos de datos tienen distintas características, por ejemplo distintas cantidades de casos, distintas cantidades de variables, distintas proporciones de casos pertenecientes a una y a otra clase, distintas cantidades de variables continuas y discretas y distintas distribuciones. Dado que clasificar mal los casos pertenecientes a una clase puede tener un costo mayor que el costo de clasificar mal los casos pertenecientes a otra clase, es de interés saber para cada conjunto de datos, cuál de los dos métodos, Análisis Discriminante de Fisher o Regresión Logística clasifica mejor los casos de mayor costo. Para decidirlo, se destacarán diferencias entre Regresión Logística y Análisis Discriminante de Fisher y cómo tratarlas para comparar estos dos métodos en los mismos conjuntos de datos. Daudin J. J. en [DAU/86] analiza la situación donde lo que más interesa no es el porcentaje de mal clasificados en general, sino que importa más el porcentaje de mal clasificados dentro de un grupo. Press James y Wilson Sandra en [PRE/78], Vlachonikolis I. G. y Marriot F. H. C. en [VLA/82] comparan Regresión Logística y Análisis Discriminante.

El ajuste del modelo logístico implica evaluar la significación de los coeficientes para así decidir qué variables quedan en el modelo, para ello se cuenta con estadísticos para fundamentarlos. Para comparar los resultados del Análisis Discriminante de Fisher y la Regresión Logística, aquí se propone agregar al Análisis Discriminante de Fisher un criterio para decidir si una variable debe quedar dentro o fuera del modelo. Se compararán los métodos analizando la mejor clasificación en cada caso, Daudin J. J. en [DAU/86], Krzanowski W. J. en [KRZ/80] y en [KRZ/83], y Vlachonikolis I. G. y Marrito F. H. C. en [VLA/82] realizaron Análisis Discriminante con ajuste del modelo.

Jonson R. A. y Wichern D. W. en [JOH/02] explican que en el Análisis Discriminante de Fisher, el autor no supone distribución normal multivariada pero si el uso de variables continuas, y también explica que implícitamente supone igualdad de matrices de varianzas-covarianzas de las poblaciones debido al uso de la estimación de la matriz de varianzas-covarianzas combinada. Aunque la distribución normal multivariada junto con una buena separación de las medias de las clases sirve para tener buenas clasificaciones. El criterio para seleccionar variables que se agrega tiene sus fundamentos para variables continuas con distribución normal multivariada. Dado que en numerosos conjuntos de datos, en los que se aplica Regresión Logística, se incluyen variables explicativas categóricas, para su comparación, en el Análisis Discriminante de Fisher se considerarán también variables de predicción categóricas, aunque no sean éstos casos ideales para su funcionamiento. En el ajuste del modelo de Regresión Logística se transforman las variables independientes categóricas en variables dummy. Para comparar los resultados del Análisis Discriminante de Fisher con Regresión Logística y no perder información, también se transformarán en dummy las variables predictoras categóricas. Krzanowski W. J. en [KRZ/75], Press James y Wilson Sandra en [PRE/78] recomiendan no usar Análisis Discriminante cuando involucra variables categóricas, sin embargo, Knoke J.

D. en [KNO/82], considera la Función Discriminante Lineal robusta ante la violación de los supuestos de normalidad multivariada e igualdad de matrices de varianzas-covarianzas.

Se analizarán conjuntos de datos sobre clasificaciones de enfermedades, en los cuales la cantidad de casos con presencia de enfermedad es menor que la cantidad de casos sin presencia. Debido a que el costo de clasificar mal un caso con presencia de enfermedad es mayor que el de clasificar mal un caso sin presencia, y que un modelo descripto por la menor cantidad posible de variables es preferible, entonces, tanto para Regresión Logística como Análisis Discriminante de Fisher, se buscará el modelo de menor dimensión dentro de los que tengan mejor clasificación respecto a la presencia de enfermedad, y se compararán las clasificaciones.

3 Los métodos que se comparan

3.1 Regresión Logística [HOS/89]

3.1.1 Conceptos generales

Los métodos de regresión analizan datos describiendo la relación entre una variable de respuesta (también llamada 'dependiente' o 'variable de salida') y una o más variables de predicción (también llamadas 'independientes', 'exploratorias' o 'covariables'). En este caso la variable de respuesta es discreta con dos o más modalidades. Cuando se busca una relación entre las variables se busca un modelo que sea el que mejor clasifique los casos respecto de la modalidad en la variable respuesta. Vamos a considerar el caso en que la variable de respuesta es binaria (generalmente con valores 0 para representar ausencias o fracasos y 1 para presencias o éxitos).

En Regresión Lineal en muchos conjuntos de datos es necesario aplicar transformaciones a las variables para poder realizar un buen ajuste, debido a que no se puede ajustar el modelo correctamente cuando hay variables de respuesta categóricas, es necesario transformar las variables independientes de forma que la variable de respuesta del nuevo modelo sea una variable continua, por lo tanto la Regresión Logística es un caso especial de transformaciones de variables para realizar ajustes del modelo de regresión, como se ve a continuación.

Se trabaja con n casos donde y_i es el valor de la variable de respuesta, y $x_i = (x_{i1}, x_{i2},..., x_{ip})$ son los valores de las variables de predicción para el i-ésimo caso. La variable de respuesta y_i es una variable aleatoria de Bernoulli con $P(y_i =1) = \pi_i$ y $P(y_i=0) = 1- \pi_i$

$E(y_i/x_i) = 1(\pi_i) + 0(1- \pi_i) = \pi_i = \pi(x_i)$

Si se usa $E(y_i/x_i) = x_i \beta$ entonces los valores del error solo pueden ser:

$\sigma_i = 1- x_i \beta$ si $y_i=1$
$\sigma_i = - x_i \beta$ si $y_i=0$

Los errores de este modelo no tienen distribución normal.

$$\sigma_{y_i}^2 = E\{y_i - E(y_i)\}^2 = (1-\pi_i)^2 \pi_i + (0-\pi_i)(1-\pi_i) = \pi_i(1-\pi_i)$$

La varianza no es constante.

Entonces se debe buscar una transformación para poder realizar un ajuste del modelo. Se utiliza distribución logística:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+....+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+....+\beta_p x_{ip}}}$$

Para linealizar esta función se debe realizar una transformación llamada **logit:**

$$g(x_i) = \ln\left(\frac{\pi(x_{i1},....,x_{ip})}{1-\pi(x_{i1},...,x_{ip})}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

3.1.2 Estimación de parámetros

Se construye una función llamada **función de verosimilitud (likelihood function).** La que expresa la verosimilitud de los datos observados en función de los parámetros desconocidos. Los estimadores de máxima verosimilitud de esos parámetros son aquellos que maximizan la función.

Se utiliza $x_i = x_{i1}, x_{i2}, \ldots, x_{ip}$ y $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$. La contribución de cada caso a la función es:

$$f_i(x_i, y_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Las observaciones se suponen independientes entonces la función de verosimilitud es:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

Buscar los valores de β que maximizan la función es lo mismo que buscar los valores que hacen cero la derivada del logaritmo de la función, es más cómodo trabajar con el logaritmo.

$$\ln L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}$$

La derivada de ln(L(β)) en β_0 es:

$$\sum_{i=0}^{n} [y_i - \pi(x_i)]$$

La derivada de ln(L(β)) en β_i es :

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)]$$

Se denominan **ecuaciones de verosimilitud (likelihood equations).**

Luego de obtener las estimaciones de los parámetros, se utilizan para calcular el valor de π en cada caso para clasificarlos en una de las modalidades.

3.1.3 Deviance

Así como en la Regresión Lineal se busca minimizar los residuos, en la Regresión Logística se busca minimizar la Deviance, que se define de la siguientes manera.

Denotamos valores observados de la variable de salida (dependiente) de la muestra, en un vector Y donde $Y^T = (Y_1, Y_2, \ldots, Y_n)$. Denotamos valores predichos por el modelo, o valores ajustados con

$\hat{y}$ *donde* $\hat{y}^T = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)$

Mientras en la Regresión Lineal el análisis de residuos se basa en los valores de la variable dependiente, tanto para valores observados como los predichos, en la Regresión Logística se base en las variables independientes, como se ve a continuación.

El termino patrón de variables independientes describe un solo conjunto de valores para las variables independientes en el modelo, donde tienen todas el mismo valor, es decir un patrón está formado por los casos que tienen los mismos valores en las variables independientes que forman parte del modelo.

Una vez ajustado el modelo puede haber menos variables independientes, por lo tanto puede haber también menos patrones, con mayor cantidad de casos en algunos o en todos los patrones.

La bondad de ajuste se evalua sobre la constelación de valores ajustados determinado por las variables independientes en el modelo, no la colección total de variables independientes. Supongamos que el modelo ajustado contiene p variables independientes, $x^T=(x_1,x_2,..., x_p)$ y J denota el número de valores distintos de x observados. Si algunos casos tienen el mismo valor de x entonces J<n (n cantidad de casos). Denotaremos el número de casos que tienen $x=x_j$ con m_j, j=1,2,3,....,J. Además $\sum m_j=n$. Y_j denota el número de respuestas positivas, Y=1, entre los m_j casos. Si el número de patrones de variables independientes crece con n entonces cada valor de m_j tenderá a ser pequeño. Estos casos se los llama m-asintóticos. Si fijamos J<n y dejamos a n hacerse grande entonces cada valor de m_j tenderá a hacerse grande.

En el desarrollo asumimos que el modelo ajustado contiene p variables independientes y que forman J patrones (covariate patterns) indexados por j=1,2,...,J.

La clave está en la suma de cuadrados de los residuos. Hay errores binomiales, entonces la varianza del error es una función del promedio condicional:

$$\text{var}(Y_i/x_i) = m_i E(Y_i/x_i) \times [1 - E(Y_i/x_i)] = m_i \pi(x_i)[1 - \pi(x_i)]$$

$$r_j = \frac{(y_j - (m_j \pi_j))}{\sqrt{m_j \pi_j (1 - \pi_j)}}$$

$$d_j = sign(y_j - m_j \pi_j) \times \sqrt{2 \times [y_j |\ln(\pi_j)| + (m_j - y_j) |\ln(1 - \pi_j)|]}$$

$$si\ y_j = 0 \Rightarrow d_j = -\sqrt{2 m_j |\ln(1 - \pi_j)|}$$

$$si\ y_j = m_j \Rightarrow \sqrt{2 m_j |\ln(\pi_j)|}$$

Si $Y_j = 0$ se usa el segundo término, si $m_j=Y_j$ entonces se usa el primer término, si no se usa la fórmula completa.

$$X^2 = \sum_{j=1}^{J} r_j^{\,2}, y, D = \sum_{j=1}^{J} d_j^{\,2}$$

3.1.4 Variables Dummy

La Regresión Logística permite el uso se variables categóricas, para poder utilizarlas hay que pasarlas a variables dummy, si la variable tiene k categorías entonces se crean k-1 variables binarias de la siguiente forma en cada caso del conjunto de datos:

variable categórica	vd_1	...	vd_i	...	vd_{k-1}	
0	0	...	0	...	0	todas valen 0
...	...	...	...	...	...	...
i	0	...	1	...	0	$vd_i = 1$ y las demás valen 0
...	...	...	...	...	...	...
$k-1$	0	...	0	...	1	$vd_{k-1} = 1$ y las demás valen 0

Nunca hay más de una variable dummy correspondientes a la transformación de una misma variable categórica que tenga valor 1 en un mismo caso y debido a que una categoría está representada por todos ceros entonces quedan representadas todas las categorías.

3.1.5 Criterios para ajustar el modelo

3.1.5.1 Test de significación individual

B = estimación del coeficiente, S.E. = error estándar.

$$\left(\frac{B}{S.E.}\right) \sim N(0,1)$$

$$\text{Wald} = \left(\frac{B}{S.E.}\right)^2 \sim \chi_1^2$$

Si Wald < α, se considera un aporte individualmente significativo.
Si Wald > α, se considera que la variable no tiene un aporte individualmente significativo.

3.1.5.2 Test de significación Global

3.1.5.2.1 Test de Hosmer y Lemeshow

El test de Hosmer y Lemeshow ve la diferencia que hay entre los casos observados y los casos esperados, pero dado que se puede clasificar bien los casos con probabilidad cerca de cero y mal los casos con probabilidad cerca de 1 o viceversa, entonces se separan los casos en capas de forma que en una capa entre todos los casos pertenecientes a uno o más patrones no pudiendo quedar casos pertenecientes a un patrón en más de 1 capa. La división se hace en 10 capas quedando 8 grados de libertad. Los casos están agrupados por orden creciente de probabilidad.

$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n_k' \bar{\pi}_k\right)^2}{n_k' \bar{\pi}_k \left(1 - \bar{\pi}_k\right)}$ donde n_k' es el número de patrones en el k^{th} grupo (capa)

$o_k = \sum_{j=1}^{n_k'} y_j$ es la cantidad de casos con $y = 1$ entre los n_k' patrones

$\bar{\pi}_k = \sum_{j=1}^{n_k'} \frac{m_j \hat{\pi}_j}{n_k'}$ es la probabilidad estimada promedio.

$\hat{C} \sim \chi^2_{g-2}$ g = cantidad de grupos (capas)

Este test permite ver la diferencia entre los casos observados y los casos esperados, pero al dividir en grupos y estar ordenados en orden creciente de probabilidad, obliga a que en todos los grupos tenga la mayoría de casos bien clasificados con relación a la probabilidad. Si en un grupo, por ejemplo el último, que debería tener los valores de probabilidad cercanos a 1, hay muchos casos mal clasificados entonces el test no da significativo aunque todos los demás casos de los otros grupos están bien clasificados.

3.1.5.2.2 Aporte a reducir la Deviance

Cuando se ajusta el modelo, la cantidad de casos dentro de un conjunto de datos, es siempre la misma y en la Regresión Logística cuando se compara el Modelo Completo (M. C.) contra el Modelo Reducido (M. R.) para analizar si aporta a reducir la Deviance, los Grados de Libertad se obtienen de la siguiente forma:

G.L. = cantidad de casos – 1 – cantidad de variables.

Por lo tanto, G.L. (M. R.) –G. L. (M. C.) = 1, y el test queda de la siguiente forma:

Deviance (M. R.) – Deviance (M. C.) $\sim \chi^2_{1,\alpha}$

Si Deviance (M. R.) – Deviance (M. C.) > α india poco aporte.

Si Deviance (M. R.) – Deviance (M. C.) < α indica un aporte significativo.

3.2 Análisis Discriminante de Fisher [JOH/02], [FIS/36], [FIS/38], [MAH/36]

3.2.1 Conceptos generales

La idea del Análisis Discriminante de Fisher es pasar de una representación multivariada de los datos a una representación univariada, de modo que las nuevas representaciones de los datos, provenientes de una población, estén separados lo máximo posible de las nuevas representaciones de los datos provenientes de la otra población.

Las condiciones para que los métodos de Análisis Discriminante puedan clasificar correctamente son: que las medias de las poblaciones deben estar separadas lo máximo posible y que las variables independientes sean continuas con distribución normal multivariada. A continuación se ve la separación de las medias de las poblaciones.

Se muestran a continuación las definiciones que involucran el concepto de distancia que serán usadas luego:

$$Sea\ P=(x_1,x_2)\ y\ O=(0,0)\ entonces\ d(O,P)=\sqrt{x_1^2+x_2^2}$$

Donde d(O,P) es la distancia medida como una línea recta desde el punto P hasta el origen O.

$$Sea\ P=(x_1,x_2,.....,x_p)\ y\ O=(0,0,....,0)\ entonces\ d(O,P)=\sqrt{x_1^2+x_2^2+.....+x_p^2}$$

Para la distancia medida como una línea recta entre dos puntos P y Q es como sigue:

$$Sea\ P=(x_1,x_2,....,x_p)\ y\ Q=(y_1,y_2,.....,y_p)\ entonces\ d(P,Q)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+.....+(x_p-y_p)^2}$$

Esta definición de distancia no es satisfactoria para propósitos de estadística porque cada coordenada representa medidas distintas entonces tienen variabilidad muy diferentes en magnitud, por eso se estandarizan las coordenadas como sigue:

$$x_1^*=\frac{x_1}{\sqrt{s_{11}}},\quad x_2^*=\frac{x_2}{\sqrt{s_{22}}}$$

Entonces la distancia de P hacia O se calcula usando los valores estandariazados como sigue:

$$d(O,P)=\sqrt{(x_1^*)^2+(x_2^*)^2}=\sqrt{\left(\frac{x_1}{\sqrt{s_{11}}}\right)^2+\left(\frac{x_2}{\sqrt{s_{22}}}\right)^2}=\sqrt{\frac{x_1^2}{s_{11}}+\frac{x_2^2}{s_{22}}}$$

La distancia entre un punto P y un punto fijo Q, donde las coordenadas varían independientemente, se define como sigue:

$$Sea\ P=(x_1,x_2)\ y\ Q=(y_1,y_2)\ etonces\ d(P,Q)=\sqrt{\frac{(x_1-y_1)^2}{s_{11}}+\frac{(x_2-y_2)^2}{s_{22}}}$$

Donde s_{11} y s_{22} son las varianzas de x_1 y x_2 reaspectivamente.

$$Sea\ P=(x_1,x_2,....,x_p)\ y\ Q=(y_1,y_2,....,y_p)\ entonces$$

$$d(P,Q)=\sqrt{\frac{(x_1-y_1)^2}{s_{11}}+\frac{(x_2-y_2)^2}{s_{22}}+....+\frac{(x_p-y_p)^2}{s_{pp}}}$$

En muchos casos esta medida de distancia no sirve porque supone independencia. Cuando hay correlación entre las variables hay que considerar que se corren los ejes, entonces la distancia desde un punto P hacia un punto fijo Q es de la forma:

$$P=(x_1,x_2)\ y\ Q=(y_1,y_2)$$

$$d(P,Q)=\sqrt{a_{11}(x_1-y_1)^2+2a_{12}(x_1-y_1)(x_2-y_2)+a_{22}(x_2-y_2)^2}$$

$$Sea\ P=(x_1,x_2,....,x_p),\ O=(0,0,....,0)\ y\ Q=(y_1,y_2,....,y_p)\ entonces$$

$$d(O;P)=\sqrt{a_{11}x_1^2+a_{22}x_2^2+....+a_{pp}x_p^2+2a_{12}x_1x_2+2a_{13}x_1x_3+.....+2a_{p-1,p}x_{p-1}x_p}$$

$$d(P,Q)=\sqrt{[a_{11}(x_1-y_1)^2+a_{22}(x_2-y_2)^2+....+a_{pp}(x_p-y_p)^2+2a_{12}(x_1-y_1)(x_2-y_2)+2a_{13}(x_1-y_1)(x_3-y_3)+...+2a_{p-1,p}(x_{p-1}-y_{p-1})(x_p-y_p)]}$$

Los coeficientes a_{ij} pueden considerarse en una matriz:

$$\begin{bmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{12} & a_{22} & ... & a_{2p} \\ ... & .... & ... & \\ a_{1p} & a_{2p} & ... & a_{pp} \end{bmatrix}$$

$$(\text{distancia})^2=a_{11}x_1^2+a_{22}x_2^2+.....+a_{pp}x_p^2+2(a_{12}x_1x_2+a_{13}x_1x_3+.....+a_{p-1,p}x_{p-1}x_p)$$

$$0<(\text{distancia})^2=[x_1,x_2,....,x_p]\begin{bmatrix} a_{11} & a_{12} & .... & a_{1p} \\ a_{21} & a_{22} & .... & a_{2p} \\ .... & .... & ..... & ..... \\ a_{p1} & a_{p2} & .... & a_{pp} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_p \end{bmatrix}$$

$$0<(\text{distancia})^2=x^tAx\ \text{para}\ x\neq 0$$

Entonces A es una matriz simétrica de p×p, que es definida positiva.

Como x^tAx es el cuadrado de la distancia desde el punto x hacia el origen, entonces el cuadrado de la distancia desde x hacia otro punto ❍ es $(x-❍)^tA(x-❍)$.

En el Análisis Discriminante de Fisher la regla de asignación es a menudo una función lineal de mediciones que maximiza la separación entre grupos relativa a la variabilidad dentro de los grupos. Por esto se ven a continuación las siguientes desigualdades:

3.2.2 Desigualdad de Cauchy-Schwarz

Sea b y d dos vectores de p×1, entonces $(b^t d)^2 \leq (b^t b)(d^t d)$ con la igualdad si y solo si b=cd o d=cb con c constante.

3.2.3 Extensión de la Desigualdad de Cauchy-Schwarz

Sea b y d dos vectores de p×1, y B una matriz definida positiva de p×p, entonces:

$(b^t d)^2 \leq (b^t B b)(d^t B^{-1} d)$ con igualdad si y solo si $b=cB^{-1}d$ o $d=cBb$ con c constante.

3.2.4 Lema de Maximización

Sea B una matriz definida positiva de p×p, y d un vector de p×1. Entonces para un vector x de p×1 arbitrario distinto de 0

$$\max_{x \neq 0} \frac{(x^t d)^2}{x^t B x} = d^t B^{-1} d$$

La distancia que se va a usar está representada en unidades de desvío estándar, por lo tanto la definición queda como sigue:

$$separación = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}$$

En donde s_y es la estimación de la matriz de varianzas-covarianzas combinada.
La siguiente combinación lineal

$$\hat{y} = \hat{a}^t x = (\bar{x}_1 - \bar{x}_2) S_c^{-1} x$$

Maximiza la siguiente razón:

$$\frac{\begin{pmatrix}\text{distancia al cuadrado entre}\\ \text{promedios muestrales de } y\end{pmatrix}}{(\text{varianza muestral de } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{a}^t \bar{x}_1 - \hat{a}^t \bar{x}_2)^2}{\hat{a}^t S_c \hat{a}} = \frac{(\hat{a}^t d)^2}{\hat{a}^t S_c \hat{a}}$$

$\forall \hat{a}$ y donde $d = (\bar{x}_1 - \bar{x}_2)$

El máximo la razón es $D^2 = (\bar{x}_1 - \bar{x}_2)^t S_c^{-1} (\bar{x}_1 - \bar{x}_2)$

3.2.5 Regla de asignación de casos a una de las clases

Asignar x_0 a la clase π_1 si se cumple:

$$\hat{y}_0 = (\bar{x}_1 - \bar{x}_2)' S_c^{-1} x_0 \geq \hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2) S_c^{-1} (\bar{x}_1 + \bar{x}_2) \quad \text{o} \quad \hat{y}_0 - \hat{m} \geq 0$$

Asignar x_0 a la clase π_2 si se cumple:

$$\hat{y}_0 < \hat{m} \quad \text{o} \quad \hat{y}_0 - \hat{m} < 0$$

Se proyectan en un eje los casos y el punto de corte m, por lo tanto si las medias de las poblaciones tienen una buena separación y las poblaciones tienen distribuciones normales multivariadas, entonces a lo sumo solo se van a superponer en el eje de proyección los casos correspondientes a las colas de las distribuciones los que representan el 5% total de los casos. Si fueran otras distribuciones se podrían llegar a superponer una cantidad mucho mayor de casos que los que corresponden a las colas de las distribuciones normales, y alguna o ambas poblaciones tendrían demasiados casos mal clasificados. En la realidad hay muchos conjuntos de datos donde las variables independientes no cumplen con una distribución normal multivariada, y conjuntos de datos donde hay variables independientes que son categóricas.

3.2.6 Criterio para ajustar el modelo en Análisis Discriminante de Fisher [JOH/02], [MON/02], [CHA/00]

3.2.6.1 Distribuciones y Grados de Libertad de SCE y SCD

n_1 = cantidad de casos pertenecientes a la clase 1.

n_2 = cantidad de casos pertenecientes a la clase 2.

$n = n_1 + n_2$ = cantidad total de casos.

x_{1i} = i – ésimo caso perteneciente a la clase 1.

$(x_{1i})_j$ = valor correspondiente a la variable en la posición j, del caso en posición i, perteneciente a la clase 1.

x_{2i} = i – ésimo caso perteneciente a la clase 2.

$(x_{1i})_j$ = valor correspondiente a la variable en la posición j, del caso en posición i, perteneciente a la clase 2.

$\bar{x}_1$ = $\bar{x}$ de los casos pertenecientes a la clase 1.

$(\bar{x}_1)_j$ = valor correspondiente a la variable en la posición j, de $\bar{x}$ de los casos pertenecientes a la clase 1.

$\bar{x}_2$ = $\bar{x}$ de los casos pertenecientes a la clase 2.

$(\bar{x}_2)_j$ = valor correspondiente a la variable en la posición j, de $\bar{x}$ de los casos pertenecientes a la clase 2.

p = cantidad de variables.

$$d_j^2 = (x_j - \bar{x})' S^{-1} (x_j - \bar{x}) \quad j = 1,\ 2,\ldots..,n$$

$X_1,\ldots..,\ X_n$ son observaciones muestrales.

Cuando la población tiene distribución normal multivariada y ambos n y n-p son mayores que 25 o 30, cada una de las distancias debería comportarse de la siguiente forma: $d_i^2 \sim \chi^2$ *donde* $1 \leq i \leq n$

Estadístico suficiente

$x_1, x_2, ..., x_n$ son muestras aleatorias de una población con distribución normal multivariada con media μ y covarianzas Σ. Entonces $\bar{x}$ y S son estadísticos suficientes.
La importacia de estadísticos suficientes para una población normal es que la información acerca de μ y Σ en la matriz de datos X están contenidas en $\bar{x}$ y S, sin considerar el tamaño n de la muestra.

Distribuciones

$X \sim \mathrm{N_p}(\mu, \Sigma)$ *con* $|\Sigma| > 0$. *Entonces*
$(a)\ (x-\mu)^t \Sigma^{-1}(x-\mu) \sim \chi^2_{\mathrm{p}}$
(b) La distribución $\mathrm{N_p}(\mu, \Sigma)$ asigna probabilidad 1-α al elipsoide $\{x : (x-\mu)^t \Sigma^{-1}(x-\mu) \le \chi^2_p(\alpha)\}$ donde $\chi^2_p(\alpha)$ denota el percentil en la posición superior 100α de la distribución χ^2_p.

También puede verse como suma de cantidad p de naturales al cuadrado estandarizados.

Sea $\mathrm{Z} = \dfrac{(Y-\mu)}{\sigma}$. Si $Y \sim N(\mu, \sigma^2)$ entonces $\mathrm{Z}^2 \sim \chi^2_1$

Sean $Y_1, Y_2, ..., Y_n$ variables aleatorias independientes normalmente distribuidas,

tales que $E(Y_i) = \mu_i$ y $\mathrm{Var}(Y_i) = \sigma_i^2$, y sean $\mathrm{Z_i} = \dfrac{Y_i - \mu_i}{\sigma_i}$

entonces $\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{Z_i^2} \sim \chi^2_{\mathrm{n}}$

Es decir, la suma de los cuadrados de n variables normalmente distribuidas y estandarizadas tiene una distribució χ con n grados de libertad, y la suma de variables aleatorias χ^2 también tiene una distribución χ^2.

Sean v~ χ^2_m y w~ χ^2_n. Si v y w son independientes entonces $\dfrac{v/m}{w/n} \sim F_{m,\,n}$.

Esto quiere decir que la razón de dos variables independientes χ^2, dividida cada una entre sus respectivos g. l., sigue una distribución F.

A continuación se ven las definiciones usadas de SCD y SCE

3.2.6.1.1 Suma de Cuadrados Entre (SCE)

$d^2\left(\overline{x_1}, \overline{x_2}\right)$ es la distancia al cuadrado en el hiperplano entre los vectores de medias de las clases 1 y 2.

$$\bar{x}_1 - \bar{\bar{x}} = \bar{x}_1 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_1}{n_1+n_2} - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} = \left(\frac{n_2}{n_1+n_2}\right)\left(\bar{x}_1 - \bar{x}_2\right)$$

$$\operatorname{cov}\left(\bar{x}_1 - \bar{\bar{x}}\right) = \operatorname{cov}\left[\left(\frac{n_2}{n_1+n_2}\right)\left(\bar{x}_1 - \bar{x}_2\right)\right] = \left[\frac{n_2^2}{(n_1+n_2)^2}\right]\operatorname{cov}\left(\bar{x}_1 - \bar{x}_2\right) = \left[\frac{n_2^2}{(n_1+n_2)^2}\right]\left(\frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma\right) =$$

$$= \left[\frac{n_2^2}{(n_1+n_2)^2}\right]\left(\frac{n_2+n_1}{n_1 n_2}\right)\Sigma = \frac{n_2}{n_1(n_1+n_2)}\Sigma$$

$$\operatorname{cov}\left(\bar{x}_1 - \bar{x}_2\right) = \operatorname{cov}\left(\bar{x}_1\right) + \operatorname{cov}\left(\bar{x}_2\right) = \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

$$Si\ \hat{\Sigma} = S_c \Rightarrow \hat{\operatorname{cov}}\left(\bar{x}_1 - \bar{x}_2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times S_c$$

$$\overline{x_2} - \bar{\bar{x}} = \overline{x_2} - \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1+n_2} = \frac{n_1\overline{x_2} + n_2\overline{x_2}}{n_1+n_2} - \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1+n_2} = \left(\frac{n_1}{n_1+n_2}\right)\left(\overline{x_2} - \overline{x_1}\right)$$

$$\operatorname{cov}\left(\overline{x_2} - \bar{\bar{x}}\right) = \operatorname{cov}\left[\left(\frac{n_1}{n_1+n_2}\right)\left(\overline{x_2} - \overline{x_1}\right)\right] = \left[\frac{n_1^2}{(n_1+n_2)^2}\right]\operatorname{cov}\left(\overline{x_2} - \overline{x_1}\right) = \left[\frac{n_1^2}{(n_1+n_2)^2}\right]\left(\frac{1}{n_2} + \frac{1}{n_1}\right)\Sigma =$$

$$= \left[\frac{n_1^2}{(n_1+n_2)^2}\right]\left(\frac{n_1+n_2}{n_1 n_2}\right)\Sigma = \left[\frac{n_1}{n_2(n_1+n_2)}\right]\Sigma$$

$$\operatorname{cov}\left(\overline{x_2} - \overline{x_1}\right) = \operatorname{cov}\left(\overline{x_2}\right) + \operatorname{cov}\left(\overline{x_1}\right) = \left(\frac{1}{n_2} + \frac{1}{n_1}\right)\Sigma$$

$$Si\ \hat{\Sigma} = S_c \Rightarrow \hat{\operatorname{cov}}\left(\bar{x}_2 - \bar{x}_1\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \times S_c$$

Suponiendo igualdad de matrices de varianzas-covarianzas tenemos:

$$d^2\left(\overline{x_1}, \overline{x_2}\right) = \left(\overline{x_1} - \overline{x_2}\right)^t\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{\Sigma}\right]^{-1}\left(\overline{x_1} - \overline{x_2}\right) = \left(\overline{x_1} - \overline{x_2}\right)^t\left[\left(\frac{n_1+n_2}{n_1 n_2}\right)\hat{\Sigma}\right]^{-1}\left(\overline{x_1} - \overline{x_2}\right) =$$

$$= \left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\overline{x_1} - \overline{x_2}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1} - \overline{x_2}\right) = \left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\frac{n_1+n_2}{n_2}\right)\left(\overline{x_1} - \bar{\bar{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\frac{n_1+n_2}{n_2}\right)\left(\overline{x_1} - \bar{\bar{x}}\right) =$$

$$= \left[\frac{n_1(n_1+n_2)}{n_2}\right]\left(\overline{x_1} - \bar{\bar{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1} - \bar{\bar{x}}\right)$$

$$d^2\left(\overline{x_1},\ \overline{x_2}\right)=\left(\overline{x_1}-\overline{x_2}\right)^t\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{\Sigma}\right]^{-1}\left(\overline{x_1}-\overline{x_2}\right)=\left(\overline{x_1}-\overline{x_2}\right)^t\left[\left(\frac{n_1+n_2}{n_1 n_2}\right)\hat{\Sigma}\right]^{-1}\left(\overline{x_1}-\overline{x_2}\right)=$$

$$=\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\overline{x_1}-\overline{x_2}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1}-\overline{x_2}\right)=\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\frac{n_1+n_2}{n_1}\right)\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\frac{n_1+n_2}{n_1}\right)\left(\overline{x_2}-\overline{\overline{x}}\right)=$$

$$=\left[\frac{n_2\left(n_1+n_2\right)}{n_1}\right]\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_2}-\overline{\overline{x}}\right)$$

$$d^2\left(\overline{x_1},\ \overline{x_2}\right)=\left(\frac{n_1+n_2}{n_1+n_2}\right)d^2\left(\overline{x_1},\overline{x_2}\right)=\left(\frac{n_1}{n_1+n_2}\right)d^2\left(\overline{x_1},\ \overline{x_2}\right)+\left(\frac{n_2}{n_1+n_2}\right)d^2\left(\overline{x_1},\ \overline{x_2}\right)$$

$$\left(\frac{n_1}{n_1+n_2}\right)d^2\left(\overline{x_1},\ \overline{x_2}\right)=\left(\frac{n_1}{n_1+n_2}\right)\left[\frac{n_2\left(n_1+n_2\right)}{n_1}\right]\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_2}-\overline{\overline{x}}\right)=n_2\times\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_2}-\overline{\overline{x}}\right)$$

$$\left(\frac{n_2}{n_1+n_2}\right)d^2\left(\overline{x_1},\ \overline{x_2}\right)=\left(\frac{n_2}{n_1+n_2}\right)\left[\frac{n_1\left(n_1+n_2\right)}{n_2}\right]\left(\overline{x_1}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1}-\overline{\overline{x}}\right)=n_1\times\left(\overline{x_1}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1}-\overline{\overline{x}}\right)$$

$$\Rightarrow d^2\left(\overline{x_1},\ \overline{x_2}\right)=\left(\overline{x_1}-\overline{x_2}\right)^t\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{\Sigma}\right]^{-1}\left(\overline{x_1}-\overline{x_2}\right)=$$

$$=n_1\times\left(\overline{x_1}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_1}-\overline{\overline{x}}\right)+n_2\times\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(\hat{\Sigma}\right)^{-1}\left(\overline{x_2}-\overline{\overline{x}}\right)=SCE$$

$$Si\ \hat{\Sigma}=S_c\Rightarrow\left(\overline{x_1}-\overline{x_2}\right)^t\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_c\right]^{-1}\left(\overline{x_1}-\overline{x_2}\right)=$$

$$=n_1\times\left(\overline{x_1}-\overline{\overline{x}}\right)^t\left(S_c\right)^{-1}\left(\overline{x_1}-\overline{\overline{x}}\right)+n_2\times\left(\overline{x_2}-\overline{\overline{x}}\right)^t\left(S_c\right)^{-1}\left(\overline{x_2}-\overline{\overline{x}}\right)=SCE$$

Dado que son p variables se lo puede ver como suma de p naturales estadarizados al cuadrado. La estandarización se realiza con la matriz de varianzas covarianzas combinada inversa.

$\Rightarrow SCE\sim\chi_p^2$

Otra forma de verlo es la siguiente:

El estadístico T^2 para el test H_0:μ_1-μ_2=0 es:

$$T^2=\left[\overline{x_1}-\overline{x_2}\right]^t\left[\frac{1}{n_1}S_1+\frac{1}{n_2}S_2\right]^{-1}\left[\overline{x_1}-\overline{x_2}\right]$$

Se compara con χ_2^2 (0.05) donde α=0.05.

$\Rightarrow SCE\sim\chi_p^2$

$$Si\ se\ usa\ \ \mathrm{cov}\left(\overline{x_1}\right)=\frac{1}{n_1}S_1\ \ y\ \ \mathrm{cov}\left(\overline{x_2}\right)=\frac{1}{n_2}S_2\Rightarrow\mathrm{cov}\left(\overline{x_1}\right)+\mathrm{cov}\left(\overline{x_2}\right)=\frac{1}{n_1}S_1+\frac{1}{n_2}S_2$$

Otra forma de verlo es que se está haciendo la suma de p variables con distribución normal estandarizadas y elevadas al cuadrado.

3.2.6.1.2 Suma de Cuadrados Dentro (SCD)

Dado que $(x-\mu)' \Sigma^{-1} (x-\mu) \sim \chi^2_{\mathrm{p}}$ y por estadísticos suficiente tenemos

$$\left.\begin{aligned} SCD_1 &= \sum_{j=1}^{n_1} \left(x_{1j} - \overline{x_1}\right)' S_c^{-1} \left(x_{1j} - \overline{x_1}\right) \sim \chi^2_{\mathrm{n_1}p} \\ SCD_2 &= \sum_{j=1}^{n_2} \left(x_{2j} - \overline{x_2}\right)' S_c^{-1} \left(x_{2j} - \overline{x_2}\right) \sim \chi^2_{\mathrm{n_2}p} \end{aligned}\right\} SCD = SCD_1 + SCD_2 \sim \chi^2_{(\mathrm{n_1}+n_2)p} \Rightarrow SCD \sim \chi^2_{\mathrm{np}}$$

Otra forma de verlo es que se está haciendo la suma de n×p variables con distribución normal estandarizadas y elevadas al cuadrado.

3.2.6.2 Estadístico

SCD(MC) $\sim \chi^2_{np}$ y SCD(MR) $\sim \chi^2_{n(p-1)}$

SCD(MC) - SCD(MR) $\sim \chi^2_{np-n(p-1)} = \chi^2_{np-np+n} = \chi^2_n$

SCE(MC) $\sim \chi^2_p$ y SCE(MR) $\sim \chi^2_{p-1}$

SCE(MC) – SCE(MR) $\sim \chi^2_{p-(p-1)} = \chi^2_1$

$$\frac{\left(\frac{SCE(MC) - SCE(MR)}{1}\right)}{\left(\frac{SCD(MC) - SCD(MR)}{n}\right)} \sim F_{1,\ n}$$

Comparación para dejar fuera del modelo una variable

$$\frac{\left(\frac{SCE(MC) - SCE(MR)}{1}\right)}{\left(\frac{SCD(MC) - SCD(MR)}{n}\right)} < \text{F de la tabla}$$

Comparación para ingresar dentro del modelo una variable

$$\frac{\left(\frac{SCE(MC) - SCE(MR)}{1}\right)}{\left(\frac{SCD(MC) - SCD(MR)}{n}\right)} > \text{F de la tabla}$$

3.2.7 Análisis Discriminante de Fisher con variables continuas y binarias

En la vida real hay muchos estudios que involucran la clasificación de pacientes como sanos o enfermos. Esos estudios necesitan representar información importante con variables categóricas. Las variables categóricas no tienen un significado numérico (por ejemplo sexo, tipo de cobertura médica), no tienen un orden entre categorías, por lo tanto si se las codifica con valores numéricos no sería una representación real. Para solucionarlo se utiliza una representación que no tenga un significado numérico. En la Regresión Logística se soluciona con la transformación de las variables categóricas, con más de 2 categorías, a variables dummy, las que representan presencia o ausencia de una categoría.

Si se usan variables categóricas en el Análisis Discriminante de Fisher entonces se está decidiendo arbitrariamente que categoría está más cerca que otra dentro de esa variable, por ejemplo si para una variable hay 4 categorías y se codifican con 0, 1, 2 y 3, dado que Análisis Discriminante de Fisher usa distancias, con esta codificación se decide arbitrariamente que la categoría representada por el 3 está más cerca de la categoría representada por el 2 que de la categoría representada por el 0, está a una distancia 3 veces mayor, entonces estamos decidiendo el aporte de la variable a la estimación de los casos. Si se cambia la codificación y la categoría representada por el 3 se codifica con 1 y la representada por 1 se codifica con 3, entonces se cambia la relación de las distancias que había entre las categorías en esa variable y por lo tanto se cambia el aporte de esa variable a la estimación de los casos. Esto puede cambiar la clasificación de los casos.

Esto se resuelve usando variables dummy. En el Anexo se puede ver que cuando se transforman las variables categóricas a variables dummy (de la misma forma como está explicado en el capítulo de Regresión Logística), la clasificación de los casos no cambian aunque se cambie la codificación de las variables binarias.

No se considera en este trabajo si hay distorsiones, en la función de Fisher, que puedan ser causadas por el uso de variables binarias en la matriz de varianzas-covarianzas, producidas en las varianzas y en las covarianzas que involucran a alguna variable binaria.

Para el cálculo del estadístico se supone distribución normal multivariada, pero se busca aplicarlo en condiciones en las que se violan los supuestos, es decir cuando en los conjuntos hay variables categóricas.

Para SCE y SCD también existe el problema de la codificación arbitraria en las variables categóricas con más de 2 categorías, al usar variables dummy, y al cambiar la codificación de las variables binarias no cambian los valores de SCE y SCD, por lo tanto no cambia el criterio que se usa para ajustar el modelo, esto puede verse en el Anexo, por lo tanto, para no perder la información representada por las variables categóricas es preferible usarlas pero transformadas a variables dummy.

En este trabajo no se considera que ocurre con las distribuciones de SCE y SCD al usar variables dummy, solo se aplica y se comparan las Tablas de Clasificaciones en Validación Cruzada con las de la Regresión Logística, poniendo énfasis en los porcentajes de casos bien clasificados pertenecientes a las clases de las enfermedades.

Las SCE y SCD están definidas con la distancia de Mahalanobis que utiliza la Matriz de Varianzas-Covarianzas Combinada. En este trabajo no se consideran las distorsiones en las varianzas y en las covarianzas producidas por el uso de variables categóricas transformadas a dummy.

La Regresión Logística es muy utilizada en los tipos de estudios que se usan en este trabajo, permitiendo representar la información de las variables categóricas. En la bibliografía se encuentran ajuste hechos con variables dummy.

Para saber si se obtienen buenos resultados al aplicar las variables dummy en el Análisis Discriminante de Fisher, se comparan los resultados con los resultados obtenidos por la Regresión Logística, este último método tiene fundamentado el uso de variables dummy.

4 Experimento

Para cada conjunto de datos se sigue el mismo proceso.

4.1 Limpieza y transformación de los datos

Primero se realiza un análisis descriptivo. Se busca si hay casos extremos y se decide si van a formar parte del análisis, para esto se utilizan Histogramas para las variables continuas y Tablas de Distribución de Frecuencias para las variables categóricas, si se encuentran valores extremos o que se puedan considerar no válidos se investiga si fueron utilizados en las publicaciones correspondientes y cómo fueron utilizados.

Después de verificar que las variables categóricas y binarias tienen los valores para las categorías correspondientes, se realiza la transformación a las variables dummy como se explica en el capítulo de Regresión Logística.

Para cada conjunto de datos queda un conjunto donde las variables categóricas fueron reemplazadas por las variables dummy correspondientes y ese conjunto de datos será utilizado tanto en la Regresión Logística como en el Análisis Discriminante de Fisher.

4.2 Ajuste de los modelos

Se aplica un método paso a paso para seleccionar variables, en el cual las variables son seleccionadas para inclusión o exclusión en el modelo en forma secuencial basado únicamente en un criterio estadístico. Hay dos versiones del procedimiento paso a paso: (1) inclusión hacia delante, con un test para eliminación hacia atrás, y (2) eliminación hacia atrás, seguido por un test para inclusión hacia adelante. En este trabajo se utiliza el (2) para la Regresión Logística y una modificación para el Análisis Discriminante de Fisher.

4.2.1 Ajuste del modelo en la Regresión Logística

En la Regresión Logística se cuenta con un test estadístico para ver si la estimación del coeficiente de una variable es individualmente significativa, es decir si es suficientemente lejano del 0. Utilizando ese test se elige para remover una variable a una de las que tenga el valor de significación del test > 0.05 y con uno de los valores más altos (no necesariamente el más grande). Una vez seleccionada la variable que queda fuera se ajusta el modelo y se compara el modelo completo que incluye a la variable en cuestión y el modelo reducido que no la incluye, para eso se utiliza un test de significación global, se busca ver si la variable que no era individualmente significativa lo es teniendo en cuenta el aporte de todas las variables, por eso se ve el aporte de la variable en cuestión al modelo una vez que las otras variables ya fueron incluidas en el modelo. Si este test da significativo entonces se decide mantener la variable dentro del modelo, si el test no es significativo pero cercano a 0.05 entonces se debe ver si el test de significación individual es o no significativo pero cercano a cero y se debe decidir si mantener o dejar fuera del modelo la variable. Si el test global no es significativo entonces se deja la variable fuera del modelo. Si bien la decisión final se toma por el test de significación global, es importante el test de significación individual ya que la mayoría de las variables que no son significativas individualmente tampoco lo son a nivel global, esto puede verse siguiendo los ajustes de los modelos en los ejemplos en el anexo.

Las variables se remueven de a una y a partir de la segunda variable que fue seleccionada para quedar fuera del modelo, se ven si las variables que quedaron fuera del modelo en pasos anteriores pueden ingresar al mismo. Para eso se realiza un ajuste del modelo con la variable incluida y se compara de la misma forma

como se explicó en el párrafo anterior. Si una variable pasó nuevamente a formar parte del modelo, esta puede ser removida nuevamente del modelo en alguno de los próximos pasos.

4.2.2 Ajuste del modelo en el Análisis Discriminante de Fisher

En el Análisis Discriminante de Fisher, en este trabajo se utiliza solamente el criterio para ajustar el modelo que se ve en el capítulo correspondiente, por eso se comienza con todas las variables dentro del modelo y por orden de las variables, de a una, se ajusta un modelo sin la variable llamado Modelo Reducido, y el modelo que contiene a la variable se llama Modelo Completo, se aplica el estadístico, si es significativo entonces la variable queda fuera del modelo y si no es significativo entonces la variable permanece en el modelo. Cuando debe ser evaluada por primera vez la variable que está en la i-ésima posición, quiere decir que ya fueron evaluadas al menos una vez las i-1 variables anteriores y todavía no fueron evaluadas las variables que se encuentran en las posiciones i+1 hasta la posición p. Si la variable que se encuentra en la i-ésima posición deja de formar parte del modelo, se deben evaluar nuevamente las i-1 variables anteriores. Si al evaluar una variable que se encuentra en la posición j, con $j < i$, esa variable deja de formar parte del modelo o pasa a formar parte del modelo (es decir produce un cambio en el modelo) no hace que se vuelvan a evaluar las j-1 variables anteriores, sino que continúa con la variable que se encuentra en la posición j+1, con $j+1 < i$, si $j+1 = i$ entonces pasa a evaluar por primera vez a la i+1-ésima variable.

Al evaluar una de las i-1 variables anteriores a la i-ésima, puede ocurrir que la variable estaba fuera del modelo, entonces se ajusta el modelo incluyéndola pasando a ser el Modelo Completo y el modelo que no la incluye pasa a ser el Modelo Reducido, se calcula el estadístico y se comparan los modelos para ver si la variable pasó a formar parte del modelo o no. Si la variable forma parte del modelo entonces se ajusta un modelo que no contenga la variable, pasando a ser ese el Modelo Reducido y el que la incluye pasa a ser el Modelo Completo, se calcula el estadístico, se compara para ver si la variable quedó fuera del modelo o sigue formando parte del mismo.

El recorrido de evaluar las variables que ya fueron evaluadas en pasos anteriores se hace sólo a partir del cambio en el modelo producido porque sale una variable que es evaluada por primera vez. Es decir los cambios en los modelos producidos porque ingresan o salen del modelo las variables que fueron evaluadas anteriormente no hacen volver atrás para ver las variables evaluadas anteriormente.

4.3 Validación Cruzada

Tanto para la Regresión logística como para el Análisis Discriminante de Fisher, se utiliza validación cruzada una vez que se tienen las variables que forman parte del modelo. Para hacerlo se deja fuera de a un caso por vez, se ajusta el modelo, se obtiene la estimación del caso que quedó fuera utilizando la función obtenida en el ajuste y se clasifica el caso que quedó fuera. En los resultados se muestra la Tabla de Clasificación en Validación Cruzada y no la clasificación caso por caso. En las comparaciones entre la Regresión Logística y el Análisis Discriminante de Fisher se deben ver las Tablas de Clasificaciones en validación Cruzada y no las otras.

Herramientas utilizadas

Para la Regresión Logística se utilizaron SPSS 12.0 para el ajuste del modelo y para validación cruzada WEKA (Waikato Environment for Knowledge Analysis) C 1999-2000.

Para el Análisis Discriminante de Fisher se utilizaron Java 1.3.1 y Excel.

5 Resultados

5.1 Ejemplo 1 Diabetic Ketoacidosis [VIL/95]

Frutos de los avances en el conocimiento de la diabetes, la cetoacidosis ha surgido como la causa más frecuente de muerte en pacientes diabéticos, de modo que un pronóstico temprano al principio de una crisis pudiera ser importante. Se dispone de los datos de 106 pacientes con cetoacidosis admitidos en la unidad de cuidado intensivo del Hospital Clínica de Barcelona colectados durante 6 años. De estos 106 individuos, 93 sobrevivieron y 13 murieron. Se registraron las variables 'age' (edad) y 'blood glucose' (glucosa en sangre) que son continuas, y 'previous diagnoses of diabetes' (diagnósticos previos de diabetes, si=1, no=0) y 'deep comma degree state' (grado de estado de coma profundo yes=1, no=0), son categóricas binarias.
Consideraremos a dos subpoblaciones: la subpoblación 1 corresponde a los pacientes que sobrevivieron y la subpoblación 2 a los que murieron.

Análisis Discriminante de Fisher

Función

$$\text{PREVIOUS DIAGNOSIS OF DIABETES} = \begin{cases} 0 \text{ - no} \\ 1 \text{ - si} \end{cases}$$

$$\text{Si PREVIOUS DIAGNOSIS OF DIABETES} \Rightarrow Y = (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE) + (\beta_3 \times 1) = (\beta_3) + (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE)$$

$$\text{Si no PREVIOUS DIAGNOSIS OF DIABETES} \Rightarrow Y = (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE) + (\beta_3 \times 0) = (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE)$$

Donde $\beta_1 = -0.08727, \beta_2 = -0.00495, \beta_3 = 3.5937$

Tabla de Clasificaciones

	predicho 1	2	
original 1	78	15	93
original 2	1	12	13
	79	27	106

Cantidad de casos bien clasificados	90
Cantidad de casos mal clasificados	16

Porcentajes

	predicho 1	2
original 1	83.87	16.13
original 2	7.69	92.31

Porcentaje de casos bien clasificados	84.91
Porcentaje de casos mal clasificados	15.09

Tabla de Clasificación en Validación Cruzada

Tabla de Clasificaciones

	predicho 1	2	
original 1	77	16	93
original 2	1	12	13
	78	28	106

Cantidad de casos bien clasificados	89
Cantidad de casos mal clasificados	17

Porcentajes

	predicho 1	2
original 1	82.80	17.20
original 2	7.69	92.31

Porcentaje de casos bien clasificados	83.96
Porcentaje de casos mal clasificados	16.04

Regresión Logística

Función

$$\text{PREVIOUS DIAGNOSIS OF DIABETES} = \begin{cases} 0 \text{ - no} \\ 1 \text{ - si} \end{cases}$$

$$\text{Si PREVIOUS DIAGNOSIS OF DIABETES} \Rightarrow Y = \beta_0 + (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE) + (\beta_3 \times 1) =$$
$$= (\beta_0 + \beta_3) + (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE)$$

$$\text{Si no PREVIOUS DIAGNOSIS OF DIABETES} \Rightarrow Y = \beta_0 + (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE) + (\beta_3 \times 0) =$$
$$= \beta_0 + (\beta_1 \times AGE) + (\beta_2 \times BLOODGLUCOSE)$$

Donde β_0 = constante = -10.628, $\beta_1 = 0.109$, $\beta_2 = 0.005$, $\beta_3 = 2.238$

Classification Table [a]

			Predicted		
			SUBP		Percentage
	Observed		1	2	Correct
Step 1	SUBP	1	90	3	96.8
		2	6	7	53.8
	Overall Percentage				91.5

a. The cut value is .500

Tabla de Clasificación en validación cruzada

Tabla de clasificaciones

	Predichos		
	1	2	total
clase menor original 1	88	5	93
clase mayor original 2	6	7	13

Porcentajes de Tabla de clasificaciones

	Predichos	
	1	2
clase menor original 1	94.62	5.38
clase mayor original 2	46.15	53.85

porcentajes de instancias bien clasificadas 89.62
porcentajes de instancias mal clasificadas 10.38

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 92.31%,** mientras que en la **Regresión logística es 53.85%.**

5.2 Ejemplo 2 "psychosocial influences in breast cancer" [KRZ/80], [NUÑ/03]

Los datos se puede ver en:

http://www.tibs.org/biometrics/datasets/QLX.DAT.

Este conjunto datos corresponde a un estudio sobre las influencias sicosociales en el cáncer de pecho, realizado en el King's Collage Hospital, London. El mismo contiene información sobre 137 mujeres con tumor de pecho, 78 de éstos benignos, $\pi 1$, y 59 malignos, $\pi 2$. Parte de este conjunto de datos fue utilizado en uno de los ejemplos en Krzanowski (1975); aquí nosotros seleccionamos un conjunto de variables diferente, elegido para reflejar el rango completo de las variables aleatorias. Las variables aleatorias 1-6 eran observaciones sicosociales: acting out hostility (actúa con hostilidad); criticism of others (crítica hacia los otros); paranoid hostility (hostilidad paranoica); self-criticism (autocrítica); guilt (culpabilidad); direction of hostility (dirección de la hostilidad). Éstas fueron codificadas en un rango 0-10 y fueron tratadas como variantes aleatorias continuas. La variable aleatoria 7 era age of menarche (edad de la menarca), también continua. Las variables 8 y 9 eran binarias: presence/absence of allergy and thyroid (presencia/ausencia de la alergia y de la tiroides); mientras que las variables aleatorias 10 y 11 eran categóricas, cada una con tres estados: temper (carácter), feelings (sentimientos). Había así siete variables continuas, dos variables categóricas binarias y dos con tres categorías.

Cada variable categórica primero fue reemplazada por dos variables binarias llamadas dummy. Había así seis variables binarias en total: x1 y x2 denotan las dos originales allergy (alergia) y thyroid (tiroides), x3 y x4 son las dos variables binarias que corresponden a la variable categórica temper (carácter), mientras que x5 y x6 son las dos variables binarias que corresponden a la otra variable categórica feelings (sentimientos). Así, si los tres estados de temper (carácter) fueron codificados con 1, 2 y 3 entonces el estado 1 fue transformado en x3 = 1, x4 = 0; El estado 2 fue transformado en x3 = 0, x4 = 1 y el estado 3 fue transformado en x3 = 0, x4 = 0.

En este trabajo se utilizan los siguientes nombres para las variables categóricas: las variables binarias siguen siendo allergy (alergia) y thyroid (tiroides), la variable categórica transformada a dummy temper (carácter) pasa a ser temper1 y temper2, y la variable categórica transformada a dummy feelings (sentimientos) pasa a ser feelings1 y feelings2.

Análisis Discriminante de Fisher

Función

$$\text{ALEERGY} = \begin{cases} 0 - \text{ausencia (no)} \\ 1 - \text{presencia (si)} \end{cases} \quad \text{THYROID} = \begin{cases} 0 - \text{ausencia (no)} \\ 1 - \text{presencia (si)} \end{cases} \quad \text{TEMPER1} = \begin{cases} 0 \text{ - ausencia (no)} \\ 1 \text{ - presencia (si)} \end{cases}$$

$$\text{FEELINGS1} = \begin{cases} 0 - \text{ausencia (no)} \\ 1 - \text{presencia (si)} \end{cases}$$

Si THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 1) + (\beta_4 \times 1) =$
$= \beta_1 + \beta_2 + \beta_3 + \beta_4$

Si THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 1) + (\beta_4 \times 0) =$
$= \beta_1 + \beta_2 + \beta_3$

Si THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (\beta_4 \times 1) =$
$= +\beta_1 + \beta_2 + \beta_4$

Si THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (\beta_4 \times 0) =$
$= \beta_1 + \beta_2$

Si THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 1) + (\beta_4 \times 1) =$
$= \beta_1 + \beta_3 + \beta_4$

Si THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 1) + (\beta_4 \times 0) =$
$= \beta_1 + \beta_3$

Si THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (\beta_4 \times 1) =$
$= \beta_1 + \beta_4$

Si THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (\beta_4 \times 0) =$
$= \beta_1$

Si no THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 1) + (\beta_4 \times 1) =$
$= \beta_2 + \beta_3 + \beta_4$

Si no THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 1) + (\beta_4 \times 0) =$
$= \beta_2 + \beta_3$

Si no THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 0) + (\beta_4 \times 1) =$
$= \beta_0 + \beta_2 + \beta_4$

Si no THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 0) + (\beta_4 \times 0) =$
$= \beta_2$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 1) + (\beta_4 \times 1) =$
$= \beta_3 + \beta_4$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\wedge$ no FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 1) + (\beta_4 \times 0) =$
$= \beta_3$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 0) + (\beta_4 \times 1) =$
$= \beta_4$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\wedge$ FEELINGS1 $\Rightarrow y = (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 0) + (\beta_4 \times 0) =$
$= 0$

Donde, $\beta_1 = 1.8967$, $\beta_2 = -1.5242$, $\beta_3 = 2.2386$, $\beta_4 = 0.7579$

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	60	18	78
original 1	18	41	59
	78	59	137

Cantidad de casos bien clasificados	101
Cantidad de casos mal clasificados	36

Porcentajes

	predicho 0	predicho 1
original 0	76.92	23.08
original 1	30.51	69.49

Porcentaje de casos bien clasificados	73.72
Porcentaje de casos mal clasificados	26.28

Tabla de Clasificación en Validación Cruzada

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	59	19	78
original 1	18	41	59
	77	60	137

Cantidad de casos bien clasificados	100
Cantidad de casos mal clasificados	37

Porcentajes

	predicho 0	predicho 1
original 0	75.64	24.36
original 1	30.51	69.49

Porcentaje de casos bien clasificados	72.99
Porcentaje de casos mal clasificados	27.01

Regresión Logística

Función

$$\text{ALEERGY} = \begin{cases} 0 - \text{ausencia (no)} \\ 1 - \text{presencia (si)} \end{cases} \quad \text{THYROID} = \begin{cases} 0 - \text{ausencia (no)} \\ 1 - \text{presencia (si)} \end{cases} \quad \text{TEMPER1} = \begin{cases} 0 \text{ - ausencia (no)} \\ 1 - \text{presencia(si)} \end{cases}$$

Si THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Si THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\beta_2 \times 1) + (\beta_3 \times 0) = \beta_0 + \beta_1 + \beta_2$

Si THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 1) = \beta_0 + \beta_1 + \beta_3$

Si THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\beta_2 \times 0) + (\beta_3 \times 0) = \beta_0 + \beta_1$

Si no THYROID $\wedge$ ALLERGY $\wedge$ TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 1) = \beta_0 + \beta_2 + \beta_3$

Si no THYROID $\wedge$ ALLERGY $\wedge$ no TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\beta_2 \times 1) + (\beta_3 \times 0) = \beta_0 + \beta_2$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 1) = \beta_0 + \beta_3$

Si no THYROID $\wedge$ no ALLERGY $\wedge$ no TEMPER1 $\Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\beta_2 \times 0) + (\beta_3 \times 0) = \beta_0$

Donde $\beta_0 = cons\tan te = -1.096, \beta_1 = 2.596, \beta_2 = -1.564, \beta_3 = 2.132$

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	65	13	83.3
		malignant	22	37	62.7
	Overall Percentage				74.5

a. The cut value is .500

Tabla de clasificación en validación cruzada

Tabla de clasificaciones

	Predichos 0	1	total
clase menor original 0	64	14	78
clase mayor original 1	22	37	59

Porcentajes de Tabla de clasificaciones

	Predichos 0	1
clase menor original 0	82.05	17.95
clase mayor original 1	37.29	62.71

porcentajes de instancias bien clasificadas 73.72
porcentajes de instancias mal clasificadas 26.28

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 69.49%**, mientras que en la **Regresión logística es 62.71%.**

5.3 Ejemplo 3 "Intensive Care Unit (ICU)" [HOS/89]

NOMBRE: The ICU Data (ICU.DAT)
KEYWORDS: Regresión Logística
DIMENSION: 200 observaciones, 21 variables

FUENTE: Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition. Los derechos de autor son de John Wiley & Sons Inc. Los datos fueron recogidos en el Centro médico Baystate en Springfield, Massachusetts.

RESUMEN DESCRIPTIVO:

El conjunto de datos de ICU consiste en una muestra de 200 personas que son parte de un estudio mucho más grande sobre la sobrevivencia de pacientes después de la admisión a una unidad de cuidado intensivo del adulto (ICU). La meta principal de este estudio fue desarrollar el modelo de regresión logística para predecir la probabilidad de sobrevida de los pacientes que llegan al hospital y estudiar los factores de riesgo asociados con la mortalidad en ICU. Un número de publicaciones aparecieron, enfocando varias facetas del problema. El lector que desea aprender más sobre los aspectos clínicos de este estudio debe comenzar con Lemeshow, Teres, Avrunin, y Pastides (1988).

LISTA DE VARIABLES:

Número de Variable	Nombre	Códigos/Valores	Abreviaciones
1	Código de identificación	ID Number	ID
1	Estado Vital	0 = sobrevivió 1 = murió	STA
3	Edad	Años	AGE
4	Sexo	0 = Masculino 1 = Femenino	SEX
5	Raza	1 = Blanca 2 = Negra 3 = Otras	RACE
6	Servicio en la admisión de ICU	0 = Médico 1 = Quirúrgico	SER
7	Presencia de Cancer	0 = No 1 = Si	CAN
8	Historia de falla renal crónica	0 = No 1 = Si	CRN
9	Probable infección en la admisión de ICU	0 = No 1 = Si	INF

10	CPR antes de la admisión de ICU	0 = No 1 = Si	CPR
11	Presión sistólica en la admisión de ICU	mm Hg	SYS
12	Ritmo cardíaco en la admisión de ICU	Beats/min	HRA
13	Admisión previa a un ICU En los 6 meses anteriores	0 = No 1 = Si	PRE
14	Tipo de admisión	0 = Electivo 1 = Emergencia	TYP
15	Hueso largo, múltiple, cuello, área simple, o fractura de la cadera	0 = No 1 = Si	FRA
16	PO2 (Presión de oxígeno) de los gases de sangre iniciales	0 = > 60 1 = < 60	PO2
17	PH (Potencial Hidrógeno) de los gases de sangre iniciales	0 => 7.25 1 =< 7.25	PH
18	PCO2 (Presión de Dióxido de Carbono)de los gases de sangre Iniciales	0 = < 45 1 = > 45	PCO
19	Bicarbonato de los gases de sangre iniciales	0 = > 18 1 = < 18	BIC
20	Creatinina de los gases de sangre iniciales	0 = < 2.0 1 = > 2.0	CRE
21	Nivel de conciencia en la admisión de ICU	0 = No Coma or Estupor 1 = Estupor profundo 2 = Coma	LOC

NOTAS PEDAGÓGICAS:

Estos datos fueron usados para ajustar un modelo de regresión logística ordinario y para varios ejercicios que incluían la regresión logística.

REFERENCIAS:

1. Hosmer and Lemeshow, Applied Logistic Regression, Wiley, (1989).

2. Lemeshow, S., Teres, D., Avrunin, J. S., Pastides, H. (1988). Predicting the Outcome of Intensive Care Unit Patients. Journal of the American.

Análisis Discriminante de Fisher

Función

$$\text{Cancer Part of Present Problem(CAN)} = \begin{cases} 0 \text{ - NO} \\ 1 \text{ - YES} \end{cases}$$

$$\text{Level of Consciousness at ICU Admission(LOC2)} = \begin{cases} 0 \text{ - otro} \\ 1 \text{ - comma} \end{cases}$$

$$\text{Level of Consciousness at ICU Admission(LOC1)} = \begin{cases} 0 \text{ - otro} \\ 1 \text{ - Deep stupor} \end{cases}$$

$$\text{Type of Admission(TYP)} = \begin{cases} 0 \text{ - Elective} \\ 1 \text{ - Emergency} \end{cases}$$

$$\text{PCO2 from initial Blood Gases(PCO)} = \begin{cases} 0 < 45 \\ 1 > 45 \end{cases}$$

$$\text{PH from Initial Blood Gases(PH)} = \begin{cases} 0 > 7.25 \\ 1 < 7.25 \end{cases}$$

$$y = (\text{AGE} \times \beta_1) + (\text{SYS} \times \beta_7) + (\text{aporte de las variables dummy})$$

aporte de las variables dummy =

$$= \begin{cases} \text{Si CAN} \Rightarrow (\beta_2 \times 1) + \text{aporte de las variables dummy 2} = \beta_2 + \text{aporte de las variables dummy 2} \\ \text{Si no CAN} \Rightarrow (\beta_2 \times 0) + \text{aporte de las variables dummy 2} = 0 + \text{aporte de las variables dummy 2} \end{cases}$$

aporte de las variables dummy 2 =

$$= \begin{cases} \text{Si Deep stupor} \Rightarrow (\beta_3 \times 1) + (\beta_4 \times 0) + \text{aporte de las variables dummy 3} = \beta_3 + \text{aporte de las variables dummy 3} \\ \text{Si Comma} \Rightarrow (\beta_3 \times 0) + (\beta_4 \times 1) + \text{aporte de las variables dummy 3} = \beta_4 + \text{aporte de las variables dummy 3} \\ \text{Si no Comma} \wedge \text{no Deep stupor} \Rightarrow (\beta_3 \times 0) + (\beta_4 \times 0) + \text{aporte de las variables dummy 3} = \\ \qquad = 0 + \text{aporte de las variables dummy 3} \end{cases}$$

aporte de las variables dummy 3 =

$$= \begin{cases} \text{Si PCO2} > 45 \Rightarrow (\beta_5 \times 1) + \text{aporte de las variables dummy 4} = \beta_5 + \text{aporte de las variables dummy 4} \\ \text{Si PCO2} < 45 \Rightarrow (\beta_5 \times 0) + \text{aporte de las variables dummy 4} = 0 + \text{aporte de las variables dummy 4} \end{cases}$$

aporte de las variables dummy 4 =

$$= \begin{cases} \text{Si PH} < 7.25 \Rightarrow (\beta_6 \times 1) + \text{aporte de las variables dummy 5} = \beta_6 + \text{aporte de las variables dummy 5} \\ \text{Si PH} > 7.25 \Rightarrow (\beta_6 \times 0) + \text{aporte de las variables dummy 5} = 0 + \text{aporte de las variables dummy 5} \end{cases}$$

aporte de las variables dummy 5 =

$$= \begin{cases} \text{Si tipo de admisión Emergencia} \Rightarrow (\beta_8 \times 1) = \beta_8 \\ \text{Si tipo de admisión Electiva} \Rightarrow (\beta_8 \times 0) = 0 \end{cases}$$

Donde $\beta_1 = -0.0440$, $\beta_2 = -2.1980$, $\beta_3 = -8.4935$, $\beta_4 = -5.5430$, $\beta_5 = 1.9156$, $\beta_6 = -2.1138$, $\beta_7 = 0.0168$, $\beta_8 = -2.2630$

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	143	17	160
original 1	14	26	40
	157	43	200

Cantidad de casos bien clasificados 169
Cantidad de casos mal clasificados 31

Porcentajes

	predicho 0	predicho 1
original 0	89.37	10.63
original 1	35.00	65.00

Porcentaje de casos bien clasificados 84.50
Porcentaje de casos mal clasificados 15.50

Tabla de Clasificación en validación cruzada

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	140	20	160
original 1	16	24	40
	156	44	200

Cantidad de casos bien clasificados 164
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	predicho 1
original 0	87.50	12.50
original 1	40.00	60.00

Porcentaje de casos bien clasificados 82.00
Porcentaje de casos mal clasificados 18.00

Regresión Logística

Función

$$\text{Cancer Part of Present Problem(CAN)} = \begin{cases} 0 - \text{NO} \\ 1 - \text{YES} \end{cases}$$

$$\text{Level of Consciousness at ICU Admission(LOC1)} = \begin{cases} 0 - \text{other} \\ 1 - \text{Deep stupor} \end{cases}$$

$$\text{Level of Consciousness at ICU Admission(LOC2)} = \begin{cases} 0 - \text{other} \\ 1 - \text{comma} \end{cases} \quad \text{Si LOC1} = 0 \text{ y LOC2} = 0 \text{ no comma or stupor}$$

$$\text{Type of Admission(TYP)} = \begin{cases} 0 - \text{Elective} \\ 1 - \text{Emergency} \end{cases}$$

Si CAN $\wedge$ comma $\wedge$ tipo de amisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 1) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_2 + \beta_3 + \beta_5) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si CAN $\wedge$ comma $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 1) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_2 + \beta_3) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si CAN $\wedge$ Deep stupor $\wedge$ tipo de admisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 1) =$
$= (\beta_0 + \beta_2 + \beta_5 + \beta_6) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si CAN $\wedge$ Deep stupor $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 1) =$
$= (\beta_0 + \beta_2 + \beta_6) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si CAN $\wedge$ no comma o stupor $\wedge$ tipo de amisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_2 + \beta_5) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si CAN $\wedge$ no comma o stupor $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 1) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_2) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ comma $\wedge$ tipo de admisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 1) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_3 + \beta_5) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ comma $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 1) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_3) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ Deep stupor $\wedge$ tipo de admisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 1) =$
$= (\beta_0 + \beta_5 + \beta_6) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ Deep stupor $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 1) =$
$= (\beta_0 + \beta_6) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ no comma o stupor $\wedge$ tipo de admisión Emergency $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 1) + (\beta_6 \times 0) =$
$= (\beta_0 + \beta_5) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Si no CAN $\wedge$ no comma o stupor $\wedge$ tipo de admisión Elective $\Rightarrow y = \beta_0 + (AGE \times \beta_1) + (\beta_2 \times 0) + (\beta_3 \times 0) + (SYS \times \beta_4) + (\beta_5 \times 0) + (\beta_6 \times 0) =$
$= (\beta_0) + (AGE \times \beta_1) + (SYS \times \beta_4)$

Donde β_0 = constante = 15.669, β_1 = 0.038, β_2 = –2.597, β_3 = –2.417, β_4= –0.018, β_5 = –3.848, β_6 = –12.240

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	157	3	98.1
		Died	23	17	42.5
	Overall Percentage				87.0

a. The cut value is .500

Tabla de Clasificaciones en Validación Cruzada

Tabla de clasificaciones

	Predichos		
	0	1	total
clase menor original 0	157	3	160
clase mayor original 1	26	14	40

Porcentajes de Tabla de clasificaciones

	Predichos	
	0	1
clase menor original 0	98.12	1.88
clase mayor original 1	65.00	35.00
porcentajes de instancias bien clasificadas	85.50	
porcentajes de instancias mal clasificadas	14.50	

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 60.00%**, mientras que en la **Regresión logística es 35.00%.**

5.4 Ejemplo 4 "Low Birth Weight Data" [HOS/89]

NOMBRE: Datos de bajo peso al nacer (LOWBWT.DAT)
KEYWORDS: Regresión Logística
DIMENSION: 189 observaciones, 11 variables

FUENTE: Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition. Los derechos de autor son de John Wiley & Sons Inc. y deben ser reconocidos y usados de acuerdo a ello. Los datos fueron colectados en el centro médico de Baystate, Springfield, Massachusetts durante 1986.

RESUMEN DESCRIPTIVO:

La meta de este estudio era identificar los factores de riesgo asociados con el nacimiento de un bebé de bajo peso (que pesa menos de 2500 gramos). Los datos fueron colectados sobre 189 mujeres, 59 de quienes tenían bebés con bajo peso de nacimiento y 130 de las cuales tenían bebés con peso de nacimiento normal. Se pensó en la importancia de 4 variables edad, el peso de la mujer en su período menstrual pasado, la raza, y el número de visita médicas durante el primer trimestre del embarazo.

NOTA:

Este conjunto de datos consiste en los datos completos. Un conjunto de datos apareado creado de estos datos de bajo peso del nacimiento puede ser encontrado en lowbwtm11.dat y un conjunto de datos apareado 3 a 1 (3-1 matched) creado de los datos de bajo peso del nacimiento se puede encontrar en mlowbwt.dat.

LISTA DE VARIABLES:

Número	Variable	Abbreviation
1	Código de identificación	ID
2	Bajo peso al nacer (0 = Peso del nacimiento >= 2500g 1 = Peso del nacimiento < 2500g)	LOW
3	Edad de la madre en años	AGE
4	Peso en libras en el último período menstrual	LWT
5	Raza (1 = Blanca, 2 = Negra, 3 = Otras)	RACE
6	Estado fumador durante el embarazo (1 = Si, 0 = No)	SMOKE
7	Historia del trabajo prematuro (0 = Ninguno 1 = Uno, etc.)	PTL
8	Historia de la hipertensión (1 = Si, 0 = No)	HT
9	Presencia de la irritabilidad uterina (1 = Si, 0 = No)	UI
10	Número de las visitas del médico durante el primer trimestre (0 = Ninguna, 1 = Una, 2 = Dos, etc.)	FTV
11	Peso del nacimiento en gramos	BWT

NOTAS PEDAGÓGICAS:

Estos datos fueron utilizados como ejemplo para ajustar modelos de regresión logística múltiple.

HISTORIA DETRÁS DE LOS DATOS:

El bajo peso al nacer es un resultado que preocupa a los médicos hace años. Esto se debe al hecho de que las proporciones de mortalidad infantil y de defectos al nacer son muy altas para los niños de bajo peso de nacimiento. El comportamiento de una mujer durante el embarazo (incluyendo la dieta, los hábitos de fumar, y recepción de cuidado prenatal) pueden alterar fuertemente las chances de tener al bebé en término y consecuentemente de peso normal.

Las variables identificadas en la hoja de código dada en la tabla se muestran por estar asociadas al bajo peso del nacimiento en la literatura obstétrica. El objetivo de este estudio fue comprobar si esas variables eran importantes en la población que era atendida en el centro médico donde los datos fueron colectados.

Referencias:

Hosmer and Lemeshow, Applied Logistic Regression, Wiley, (1989).

ACLARACION: la variable Birth Weight in Grams (BWT) tiene correlación con la variable Low Birth Weight (LOW) por lo tanto no es incluida en el análisis.

Análisis Discriminante de Fisher

Función

$$\text{History of Hypertension(HT)} = \begin{cases} 0 - \text{NO} \\ 1 - \text{SI} \end{cases}$$

Weight in Pounds at the Last Menstrual Period(LWT) es continua

$$\text{Race(RACE1)} = \begin{cases} 0 - \text{otra} \\ 1 - \text{raza blanca} \end{cases}$$

$$\text{Smoking Status During Pregnancy(SMOKE)} = \begin{cases} 0 - \text{NO} \\ 1 - \text{SI} \end{cases}$$

$$\text{Presence of Uterine Irritability(UI)} = \begin{cases} 0 - \text{NO} \\ 1 - \text{SI} \end{cases}$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_1 + \beta_3 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_1 + \beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_1 + \beta_3 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_1 + \beta_3) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_1 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_1 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{noRAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_1 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_1) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_3 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_3 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_3) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\text{LWT} \times \beta_2)$$

Donde $\beta_1 = -2.0158$, $\beta_2 = 0.0139$, $\beta_3 = 0.9758$, $\beta_4 = -1.0429$, $\beta_5 = -1.0189$

Tabla de Clasificaciones

	predicho 0	1	
original 0	94	36	130
original 1	21	38	59
	115	74	189

Cantidad de casos bien clasificados	132
Cantidad de casos mal clasificados	57

Porcentajes

	predicho 0	1
original 0	72.31	27.69
original 1	35.59	64.41

Porcentaje de casos bien clasificados	69.84
Porcentaje de casos mal clasificados	30.16

Tabla de Clasificación en validación cruzada

Tabla de Clasificaciones

	predicho 0	1	
original 0	91	39	130
original 1	26	33	59
	117	72	189

Cantidad de casos bien clasificados	124
Cantidad de casos mal clasificados	65

Porcentajes

	predicho 0	1
original 0	70.00	30.00
original 1	44.07	55.93

Porcentaje de casos bien clasificados	65.61
Porcentaje de casos mal clasificados	34.39

Regresión Logística

Función

$$\text{History of Hypertension(HT)} = \begin{cases} 0 \text{ - NO} \\ 1 \text{ - SI} \end{cases}$$

Weight in Pounds at the Last Menstrual Period(LWT) es continua

$$\text{Race(RACE1)} = \begin{cases} 0 \text{ - otra} \\ 1 \text{ - raza blanca} \end{cases}$$

$$\text{Smoking Status During Pregnancy(SMOKE)} = \begin{cases} 0 \text{ - NO} \\ 1 \text{ - SI} \end{cases}$$

$$\text{Presence of Uterine Irritability(UI)} = \begin{cases} 0 \text{ - NO} \\ 1 \text{ - SI} \end{cases}$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_0 + \beta_1 + \beta_3 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_0 + \beta_1 + \beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_0 + \beta_1 + \beta_3 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_0 + \beta_1 + \beta_3) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_0 + \beta_1 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_0 + \beta_1 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{noRAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_0 + \beta_1 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 1) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_0 + \beta_1) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_0 + \beta_3 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_0 + \beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_0 + \beta_3 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_0 + \beta_3) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 1) + (\beta_5 \times 1) = (\beta_0 + \beta_4 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 1) + (\beta_4 \times 1) + (\beta_5 \times 0) = (\beta_0 + \beta_3 + \beta_4) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 1) = (\beta_0 + \beta_5) + (\text{LWT} \times \beta_2)$$

$$\text{Si no HT} \wedge \text{no RAZA BLANCA} \wedge \text{no SMOKE} \wedge \text{no UI} \Rightarrow y = \beta_0 + (\beta_1 \times 0) + (\text{LWT} \times \beta_2) + (\beta_3 \times 0) + (\beta_4 \times 0) + (\beta_5 \times 0) = (\beta_0) + (\text{LWT} \times \beta_2)$$

Donde β_0 = constante = 3.677, $\beta_1 = -1.853$, $\beta_2 = -0.015$, $\beta_3 = 1.065$, $\beta_4 = -1.091$, $\beta_5 = -0.893$

Classification Table[a]

Observed			Predicted		
			LOW		Percentage Correct
			0	1	
Step 1	LOW	0	117	13	90.0
		1	38	21	35.6
	Overall Percentage				73.0

a. The cut value is .500

Tabla de Clasificación en validación cruzada

Tabla de clasificaciones

	Predichos		
	0	1	total
clase menor original 0	112	18	130
clase mayor original 1	39	20	59

Porcentajes de Tabla de clasificaciones

	Predichos	
	0	1
clase menor original 0	86.15	13.85
clase mayor original 1	66.10	33.90

porcentajes de instancias bien clasificadas 69.84
porcentajes de instancias mal clasificadas 30.16

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 55.93%**, mientras que en la **Regresión logística es 33.90%.**

5.5 Ejemplo 5 "South African Heart Disease"

La explicación del estudio y las variables se encuentran en: southafrica.txt y en: http://www.stat.purdue.edu/~josong/stat598m/southafrica.html
Los datos se encuentran en: SAheart_data.txt.

La explicación del problema se encuentra en :
http://www.stat.purdue.edu/~josong/stat598m/southafrica.html

Este es un ejemplo de regresión logística sobre enfermedad cardiaca en Sudáfrica. El objetivo del estudio: ""Establecer la intensidad de los factores de riesgo de isquemia coronaria en una región de alta incidencia". Los datos representan a hombres de raza blanca entre 15 y 64 años, la variable de la respuesta es la presencia o la ausencia del infarto del miocardio durante el tiempo de este estudio.

Las variables predictoras son : sbp (presión sistólica), tobacco (tabaco), ldl (lipoproteína ldl), famhist (historia faminliar), obesity (obesidad), alcohol, age (edad), donde famhist es binaria y las demás son continuas.

La variable de respuesta es chd.

Análisis discriminante de Fisher

Función

$$\text{FAMISHISTP} = \begin{cases} 0 \text{ - ausente (no)} \\ 1 - \text{presente (si)} \end{cases}$$

$$\text{Si FAMISHISTP} \Rightarrow y = (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_3 \times 1) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE) =$$
$$= (\beta_3) + (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE)$$

$$\text{Si no FAMISHISTP} \Rightarrow y = (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_3 \times 0) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE) =$$
$$= (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE)$$

Donde $\beta_1 = -0.0957, \beta_2 = -0.1738, \beta_3 = -0.9730, \beta_4 = -0.0318, \beta_5 = -0.0424$

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	208	94	302
original 1	45	115	160
	253	209	462

Cantidad de casos bien clasificados	323
Cantidad de casos mal clasificados	139

Porcentajes

	predicho 0	predicho 1
original 0	68.87	31.13
original 1	28.13	71.87

Porcentaje de casos bien clasificados	69.91
Porcentaje de casos mal clasificados	30.09

Tabla de Clasificación en validación Cruzada

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	208	94	302
original 1	46	114	160
	254	208	462

Cantidad de casos bien clasificados 322
Cantidad de casos mal clasificados 140

Porcentajes

	predicho 0	predicho 1
original 0	68.87	31.13
original 1	28.75	71.25

Porcentaje de casos bien clasificados 69.70
Porcentaje de casos mal clasificados 30.30

Regresión Logística

Función

$$\text{FAMISHISTP} = \begin{cases} 0 \text{ - ausente (no)} \\ 1 - \text{presente (si)} \end{cases}$$

$$\text{Si FAMISHISTP} \Rightarrow y = \beta_0 + (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_3 \times 1) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE) =$$
$$= (\beta_0 + \beta_3) + (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE)$$

$$\text{Si no FAMISHISTP} \Rightarrow y = \beta_0 + (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_3 \times 0) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE) =$$
$$= \beta_0 + (\beta_1 \times \text{TOBACCO}) + (\beta_2 \times \text{LDL}) + (\beta_4 \times \text{TYPEA}) + (\beta_5 \times AGE)$$

Donde β_0 = constante = -5.538, $\beta_1 = 0.080$, $\beta_2 = 0.162$, $\beta_3 = -0.908$, $\beta_4 = 0.037$, $\beta_5 = 0.050$

Classification Table[a]

			Predicted		
			CHD		Percentage Correct
	Observed		0	1	
Step 1	CHD	0	256	46	84.8
		1	73	87	54.4
	Overall Percentage				74.2

a. The cut value is .500

Tabla de Clasificación en validación cruzada

Tabla de clasificaciones

	Predichos 0	Predichos 1	total
clase menor original 0	254	48	302
clase mayor original 1	74	86	160

Porcentajes de Tabla de clasificaciones

	Predichos 0	Predichos 1
clase menor original 0	84.11	15.89
clase mayor original 1	46.25	53.75

porcentajes de instancias bien clasificadas 73.59
porcentajes de instancias mal clasificadas 26.41

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 71.25%**, mientras que en la **Regresión logística es 53.75%.**

5.6 Ejemplo 6 "French Wine - Dementia Study"

University of Massachusetts
School of Public Health and Health Sciences
BioEpi 640
Examen Final
Lunes 17 de Mayo de 1999
Dr. Hosmer

Este examen se basa en un estudio de los factores de riesgo asociados a demencia (es decir, enfermedad de Alzheimer) en una muestra aleatoria de 2076 sujetos mayores en la región del sudoeste de Francia. Para ser incluidos en este estudio los sujetos debían ser mayores de 60 años de edad y libres de demencia (basada en una batería de pruebas psicológicas) en el momento de ser incluidos en el estudio. El resultado de interés era si el sujeto tenía muestras clínicas de demencia después de tres años de seguimiento. El primer factor de riesgo considerado fue el consumo de vino que fue agrupada en tres categorías según lo mostrado en la tabla 1 junto con las otras variables bajo consideración:

Los datos del archivo wine_data.txt consisten en todos los sujetos con demencia y una muestra escogida al azar de 10 por ciento de sujetos sin demencia.

Tabla 1: Descripción de las variables usadas en "Estudio de demencia asociado al consumo de vino en Francia"

Variable	Descripción	Códigos/Valores
ID	Código de identificación	1 - 272
AGE	Edad	Years
WINE	Consumo de vino	0 = No consume vino 1 = hasta 1/4 litro por día 2 = más de 1/4 litro por día
MMSE	Mini examen del estado mental	0-30
HIGHBP	Presión diastólica >90	1 = Presión arterial alta 0 = Sin Presión arterial alta
T3DEMEN	Episodios de demencia en el lapso de 3 años después de el ingreso al estudio	1 = Yes 0 = No

Estos datos están disponibles en el Web site del curso como wine_data.txt

Análisis discriminante de Fisher

Función

$$\text{WINE2} = \begin{cases} 1 \text{ - Más que 1/4 litro por día} \\ 0 \text{ – Hasta 1/4 litro o menos por día} \end{cases}$$

Si Más que 1/4 litro por día $\Rightarrow y = (\beta_1 \times \text{AGE}) + (\beta_2 \times 1) + (\beta_3 \times \text{MINI MENTAL STATE EXAM}) =$
$= (\beta_2) + (\beta_1 \times \text{AGE}) + (\beta_3 \times \text{MINI MENTAL STATE EXAM})$

Si Hasta 1/4 litro o menos por día $\Rightarrow y = (\beta_1 \times \text{AGE}) + (\beta_2 \times 0) + (\beta_3 \times \text{MINI MENTAL STATE EXAM}) =$
$= (\beta_1 \times \text{AGE}) + (\beta_3 \times \text{MINI MENTAL STATE EXAM})$

Donde $\beta_1 = -0.1240, \beta_2 = 0.9081, \beta_3 = 0.3474$

Tabla de Clasificaciones

	predicho 0	1	
original 0	164	36	200
original 1	19	53	72
	183	89	272

Cantidad de casos bien clasificados	217
Cantidad de casos mal clasificados	55

Porcentajes

	predicho 0	1
original 0	82.00	18.00
original 1	26.39	73.61

Porcentaje de casos bien clasificados	79.78
Porcentaje de casos mal clasificados	20.22

Tabla de Clasificación en validación cruzada

Tabla de Clasificaciones

	predicho 0	1	
original 0	164	36	200
original 1	19	53	72
	183	89	272

Cantidad de casos bien clasificados	217
Cantidad de casos mal clasificados	55

Porcentajes

	predicho 0	1
original 0	82.00	18.00
original 1	26.39	73.61

Porcentaje de casos bien clasificados	79.78
Porcentaje de casos mal clasificados	20.22

Regresión Logística

Función

$$\text{WINE2} = \begin{cases} 1 \text{ - Más que 1/4 litro por día} \\ 0 \text{ – Hasta 1/4 litro por día o no consume} \end{cases}$$

Si Más que 1/4 litro por día $\Rightarrow y = \beta_0 + (\beta_1 \times \text{AGE}) + (\beta_2 \times 1) + (\beta_3 \times \text{MINI MENTAL STATE EXAM}) =$

$$= (\beta_0 + \beta_2) + (\beta_1 \times \text{AGE}) + (\beta_3 \times \text{MINI MENTAL STATE EXAM})$$

Si Hasta 1/4 litro por día o no consume $\Rightarrow y = \beta_0 + (\beta_1 \times \text{AGE}) + (\beta_2 \times 0) + (\beta_3 \times \text{MINI MENTAL STATE EXAM}) =$

$$= \beta_0 + (\beta_1 \times \text{AGE}) + (\beta_3 \times \text{MINI MENTAL STATE EXAM})$$

Donde β_0 = constante = -3.502, $\beta_1 = 0.114$, $\beta_2 = 1.309$, $\beta_3 = -0.302$

Classification Table[a]

			Predicted		
			Incident Dementia		Percentage
	Observed		No	Yes	Correct
Step 1	Incident Dementia	No	184	16	92.0
		Yes	34	38	52.8
	Overall Percentage				81.6

a. The cut value is .500

Tabla de Clasificación en validación cruzada

Tabla de clasificaciones

	Predichos		
	0	1	total
clase menor original 0	183	17	200
clase mayor original 1	34	38	72

Porcentajes de Tabla de clasificaciones

	Predichos	
	0	1
clase menor original 0	91.50	8.50
clase mayor original 1	47.22	52.78

porcentajes de instancias bien clasificadas 81.25
porcentajes de instancias mal clasificadas 18.75

En las Tablas de Clasificaciones se puede ver que el porcentaje de los casos bien clasificados, de los correspondientes a la clase de la enfermedad, en el **Análisis Discriminante de Fisher es 73.61%**, mientras que en la **Regresión logística es 52.78%.**

6 Conclusiones y Trabajos Futuros

En este trabajo se presentó un test estadístico para ajustar el modelo en el método Análisis discriminante de Fisher. Este test es global, es decir que se tienen en cuenta el aporte de todas las variables que forman parte del Modelo Completo (M.C.) y el aporte de todas las variables que forman parte del Modelo Reducido (M.R.). La Regresión logística además de un test estadístico global para ajustar el modelo, cuenta con test estadístico individual, es decir para ver si la estimación del coeficiente de cada una de las variables es significativamente distinta de cero. En este trabajo no se ven test estadísticos individuales para ajustar el modelo en el Análisis Discriminante de Fisher.

La Regresión Logística permite el uso de variables continuas y categóricas transformadas a dummy, mientras que el Análisis Discriminante de Fisher originalmente utiliza solo variables continuas y el test estadístico presentado tiene el supuesto de distribución normal multivariada. Sin embargo para no perder la información representada por las variables categóricas también se aplicó el Análisis Discriminante de Fisher con transformaciones a dummy, logrando en cada conjunto de datos mayor cantidad de casos bien clasificados de los correspondientes a las clases de los enfermos o de los que murieron. Cabe aclarar que clasificar mal a un enfermo clasificándolo como sano puede traer graves consecuencias. En este trabajo no se ve si hay distorsiones en las varianzas o en las covarianzas de la matriz de varianzas-covarianzas provocadas por el uso de variables binarias.

En los ejemplos 1, 4 ,5 y 6 los modelos obtenidos por ambos métodos tienen las mismas variables. En el ejemplo 2, cuenta con 7 continuas y 6 binarias, el modelo obtenido por Análisis Discriminante de Fisher retiene las tres variables binarias que se obtienen con Regresión Logística y una binaria más. En el ejemplo 3, cuenta con 3 continuas y 18 binarias, el modelo obtenido por Análisis Discriminante de Fisher retiene las 2 variables continuas y las 4 variables binarias que se obtienen con Regresión Logística y 2 variables binarias más. En total solo difieren en 3 variables binarias. El criterio presentado se comporta en forma robusta en estos ejemplos.

En los Resultados se puede ver que en cada uno de los ejemplos vistos, en la Tabla de clasificación en Validación Cruzada, el porcentaje de casos bien clasificados de los correspondientes a la clase de los enfermos o de los que murieron, en el ajuste realizado con el Análisis Discriminante de Fisher es mayor que en el ajuste realizado con la Regresión Logística. En el ejemplo 1, 92.31% contra 53.85%, en el ejemplo 2, 69.49% contra 62.71%, en el ejemplo 3, 60.00% contra 35.00%, en el ejemplo 4, 55.93% contra 33.90%, en el ejemplo 5, 71.25% contra 53.75%, en el ejemplo 6, 73.61% contra 52.78%.

Hay conjuntos de datos donde hay casos bien clasificados por ambos métodos y casos mal clasificados por ambos métodos, estos casos tienen un comportamiento bien determinado, los casos bien clasificados tienen los valores de sus coeficientes cercanos a casos de su misma clase, los casos mal clasificados tienen los valores de sus coeficientes cercanos a casos de la otra clase. Pero hay casos que están bien clasificados por un método y mal clasificados por el otro, esto se debe a que esos casos que tienen los valores de sus coeficientes cercanos a casos pertenecientes a su misma clase y también cercanos a otros casos pertenecientes a la otra clase. Se podría usar otro método para estudiar el comportamiento de esos casos.

En este trabajo no se utilizaron conjuntos de entrenamiento y de test porque la cantidad de casos es pequeña y se pierden casos para ser clasificados y se presenta la dificultad de encontrar conjuntos bien distribuidos. Una propuesta para un trabajo a futuro, es dividir el conjunto de casos en tres conjuntos disjuntos y con 1/3 de los casos cada uno. Si los llamamos c1, c2 y c3 entonces quedarían tres pares de conjuntos de entrenamiento y test, en un par el conjunto de entrenamiento está formado por los casos pertenecientes a c1 y c2 y el de test es c3, en otro par el conjunto de entrenamiento está formado por los casos pertenecientes a los conjuntos c1 y c3 y el de test es c2, en el tercer par el conjunto de entrenamiento está formado por los casos pertenecientes a c2 y c3 y el de test es c1. Se realiza el ajuste del modelo y el test con cada par de conjuntos y si quedan las mismas variables en cada modelo y los porcentajes de casos bien clasificados es parecido entonces estarían bien distribuidos los conjuntos.

En este trabajo no se ve el aporte da cada variable a la estimación de la variable de respuesta, para trabajo a futuro, se debe comparar la estimación de la variable de respuesta y de los coeficientes de la función en el Modelo Completo contra el Modelo Reducido de acuerdo a la distribución y los grados de libertad.

7 Anexos

7.1 Análisis de los conjuntos de datos

7.1.1 Ejemplo 1 "Diabetic Ketoacidosis" [VIL/95]

Estadísticas Descriptivas

Tablas de Frecuencias de las variables categóricas

PreviousDiagnoses

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	9	8.5	8.5	8.5
	1	97	91.5	91.5	100.0
	Total	106	100.0	100.0	

DeepComaDegree

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	20	18.9	18.9	18.9
	1	86	81.1	81.1	100.0
	Total	106	100.0	100.0	

La variable de respuesta es "subp" y la tabla de frecuencias es la siguiente:

SUBP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	93	87.7	87.7	87.7
	2	13	12.3	12.3	100.0
	Total	106	100.0	100.0	

Estadísticas Descriptivas e Histogramas de las variables continuas

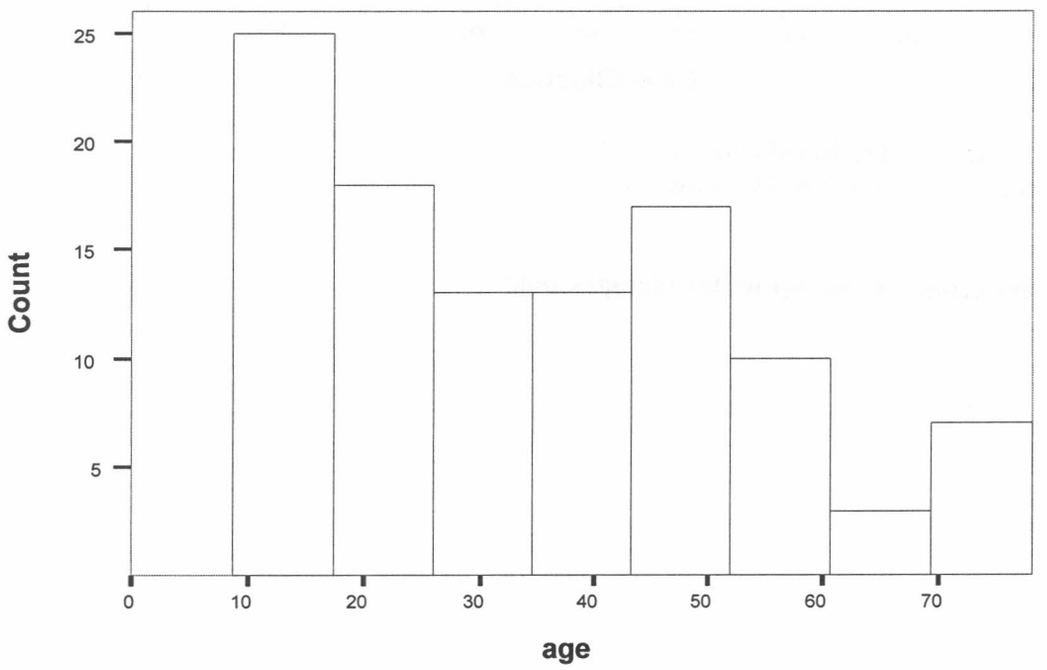
Estadísticas Descriptivas

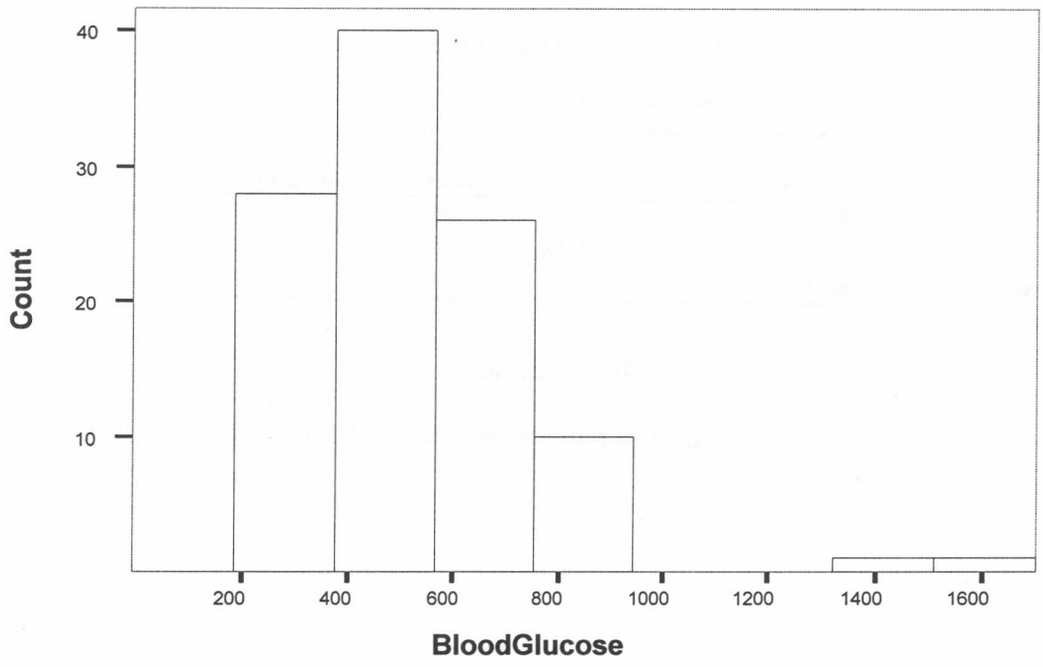
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	106	10	78	34.86	18.40
BloodGlucose	106	212	1700	524.86	221.36
Valid N (listwise)	106				

Histogramas

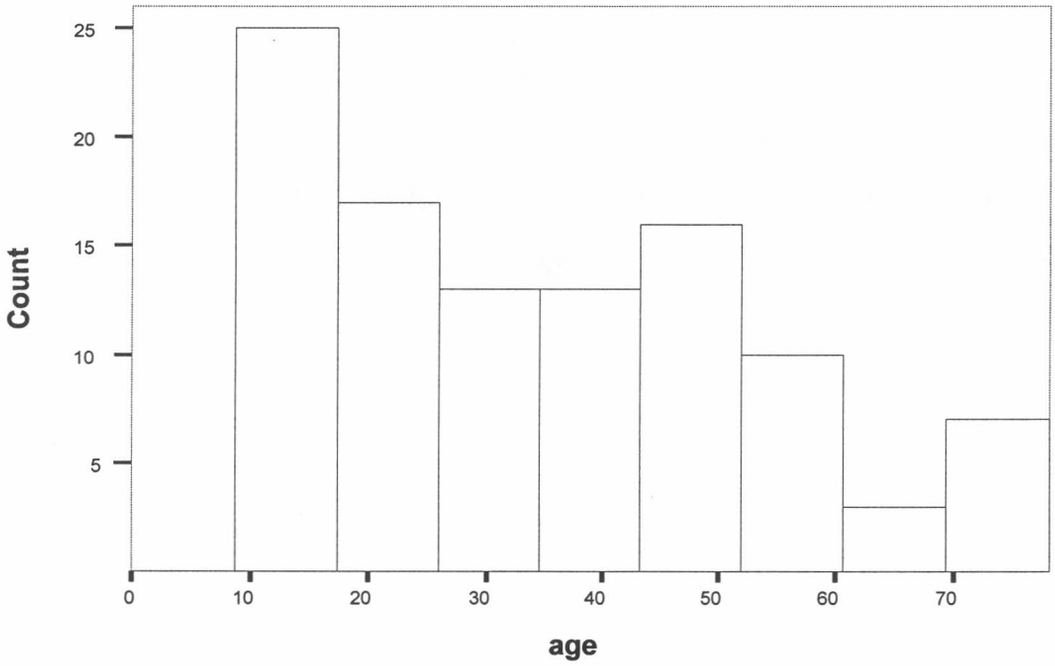
Para calcular el número de intervalos se utiliza la fórmula empírica 1+(3.3*(log(106))= 7.68 se utilizan 8 intervalos en los histogramas.

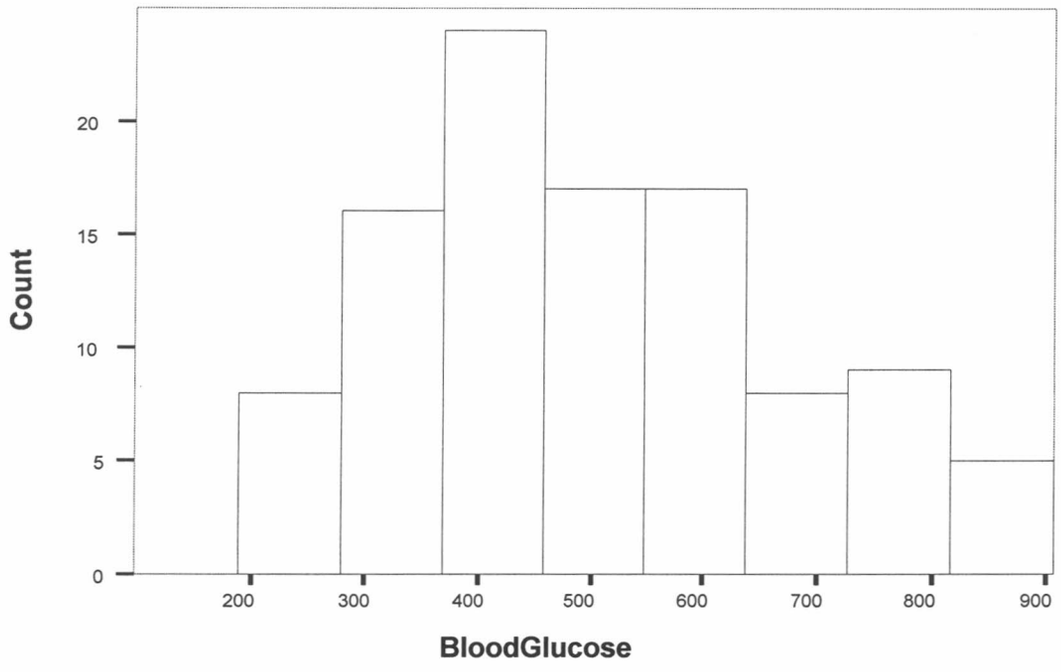




El caso 105 corresponde a Blood Glucose = 1404.
El caso 106 corresponde a Blood Glucose = 1700.

Sin los outliers tenemos los siguientes histogramas:





En la publicación realizaron el análisis con todos los datos, eso se observa de los resultados publicados, por lo tanto en el análisis tanto de Regresión Logística como de Análisis Discriminante de Fisher, se usan todos los datos.

Análisis Discriminante de Fisher

Vectores de Medias y Matriz de varianzas-covarianzas combinada

Vectores de medias

Nombres de variables	media de la clase 1	media de la clase 2
age	31.7634	57.0000
BloodGlucose	492.0215	759.7692
PreviousDiagnoses	0.9462	0.6923
DeepComaDegree	0.8602	0.4615

Matriz de varianzas-covarianzas combinada

	age	BloodGlucose	PreviousDiagnoses	DeepComaDegree
age	271.9114	525.2449	0.3059	-0.9430
BloodGlucose	525.2449	41607.2141	-4.3251	-25.8109
PreviousDiagnoses	0.3059	-4.3251	0.0721	0.0110
DeepComaDegree	-0.9430	-25.8109	0.0110	0.1386

Criterio para ajustar el modelo en cada paso

F de la tabla = 3.9306

Número de Paso	Variable	SCD						SCE		SCD	SCE	razón
	entra/sale	MC			MR			MC	MR	MC-MR	MC-MR	
			SCD1	SCD2		SCD1	SCD2					
1	sale: age NO	416	297.4013	118.5987	312	212.0233	99.9767	52.7785	31.7766	104	21.002	21.406
1	sale:BloodGlucose NO	416	297.4013	118.5987	312	237.1771	74.8229	52.7785	45.3265	104	7.452	7.5953
1	sale:PreviousDiagnoses NO	416	297.4013	118.5987	312	228.0397	83.9603	52.7785	43.3001	104	9.4784	9.6607
1	sale: DeepComaDegree SI	416	297.4013	118.5987	312	212.5517	99.4483	52.7785	50.6757	104	2.1028	2.1432
2	sale: age NO	312	212.5517	99.44831	208	127.2599	80.7401	50.6757	27.7871	104	22.889	23.329
2	sale:BloodGlucose NO	312	212.5517	99.44831	208	157.026	50.974	50.6757	39.3809	104	11.295	11.512
2	sale: PreviousDiagnoses NO	312	212.5517	99.44831	208	143.946	64.054	50.6757	40.1908	104	10.485	10.687

Resumen de clasificaciones

Paso número 0 (con todas las variables)

p = 0.95 F de la tabla = 3.9306

Variables dentro del modelo con las estimaciones de los coeficientes de la función:

Age	-0.0841
BloodGlucose	-0.0042
PreviousDiagnoses	3.4337
DeepComaDegree	1.2397

Tabla de Clasificaciones

	predicho 1	2	
original 1	77	16	93
original 2	1	12	13
	78	28	106

Cantidad de casos bien clasificados 89
Cantidad de casos mal clasificados 17

Porcentajes

	predicho 1	2
original 1	82.80	17.20
original 2	7.69	92.31

Porcentaje de casos bien clasificados 83.96
Porcentaje de casos mal clasificados 16.04

Paso número 1

F = 2.1432

Variables dentro del modelo con las estimaciones de los coeficientes de la función, la variable DeepComaDegree queda fuera del modelo y se muestra en blanco su posición:

Age	-0.0872
BloodGlucose	-0.0049
PreviousDiagnoses	3.5937

Tabla de Clasificaciones

	predicho 1	2	
original 1	78	15	93
original 2	1	12	13
	79	27	106

Cantidad de casos bien clasificados 90
Cantidad de casos mal clasificados 16

Porcentajes

	predicho 1	2
original 1	83.87	16.13
original 2	7.69	92.31

Porcentaje de casos bien clasificados 84.91
Porcentaje de casos mal clasificados 15.09

Regresión Logística del juego de datos "Diabetic Ketoacidosis"

Prueba con todas las variables dentro del modelo

-2 Log likelihood = Deviance = 40.300

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.311	8	.138

Dado que sig = 0.138 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			SUBP		Percentage
	Observed		1	2	Correct
Step 1	SUBP	1	90	3	96.8
		2	5	8	61.5
	Overall Percentage				92.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.108	.032	11.244	1	.001	1.114
	BLOODGLU	.004	.002	4.975	1	.026	1.004
	PREVIOUS(1)	2.093	1.076	3.785	1	.052	8.109
	DEEPCOMA(1)	.574	.876	.429	1	.512	1.775
	Constant	-10.452	2.658	15.458	1	.000	.000

a. Variable(s) entered on step 1: AGE, BLOODGLU, PREVIOUS, DEEPCOMA.

Dado que la variable "DEEPCOMA" tiene nivel de significación sig = 0.512 > 0.05, la estimación del coeficiente no es significativa.

Tenemos entonces los resultados de la prueba dejando fuera del modelo la variable "DeepComaDegree"

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.645	8	.091

Dado que sig = 0.091 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			SUBP		Percentage
	Observed		1	2	Correct
Step 1	SUBP	1	90	3	96.8
		2	6	7	53.8
	Overall Percentage				91.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.109	.032	11.612	1	.001	1.116
	BLOODGLU	.005	.002	6.585	1	.010	1.005
	PREVIOUS(1)	2.238	1.053	4.520	1	.033	9.375
	Constant	-10.628	2.649	16.091	1	.000	.000

a. Variable(s) entered on step 1: AGE, BLOODGLU, PREVIOUS.

Para todas las variables la estimación de los coeficientes es significativa pues el nivel de significación sig < 0.05.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	101	40.300
Reducido	3	102	40.725

$\chi^2_{1,\alpha}$	α
0.425	0.51455

0.51455 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables que son individualmente significativas

Modelo	Cantidad de variables	G. L.	Deviance
Completo	3	102	40.725
Reducido	2	103	Ver cuadro siguiente

Nombre de Variable	Deviance	$\chi^2_{1,\alpha}$	α
AGE	61.282	20.557	0.00001
BLOODGLU	50.557	9.852	0.00170
PREVIOUS	45.412	4.687	0.03039

En cada prueba $\alpha < 0.05$, indica que la variable que no fue incluida aporta significativamente al ajuste del modelo una vez que las otras variables habían sido incluidas en el mismo.

Decisión: las variables AGE, BLOODGLU y PREVIOUS forman parte del modelo.

7.1.2 Ejemplo 2 "psychosocial influences in breast cancer" [KRZ/80], [NUÑ/03]

Estadísticas Descriptivas

Tablas de Frecuencias de las variables categóricas

THYROID

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	absence	7	5.1	5.1	5.1
	presence	130	94.9	94.9	100.0
	Total	137	100.0	100.0	

ALLERGY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	absence	30	21.9	21.9	21.9
	presence	107	78.1	78.1	100.0
	Total	137	100.0	100.0	

La variable TEMPER fue transformada en dos variables dummy "TEMPER1" y "TEMPER2".

TEMPER

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	41	29.9	29.9	29.9
	1	75	54.7	54.7	84.7
	2	21	15.3	15.3	100.0
	Total	137	100.0	100.0	

La variable FEELINGS fue transformada en dos variables dummy "FEELINGS1" y "FEELINGS2".

FEELINGS

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	20	14.6	14.6	14.6
	1	62	45.3	45.3	59.9
	2	55	40.1	40.1	100.0
	Total	137	100.0	100.0	

Tabla de Frecuencia de la variable de respuesta "TUMOR"

TUMOR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	benign	78	56.9	56.9	56.9
	malignant	59	43.1	43.1	100.0
	Total	137	100.0	100.0	

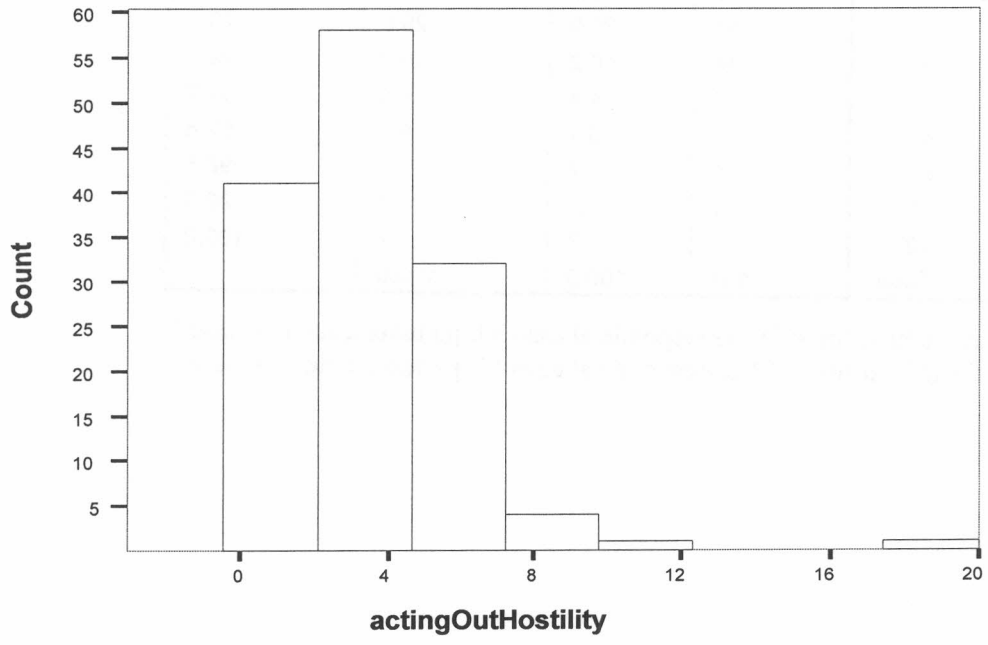
Estadísticas Descriptivas e Histogramas de las variables continuas

Descriptive Statistics

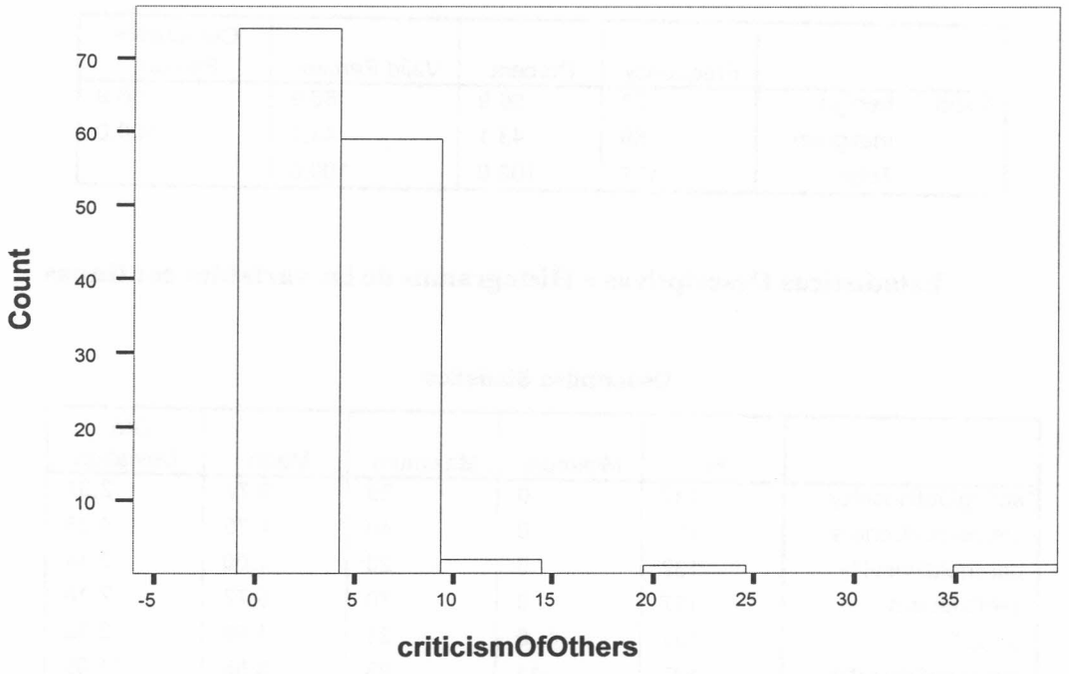
	N	Minimum	Maximum	Mean	Std. Deviation
actingOutHostility	137	0	20	3.77	2.31
criticismOfOthers	137	0	40	4.75	4.23
paranoidHostility	137	0	20	1.09	2.14
selfCriticism	137	0	70	5.73	7.19
GUILT	137	0	21	1.69	2.32
directionOfHostility	137	-11	93	3.55	11.05
ageOfMenarche	137	0	31	13.46	2.50
Valid N (listwise)	137				

Para calcular el número de intervalos se utiliza la fórmula empírica 1+(3.3*log(137))= 8.05 se utilizan 8 intervalos.

Todas las variables son enteras



Acting out hostility =20 corresponde al caso 33 perteneciente a la clase 1.



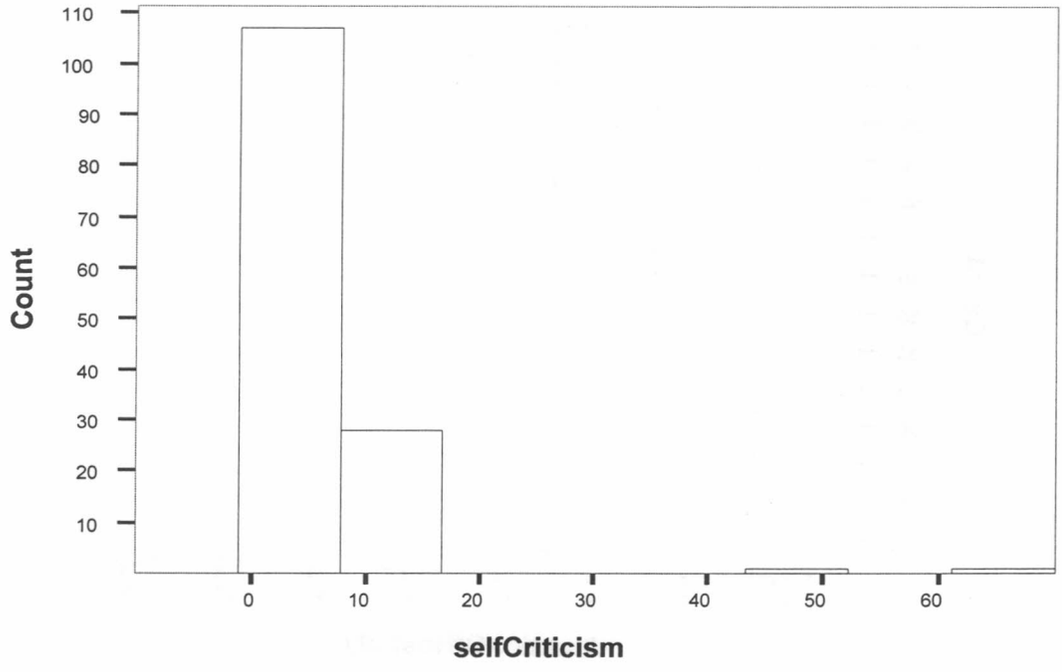
Criticism of others = 40 corresponde al caso 11, perteneciente a la clase 1.

La siguiente variables es preferible verla en una tabla de distribución de frecuencias porque tiene pocos valores distintos:

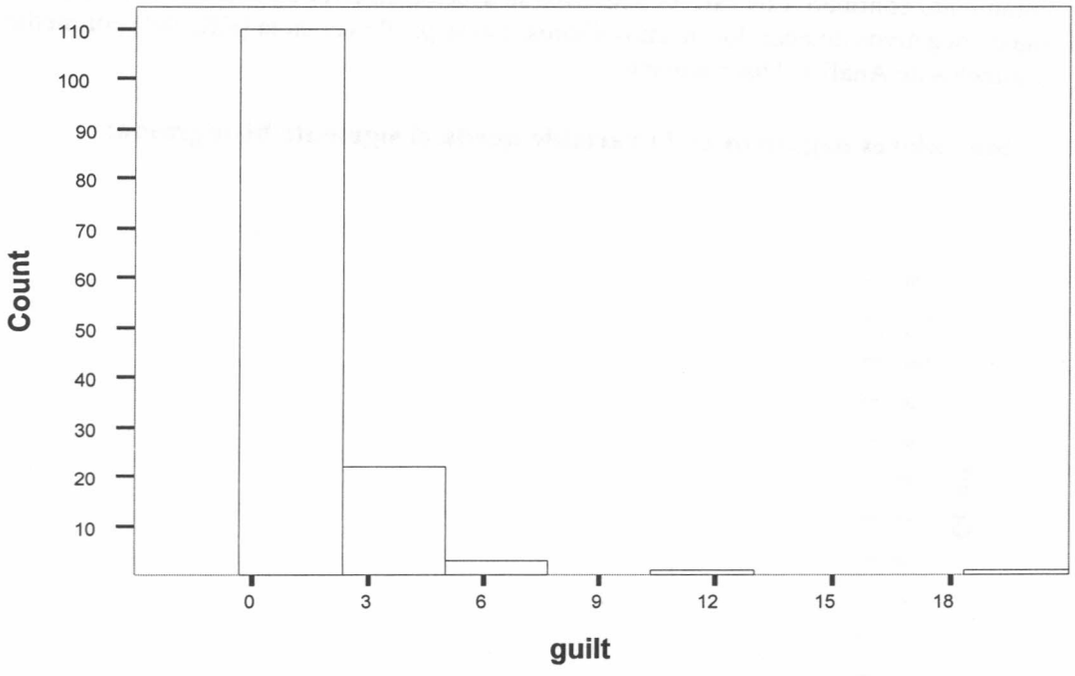
paranoidHostility

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	66	48.2	48.2	48.2
	1	41	29.9	29.9	78.1
	2	14	10.2	10.2	88.3
	3	6	4.4	4.4	92.7
	4	7	5.1	5.1	97.8
	5	1	.7	.7	98.5
	10	1	.7	.7	99.3
	20	1	.7	.7	100.0
	Total	137	100.0	100.0	

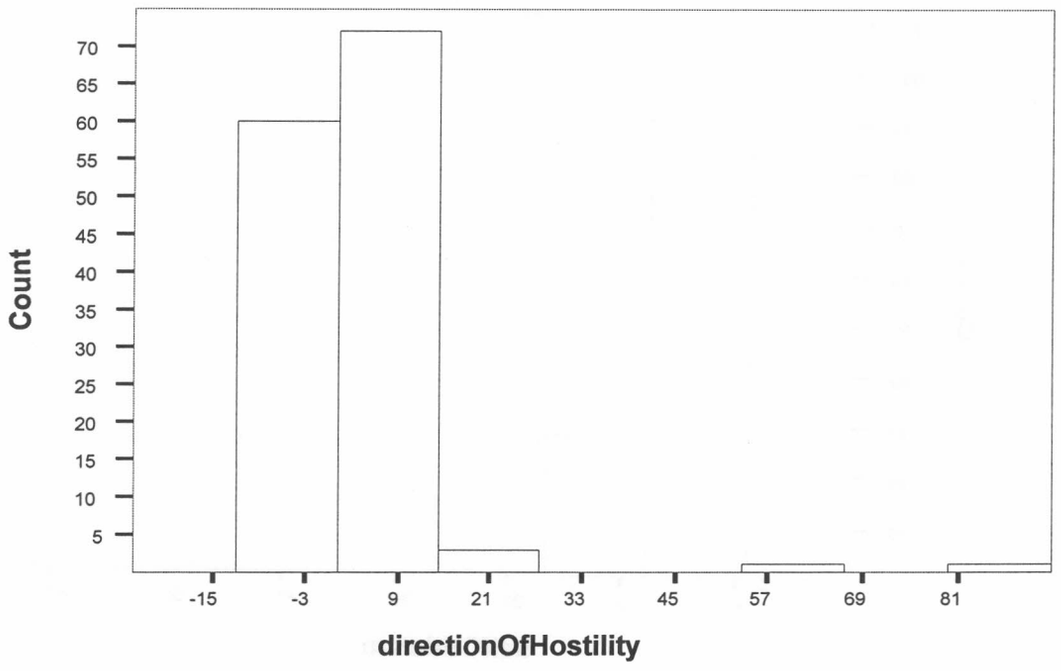
Paranoid hostility = 20 corresponde al caso 33, perteneciente a la clase 1.
Paranoid hostility = 10 corresponde al caso 11, perteneciente a la clase 1.



self Criticism = 70 corresponde la caso 11, perteneciente a la clase 1.
self Criticism = 50 corresponde la caso 33, perteneciente a la clase 1.



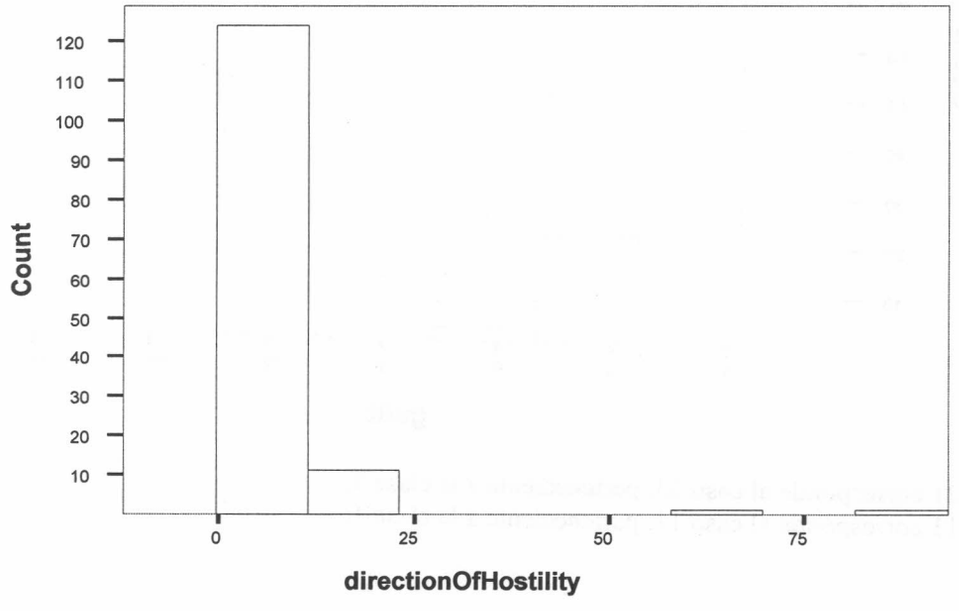
guilt = 21 corresponde al caso 33, perteneciente a la clase 1.
guilt = 11 corresponde al caso 11, perteneciente a la clase 1.



direction of hostility = 93 corresponde con el caso 11, perteneciente a la clase 1.
direction of hostility = 64 corresponde con el caso 33, perteneciente a la clase 1.

La variable "direction Of Hostility" está definida en la publicación en el rango [0, 10] y con tratamiento continuo. Los valores negativos se deben a un error de edición, se le deben sacar los signos negativos dejando los mismos dígitos, eso se puede ver en la [KRZ/80] por medio de los resultados de Análisis Discriminante.

Sin valores negativos en la variable queda el siguiente histograma:



La siguiente variable es preferible verla en una tabla de frecuencias porque tiene pocos valores distintos:

ageOfMenarche

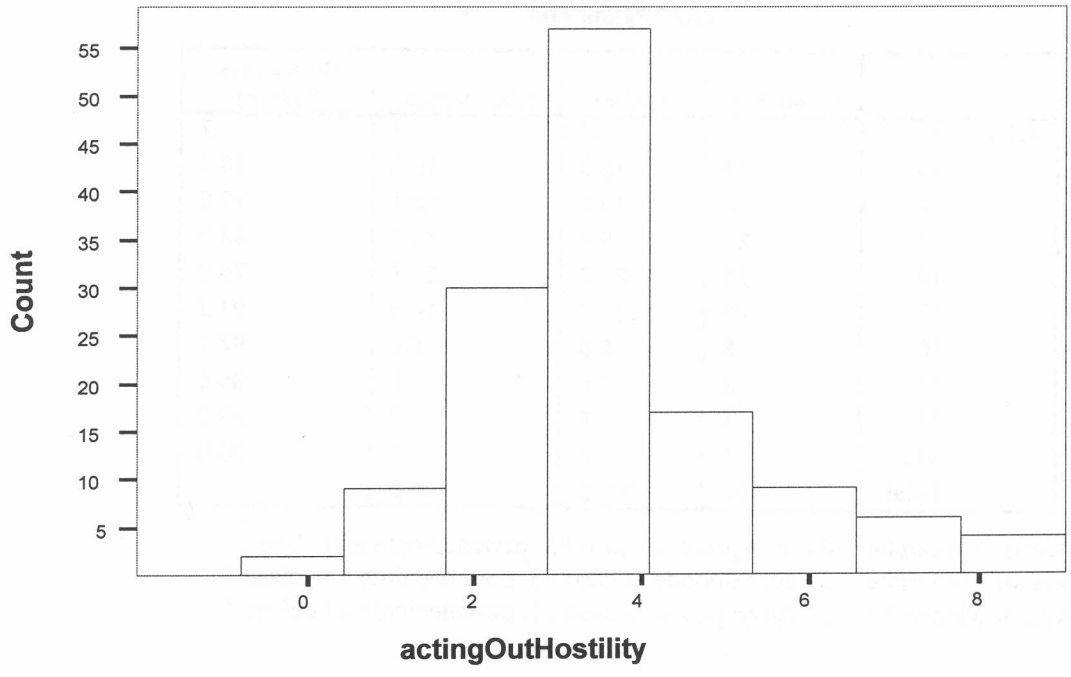
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	.7	.7	.7
	11	21	15.3	15.3	16.1
	12	18	13.1	13.1	29.2
	13	27	19.7	19.7	48.9
	14	38	27.7	27.7	76.6
	15	20	14.6	14.6	91.2
	16	8	5.8	5.8	97.1
	17	2	1.5	1.5	98.5
	21	1	.7	.7	99.3
	31	1	.7	.7	100.0
	Total	137	100.0	100.0	

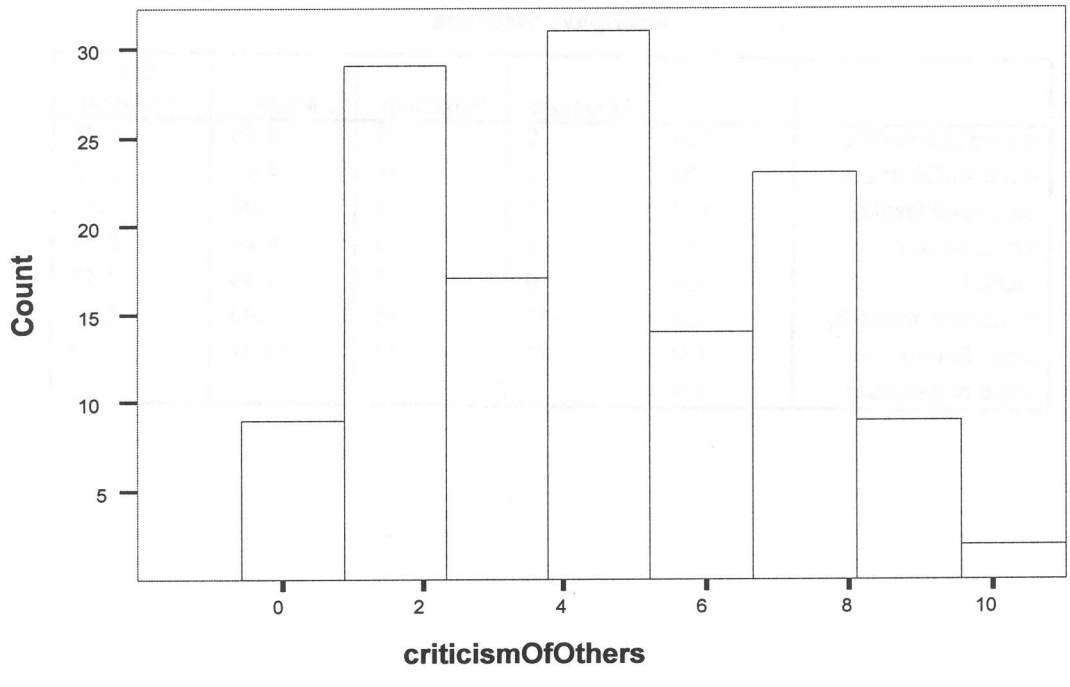
Age of Menarche = 0 corresponde al caso 83, perteneciente a la clase 0.
Age of Menarche = 31 corresponde al caso 33, perteneciente a la clase 1.
Age of Menarche = 21 corresponde al caso 11, perteneciente a la clase 1.

Sin los casos extremos tenemos los siguientes histogramas:

Descriptive Statistics

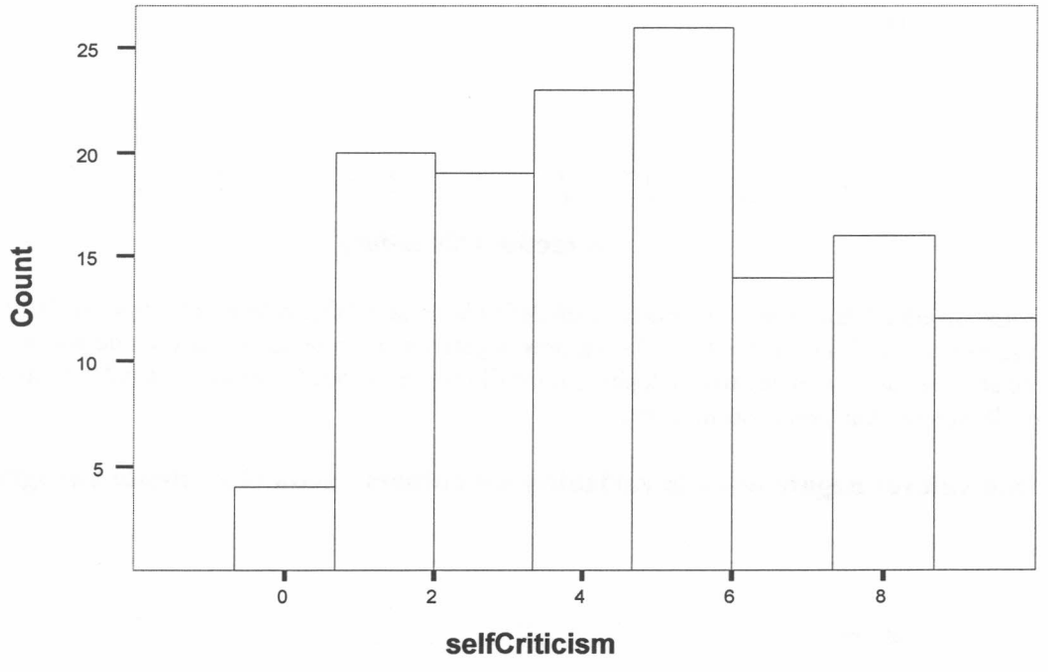
	N	Minimum	Maximum	Mean	Std. Deviation
actingOutHostility	134	0	9	3.60	1.78
criticismOfOthers	134	0	11	4.41	2.63
paranoidHostility	134	0	5	.90	1.17
selfCriticism	134	0	10	4.94	2.53
GUILT	134	0	7	1.48	1.42
directionOfHostility	134	-11	16	2.43	5.99
ageOfMenarche	134	11	17	13.37	1.52
Valid N (listwise)	134				


Count
55
50
45
40
35
30
25
20
15
10
5
0
2
4
6
8
actingOutHostility


Count
30
25
20
15
10
5
0
2
4
6
8
10
criticismOfOthers

paranoidHostility

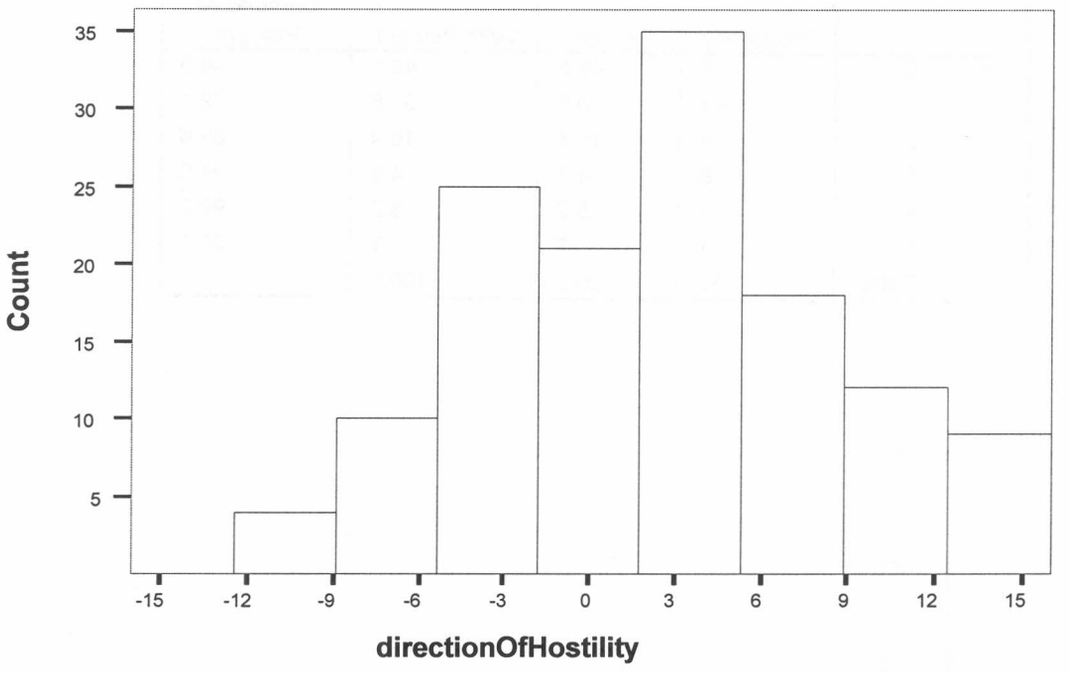
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	65	48.5	48.5	48.5
	1	41	30.6	30.6	79.1
	2	14	10.4	10.4	89.6
	3	6	4.5	4.5	94.0
	4	7	5.2	5.2	99.3
	5	1	.7	.7	100.0
	Total	134	100.0	100.0	



La siguiente variable es preferible verla en una tabla de distribución de frecuencias porque tiene pocos valores distintos:

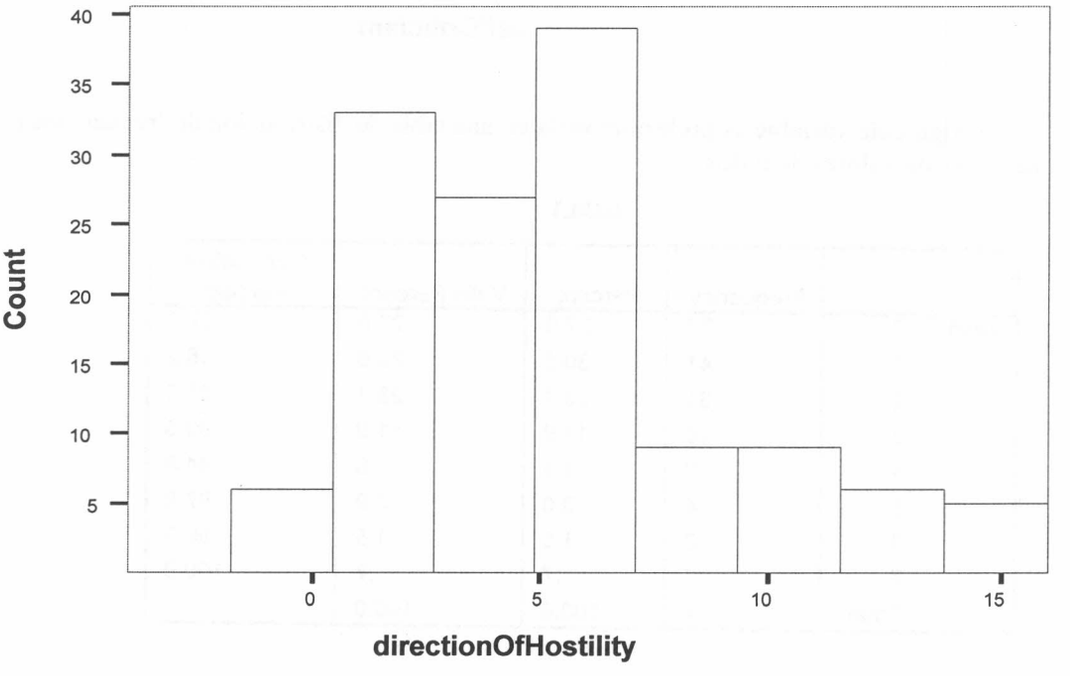
GUILT

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	37	27.6	27.6	27.6
	1	41	30.6	30.6	58.2
	2	31	23.1	23.1	81.3
	3	16	11.9	11.9	93.3
	4	2	1.5	1.5	94.8
	5	4	3.0	3.0	97.8
	6	2	1.5	1.5	99.3
	7	1	.7	.7	100.0
	Total	134	100.0	100.0	



La variable "direction of hostility" está definida en la publicación en el rango [0, 10] y con tratamiento continuo, por lo tanto los valores negativos corresponden a errores de edición y hay que sacarle los signos negativos dejándole los dígitos. Esto puede verse en [KRZ/80] en los resultados de Análisis Discriminante.

Sin valores negativos en la variable y sin outliers queda el siguiente histograma:



La variable "Age of Menarche" tienre pocos valores distintos por lo tanto es preferible verla en una tabla de distribución de frecuencias:

ageOfMenarche

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	11	21	15.7	15.7	15.7
	12	18	13.4	13.4	29.1
	13	27	20.1	20.1	49.3
	14	38	28.4	28.4	77.6
	15	20	14.9	14.9	92.5
	16	8	6.0	6.0	98.5
	17	2	1.5	1.5	100.0
	Total	134	100.0	100.0	

Análisis Discriminante de Fisher

Vectores de Medias y Matriz de Varianzas-Covarianzas Combinada

Vectores de Medias	Media de la clase 0	Media de la clase 1
actingOutHostility	3.7308	3.8305
criticismOfOthers	4.2436	5.4237
paranoidHostility	0.8590	1.4068
selfCriticism	4.9487	6.7627
guilt	1.5256	1.9153
directionOfHostility	5.0641	7.8475
ageOfMenarche	13.2949	13.6780
thyroid	0.9872	0.8983
allergy	0.7179	0.8644
temper1	0.7436	0.2881
temper2	0.1026	0.2203
feelings1	0.5769	0.2881
feelings2	0.3077	0.5254

Matriz de Varianzas-Covarianzas

5.3752	5.9359	3.3934	9.1522	3.7051	10.3246	2.3479	-0.0021	-0.0391	-0.0852	0.2545	-0.1111	0.0497
5.9359	17.6947	6.0482	23.6808	5.8158	27.9822	4.9292	-0.1794	-0.2537	-0.2321	0.3299	-0.1642	-0.0295
3.3934	6.0482	4.5458	11.0380	3.7320	13.3212	3.2368	-0.0570	-0.0655	-0.0943	0.1099	-0.0783	0.0057
9.1522	23.6808	11.0380	51.2776	12.2586	64.5268	9.3087	-0.3739	-0.4446	-0.3481	0.3222	-0.4790	0.1599
3.7051	5.8158	3.7320	12.2586	5.4076	15.5082	3.0022	-0.0221	-0.0157	-0.2152	0.0955	-0.1275	0.1186
10.3246	27.9822	13.3212	64.5268	15.5082	93.9726	11.9824	-0.5100	-0.6208	-0.3268	0.2998	-0.4318	0.0977
2.3479	4.9292	3.2368	9.3087	3.0022	11.9824	6.2748	-0.0047	0.0586	-0.0861	0.0358	-0.0355	-0.0007
-0.0021	-0.1794	-0.0570	-0.3739	-0.0221	-0.5100	-0.0047	0.0472	0.0141	0.0035	0.0031	0.0171	-0.0188
-0.0391	-0.2537	-0.0655	-0.4446	-0.0157	-0.6208	0.0586	0.0141	0.1682	0.0271	-0.0221	0.0296	-0.0298
-0.0852	-0.2321	-0.0943	-0.3481	-0.2152	-0.3268	-0.0861	0.0035	0.0271	0.1998	-0.0718	0.0566	-0.0428
0.2545	0.3299	0.1099	0.3222	0.0955	0.2998	0.0358	0.0031	-0.0221	-0.0718	0.1283	-0.0027	-0.0392
-0.1111	-0.1642	-0.0783	-0.4790	-0.1275	-0.4318	-0.0355	0.0171	0.0296	0.0566	-0.0027	0.2307	-0.1687
0.0497	-0.0295	0.0057	0.1599	0.1186	0.0977	-0.0007	-0.0188	-0.0298	-0.0428	-0.0392	-0.1687	0.2321

Criterio para ajustar el modelo en cada paso

F de la tabla = 3.9102, p = 0.95

Número de paso	Variable sale/entra	SCD						SCE		SCD	SCE	razón
		MC			MR			MC	MR	MC-MR	MC-MR	
			SCD0	SCD1		SCD0	SCD1					
1	sale: actingOutHostility SI	1755	839.6909	915.3091	1620	762.7672	857.2328	63.3527	63.1599	135	0.1928	0.1956
2	sale: criticismOfOthers SI	1620	762.7672	857.2328	1485	702.8274	782.1726	63.1599	62.5623	135	0.5976	0.6065
3	entra: actingOutHostility NO	1620	780.2049	839.7951	1485	702.8274	782.1726	62.6082	62.5623	135	0.0459	0.0466
3	sale: paranoidHostility SI	1485	702.8274	782.1726	1350	634.5744	715.4256	62.5623	59.6007	135	2.9616	3.0055
4	entra: actingOutHostility NO	1485	709.068	775.932	1350	634.5744	715.4256	59.7121	59.6007	135	0.1115	0.1131
4	entra: criticismOfOthers NO	1485	698.708	786.292	1350	634.5744	715.4256	60.7933	59.6007	135	1.1927	1.2103
4	sale: selfCriticism SI	1350	634.5744	715.4256	1215	567.9248	647.0752	59.6007	59.3606	135	0.2401	0.2436
5	entra: actingOutHostility NO	1350	642.9542	707.0458	1215	567.9248	647.0752	59.4154	59.3606	135	0.0548	0.0557
5	entra: criticismOfOthers NO	1350	626.7373	723.2627	1215	567.9248	647.0752	59.8616	59.3606	135	0.501	0.5084
5	entra: paranoidHostility NO	1350	642.3244	707.6756	1215	567.9248	647.0752	61.6929	59.3606	135	2.3323	2.3668
5	sale: guilt SI	1215	567.9248	647.0752	1080	494.3933	585.6067	59.3606	56.1485	135	3.2121	3.2597
6	entra: actingOutHostility NO	1215	554.8844	660.1156	1080	494.3933	585.6067	56.8495	56.1485	135	0.701	0.7114
6	entra: criticismOfOthers NO	1215	557.0872	657.9128	1080	494.3933	585.6067	56.2514	56.1485	135	0.1029	0.1044
6	entra: paranoidHostility NO	1215	556.035	658.965	1080	494.3933	585.6067	56.2651	56.1485	135	0.1166	0.1183
6	entra: selfCriticism NO	1215	559.6286	655.3714	1080	494.3933	585.6067	57.2959	56.1485	135	1.1474	1.1644
6	sale: directionOfHostility SI	1080	494.3933	585.6067	945	447.1007	497.8993	56.1485	54.7984	135	1.3501	1.3701
7	entra: actingOutHostility NO	1080	509.8384	570.1616	945	447.1007	497.8993	54.9516	54.7984	135	0.1532	0.1554
7	entra: criticismOfOthers NO	1080	498.0705	581.9295	945	447.1007	497.8993	55.6249	54.7984	135	0.8265	0.8387
7	entra: paranoidHostility NO	1080	509.2279	570.7721	945	447.1007	497.8993	55.5115	54.7984	135	0.7131	0.7236
7	entra: selfCriticism NO	1080	492.8203	587.1797	945	447.1007	497.8993	55.1179	54.7984	135	0.3195	0.3242
7	entra: guilt NO	1080	514.5081	565.4919	945	447.1007	497.8993	55.2977	54.7984	135	0.4993	0.5067
7	sale: ageOfMenarche SI	945	447.1007	497.8993	810	395.6028	414.3972	54.7984	54.7747	135	0.0237	0.024
8	entra: actingOutHostility NO	945	434.0352	510.9648	810	395.6028	414.3972	54.8582	54.7747	135	0.0834	0.0847

8	entra: criticismOfOthers NO	945	428.2456	516.7544	810	395.6028	414.3972	55.5292	54.7747	135	0.7545	0.7656
8	entra: paranoidHostility NO	945	417.9152	527.0848	810	395.6028	414.3972	55.3536	54.7747	135	0.5789	0.5874
8	entra: selfCriticism NO	945	409.9174	535.0826	810	395.6028	414.3972	55.0865	54.7747	135	0.3118	0.3164
8	entra: guilt NO	945	429.8424	515.1576	810	395.6028	414.3972	55.0524	54.7747	135	0.2777	0.2818
8	entra: directionOfHostility NO	945	411.529	533.471	810	395.6028	414.3972	55.9292	54.7747	135	1.1545	1.1716
8	sale: thyroid NO	810	395.6028	414.3972	675	373.293	301.707	54.7747	49.4049	135	5.3698	5.4494
8	sale: allergy NO	810	395.6028	414.3972	675	303.756	371.244	54.7747	42.7178	135	12.057	12.236
8	sale: temper1 NO	810	395.6028	414.3972	675	321.8742	353.1258	54.7747	30.8084	135	23.966	24.321
8	sale: temper2 SI	810	395.6028	414.3972	675	334.7579	340.2421	54.7747	54.7686	135	0.0062	0.0063
9	entra: actingOutHostility NO	810	374.9174	435.0826	675	334.7579	340.2421	54.8578	54.7686	135	0.0892	0.0905
9	entra: criticismOfOthers NO	810	366.3982	443.6018	675	334.7579	340.2421	55.4844	54.7686	135	0.7158	0.7264
9	entra: paranoidHostility NO	810	356.642	453.358	675	334.7579	340.2421	55.3311	54.7686	135	0.5625	0.5709
9	entra: selfCriticism NO	810	348.6978	461.3022	675	334.7579	340.2421	55.0733	54.7686	135	0.3047	0.3092
9	entra: guilt NO	810	369.2008	440.7992	675	334.7579	340.2421	55.0499	54.7686	135	0.2814	0.2855
9	entra: directionOfHostility NO	810	350.763	459.237	675	334.7579	340.2421	55.9165	54.7686	135	1.1479	1.1649
9	entra: ageOfMenarche NO	810	386.3028	423.6972	675	334.7579	340.2421	54.7921	54.7686	135	0.0236	0.0239
9	sale: thyroid NO	675	334.7579	340.2421	540	312.4332	227.5668	54.7686	49.3824	135	5.3861	5.4659
9	sale: allergy NO	675	334.7579	340.2421	540	241.523	298.477	54.7686	42.3522	135	12.416	12.6
9	sale: temper1 NO	675	334.7579	340.2421	540	256.5526	283.4474	54.7686	23.8957	135	30.873	31.33
9	sale: feelings1 SI	675	334.7579	340.2421	540	263.2319	276.7681	54.7686	52.8896	135	1.879	1.9068
10	entra: actingOutHostility NO	675	301.6135	373.3865	540	263.2319	276.7681	52.9217	52.8896	135	0.0321	0.0326
10	entra: criticismOfOthers NO	675	294.252	380.748	540	263.2319	276.7681	53.857	52.8896	135	0.9674	0.9817
10	entra: paranoidHostility NO	675	283.9567	391.0433	540	263.2319	276.7681	53.6318	52.8896	135	0.7422	0.7532
10	entra: selfCriticism NO	675	275.8743	399.1257	540	263.2319	276.7681	53.4137	52.8896	135	0.5241	0.5319
10	entra: guilt NO	675	297.4513	377.5487	540	263.2319	276.7681	53.1412	52.8896	135	0.2516	0.2554
10	entra: directionOfHostility NO	675	278.2662	396.7338	540	263.2319	276.7681	54.307	52.8896	135	1.4174	1.4384
10	entra: ageOfMenarche NO	675	315.3353	359.6647	540	263.2319	276.7681	52.9287	52.8896	135	0.0391	0.0397
10	sale: thyroid NO	540	263.2319	276.7681	405	241.1361	163.8639	52.8896	47.2068	135	5.6828	5.767
10	sale: allergy NO	540	263.2319	276.7681	405	170.0234	234.9766	52.8896	40.772	135	12.118	12.297
10	sale: temper1 NO	540	263.2319	276.7681	405	184.8981	220.1019	52.8896	18.3152	135	34.574	35.087
10	entra: temper2 NO	675	320.2447	354.7553	540	263.2319	276.7681	52.9355	52.8896	135	0.0458	0.0465

10	sale: feelings2 SI	540	263.2319	276.7681	405	191.1748	213.8252	52.8896	50.7639	135	2.1257	2.1572
11	entra: actingOutHostility NO	540	230.0998	309.9002	405	191.1748	213.8252	50.7837	50.7639	135	0.0198	0.0201
11	entra: criticismOfOthers NO	540	220.1534	319.8466	405	191.1748	213.8252	51.4878	50.7639	135	0.724	0.7347
11	entra: paranoidHostility NO	540	211.6368	328.3632	405	191.1748	213.8252	51.4052	50.7639	135	0.6414	0.6509
11	entra: selfCriticism NO	540	203.4038	336.5962	405	191.1748	213.8252	51.2298	50.7639	135	0.4659	0.4728
11	entra: guilt NO	540	226.0887	313.9113	405	191.1748	213.8252	50.9292	50.7639	135	0.1654	0.1678
11	entra: directionOfHostility NO	540	205.5452	334.4548	405	191.1748	213.8252	52.0094	50.7639	135	1.2455	1.264
11	entra: ageOfMenarche NO	540	243.3119	296.6881	405	191.1748	213.8252	50.7972	50.7639	135	0.0333	0.0338
11	sale: thyroid NO	405	191.1748	213.8252	270	169.8522	100.1478	50.7639	43.7352	135	7.0286	7.1328
11	sale: allergy NO	405	191.1748	213.8252	270	95.64467	174.3553	50.7639	39.5348	135	11.229	11.395
11	sale: temper1 NO	405	191.1748	213.8252	270	115.1247	154.8753	50.7639	11.7466	135	39.017	39.595
11	entra: temper2 NO	540	238.0792	301.9208	405	191.1748	213.8252	50.8809	50.7639	135	0.1171	0.1188
11	entra: feelings1 SI	540	275.1847	264.8153	405	191.1748	213.8252	54.764	50.7639	135	4.0001	4.0594

Resumen de Clasificaciones

En cada paso se muestran las variables dentro del modelo con las estimaciones de los coeficientes de la función. Se usa P=0.95, y corresponde F de la tabla = 3.9102.

Paso número 0 (con todas las variables)

actingOutHostility	0.0556	guilt	0.2552	allergy	-1.7858	feelings2	-0.3222
criticismOfOthers	-0.0638	directionOfHostility	-0.0935	temper1	2.4176		
paranoidHostility	-0.2459	ageOfMenarche	0.0572	temper2	0.0378		
selfCriticism	0.1001	thyroid	1.2440	feelings1	0.6413		

Tabla de Clasificaciones

	predicho 0	1	
original 0	63	15	78
original 1	16	43	59
	79	58	137

Porcentajes

	predicho 0	1
original 0	80.77	19.23
original 1	27.12	72.88

Cantidad de casos bien clasificados 106
Cantidad de casos mal clasificados 31
Porcentaje de casos bien clasificados 77.37
Porcentaje de casos mal clasificados 22.63

Paso número 1

La variable actingOutHostility queda fuera del modelo y se muestra en blanco su posición:

		guilt	0.2816	allergy	-1.7798	feelings2	-0.2877
criticismOfOthers	-0.0550	directionOfHostility	-0.0960	temper1	2.4784		
paranoidHostility	-0.2277	ageOfMenarche	0.0555	temper2	0.1440		
selfCriticism	0.0992	thyroid	1.2739	feelings1	0.6431		

Tabla de Clasificaciones

	predicho 0	1	
original 0	62	16	78
original 1	15	44	59
	77	60	137

Porcentajes

	predicho 0	1
original 0	79.49	20.51
original 1	25.42	74.58

Cantidad de casos bien clasificados 106
Cantidad de casos mal clasificados 31
Porcentaje de casos bien clasificados 77.37
Porcentaje de casos mal clasificados 22.63

Paso número 2

La variable criticismOfOthers queda fuera del modelo y se muestra en blanco su posición:

		guilt	0.2877	allergy	-1.7635	feelings2	-0.2852
		directionOfHostility	-0.0892	temper1	2.4746		
paranoidHostility	-0.2499	ageOfMenarche	0.0514	temper2	0.0760		
selfCriticism	0.0699	thyroid	1.3071	feelings1	0.6269		

Tabla de Clasificaciones

	predicho 0	1	
original 0	62	16	78
original 1	15	44	59
	77	60	137

Porcentajes

	predicho 0	1
original 0	79.49	20.51
original 1	25.42	74.58

Cantidad de casos bien clasificados 106
Cantidad de casos mal clasificados 31
Porcentaje de casos bien clasificados 77.37
Porcentaje de casos mal clasificados 22.63

Paso número 3

La variable paranoidHostility queda fuera del modelo y se muestra en blanco su posición:

		guilt	0.1813	allergy	-1.6961	feelings2	-0.2078
		directionOfHostility	-0.0752	temper1	2.3830		
		ageOfMenarche	-0.0032	temper2	-0.0084		
selfCriticism	0.0354	thyroid	1.4433	feelings1	0.6583		

Tabla de Clasificaciones

	predicho 0	1	
original 0	62	16	78
original 1	17	42	59
	79	58	137

Cantidad de casos bien clasificados 104
Cantidad de casos mal clasificados 33

Porcentajes

	predicho 0	1
original 0	79.49	20.51
original 1	28.81	71.19

Porcentaje de casos bien clasificados 75.91
Porcentaje de casos mal clasificados 24.09

Paso número 4

La variable selfCriticism queda fuera del modelo y se muestra en blanco su posición:

		guilt	0.2013	allergy	-1.7078	feelings2	-0.2274
		directionOfHostility	-0.0554	temper1	2.3952		
		ageOfMenarche	0.0017	temper2	0.0171		
		thyroid	1.3935	feelings1	0.6219		

Tabla de Clasificaciones

	predicho 0	1	
original 0	62	16	78
original 1	19	40	59
	81	56	137

Cantidad de casos bien clasificados 102
Cantidad de casos mal clasificados 35

Porcentajes

	predicho 0	1
original 0	79.49	20.51
original 1	32.20	67.80

Porcentaje de casos bien clasificados 74.45
Porcentaje de casos mal clasificados 25.55

Paso número 5

La variable guilt queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.6036	feelings2	-0.0699
		directionOfHostility	-0.0252	temper1	2.2588		
		ageOfMenarche	0.0378	temper2	0.0715		
		thyroid	1.6418	feelings1	0.6902		

Tabla de Clasificaciones

	predicho 0	1	
original 0	62	16	78
original 1	19	40	59
	81	56	137

Cantidad de casos bien clasificados 102
Cantidad de casos mal clasificados 35

Porcentajes

	predicho 0	1
original 0	79.49	20.51
original 1	32.20	67.80

Porcentaje de casos bien clasificados 74.45
Porcentaje de casos mal clasificados 25.55

Paso número 6

La variable directionOfHostility queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.5157	feelings2	-0.0194
				temper1	2.2512		
		ageOfMenarche	-0.0106	temper2	0.0473		
		thyroid	1.8881	feelings1	0.7390		

Tabla de Clasificaciones

	predicho 0	1	
original 0	60	18	78
original 1	18	41	59
	78	59	137

Cantidad de casos bien clasificados 101
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	1
original 0	76.92	23.08
original 1	30.51	69.49

Porcentaje de casos bien clasificados 73.72
Porcentaje de casos mal clasificados 26.28

Paso número 7

La variable ageOfMenarche queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.5205	feelings2	-0.0171
				temper1	2.2561		
				temper2	0.0470		
		thyroid	1.8902	feelings1	0.7416		

Tabla de Clasificaciones

	predicho 0	1	
original 0	60	18	78
original 1	18	41	59
	78	59	137

Cantidad de casos bien clasificados 101
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	1
original 0	76.92	23.08
original 1	30.51	69.49

Porcentaje de casos bien clasificados 73.72
Porcentaje de casos mal clasificados 26.28

Paso número 8

La variable temper2 queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.5256	feelings2	-0.0354
				temper1	2.2385		
		thyroid	1.8922	feelings1	0.7325		

Tabla de Clasificaciones

	predicho 0	1	
original 0	60	18	78
original 1	18	41	59
	78	59	137

Cantidad de casos bien clasificados 101
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	1
original 0	76.92	23.08
original 1	30.51	69.49

Porcentaje de casos bien clasificados 73.72
Porcentaje de casos mal clasificados 26.28

Paso número 9

La variable feelings1 queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.5064	feelings2	-0.5440
				temper1	2.3335		
		thyroid	1.9417				

Tabla de Clasificaciones

	predicho 0	1	
original 0	61	17	78
original 1	19	40	59
	80	57	137

Cantidad de casos bien clasificados 101
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	1
original 0	78.21	21.79
original 1	32.20	67.80

Porcentaje de casos bien clasificados 73.72
Porcentaje de casos mal clasificados 26.28

Paso número 10

La variable feelings2 queda fuera del modelo y se muestra en blanco su posición:

				allergy	-1.4427		
				Temper1	2.4381		
		thyroid	2.1316				

Tabla de Clasificaciones

	predicho 0	1	
original 0	65	13	78
original 1	22	37	59
	87	50	137

Cantidad de casos bien clasificados 102
Cantidad de casos mal clasificados 35

Porcentajes

	predicho 0	1
original 0	83.33	16.67
original 1	37.29	62.71

Porcentaje de casos bien clasificados 74.45
Porcentaje de casos mal clasificados 25.55

Paso número 11

La variable feelings1 pasa a formar parte del modelo y su posición que se encontraba en blanco, pasa a contener el coeficiente correspondiente a la función:

				allergy	-1.5242		
				temper1	2.2386		
		thyroid	1.8967	feelings1	0.7579		

Tabla de Clasificaciones

	predicho 0	1	
original 0	60	18	78
original 1	18	41	59
	78	59	137

Cantidad de casos bien clasificados 101
Cantidad de casos mal clasificados 36

Porcentajes

	predicho 0	1
original 0	76.92	23.08
original 1	30.51	69.49

Porcentaje de casos bien clasificados 73.72
Porcentaje de casos mal clasificados 26.28

Regresión Logística del juego de datos "psychosocial influences in breast cancer"

Prueba con todas las variables dentro del modelo

-2 Log likelihood = Deviance = 135.788

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.429	8	.816

Dado que sig = 0.816 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	64	14	82.1
		malignant	18	41	69.5
	Overall Percentage				76.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ACTINGOU	-.091	.158	.331	1	.565	.913
	CRITICIS	.112	.107	1.085	1	.298	1.118
	PARANOID	.161	.201	.644	1	.422	1.175
	SELFCRIT	-.073	.103	.506	1	.477	.930
	GUILT	-.308	.204	2.276	1	.131	.735
	DIRECTIO	.122	.066	3.439	1	.064	1.129
	AGEOFMEN	-.061	.114	.287	1	.592	.941
	THYROID(1)	1.821	1.371	1.763	1	.184	6.180
	ALLERGY(1)	-1.914	.611	9.805	1	.002	.148
	TEMPER1(1)	2.053	.554	13.724	1	.000	7.795
	TEMPER2(1)	-.258	.773	.112	1	.738	.772
	FEELING1(1)	.663	.647	1.049	1	.306	1.940
	FEELING2(1)	-.211	.697	.091	1	.762	.810
	Constant	-.346	2.001	.030	1	.863	.707

a. Variable(s) entered on step 1: ACTINGOU, CRITICIS, PARANOID, SELFCRIT, GUILT, DIRECTIO, AGEOFMEN, THYROID, ALLERGY, TEMPER1, TEMPER2, FEELING1, FEELING2.

La variable "TEMPER2" tiene nivel de significación: sig = 0.738 > 0.05, la estimación del coeficiente no es significativa.

Prueba dejando la variable "TEMPER2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.170	8	.518

Dado que sig = 0.518 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage Correct
	Observed		benign	malignant	
Step 1	TUMOR	benign	65	13	83.3
		malignant	18	41	69.5
	Overall Percentage				77.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ACTINGOU	-.070	.145	.232	1	.630	.932
	CRITICIS	.108	.107	1.027	1	.311	1.114
	PARANOID	.163	.201	.662	1	.416	1.177
	SELFCRIT	-.074	.102	.523	1	.470	.929
	GUILT	-.312	.203	2.372	1	.124	.732
	DIRECTIO	.119	.065	3.358	1	.067	1.126
	AGEOFMEN	-.061	.113	.291	1	.590	.941
	THYROID(1)	1.810	1.379	1.721	1	.190	6.108
	ALLERGY(1)	-1.881	.602	9.769	1	.002	.152
	TEMPER1(1)	2.145	.485	19.555	1	.000	8.541
	FEELING1(1)	.693	.639	1.176	1	.278	2.000
	FEELING2(1)	-.129	.651	.039	1	.843	.879
	Constant	-.707	1.684	.176	1	.675	.493

a. Variable(s) entered on step 1: ACTINGOU, CRITICIS, PARANOID, SELFCRIT, GUILT, DIRECTIO, AGEOFMEN, THYROID, ALLERGY, TEMPER1, FEELING1, FEELING2.

La variable "FELLING2" tiene nivel de significación: sig = 0.843 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de Variables	G. L.	Deviance
Completo	13	123	135.788
Reducido	12	124	135.900

$\chi^2_{1,\alpha}$	α
0.112	0.73788

0.73788 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba dejando la variable "FEELING2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.804	8	.778

Dado que sig = 0.778 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted		
		TUMOR		Percentage Correct
		benign	malignant	
Step 1 TUMOR	benign	65	13	83.3
	malignant	18	41	69.5
Overall Percentage				77.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ACTINGOU	-.072	.145	.242	1	.622	.931
	CRITICIS	.110	.106	1.081	1	.299	1.117
	PARANOID	.156	.197	.626	1	.429	1.169
	SELFCRIT	-.074	.102	.524	1	.469	.929
	GUILT	-.309	.202	2.326	1	.127	.734
	DIRECTIO	.121	.065	3.483	1	.062	1.128
	AGEOFMEN	-.061	.113	.292	1	.589	.941
	THYROID(1)	1.857	1.360	1.863	1	.172	6.405
	ALLERGY(1)	-1.873	.600	9.738	1	.002	.154
	TEMPER1(1)	2.145	.485	19.548	1	.000	8.541
	FEELING1(1)	.781	.459	2.901	1	.089	2.184
	Constant	-.846	1.527	.307	1	.580	.429

a. Variable(s) entered on step 1: ACTINGOU, CRITICIS, PARANOID, SELFCRIT, GUILT, DIRECTIO, AGEOFMEN, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "AGEOFMEN" (ageOfMenarche) tiene nivel de significación: sig = 0.589 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	12	124	135.900
Reducido	11	125	135.939

$\chi^2_{1,\alpha}$	α
0.039	0.84345

0.84345 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	12	125	Ver cuadro siguiente
Reducido	11	124	135.939

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.807	135.879	0.060	0.80650

El Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variable TEMPER2 no forma parte del modelo.

Prueba dejando la variable "AGEOFMEN" (ageOfMenarche) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.048	8	.853

Dado que sig = 0.853 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			TUMOR		Percentage Correct
			benign	malignant	
Step 1	TUMOR	benign	65	13	83.3
		malignant	18	41	69.5
	Overall Percentage				77.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ACTINGOU	-.066	.144	.207	1	.649	.937
	CRITICIS	.102	.105	.947	1	.331	1.108
	PARANOID	.132	.192	.472	1	.492	1.141
	SELFCRIT	-.075	.102	.540	1	.462	.928
	GUILT	-.311	.203	2.353	1	.125	.733
	DIRECTIO	.116	.064	3.299	1	.069	1.123
	THYROID(1)	1.892	1.338	2.002	1	.157	6.636
	ALLERGY(1)	-1.835	.595	9.523	1	.002	.160
	TEMPER1(1)	2.144	.486	19.490	1	.000	8.530
	FEELING1(1)	.789	.459	2.958	1	.085	2.201
	Constant	-1.608	.623	6.658	1	.010	.200

a. Variable(s) entered on step 1: ACTINGOU, CRITICIS, PARANOID, SELFCRIT, GUILT, DIRECTIO, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "ACTINGOU" (actingOutHostility) tiene nivel de significación: sig = 0.649 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	11	125	135.939
Reducido	10	126	136.220

$\chi^2_{1,\alpha}$	α
0.281	0.59605

0.59605 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	11	125	Ver cuadro siguiente
Reducido	10	126	136.220

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.806	136.159	0.061	0.80492
FEELING2	0.841	136.180	0.040	0.84148

En cada prueba, el Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2 y FEELING2 no forman parte del modelo.

Prueba dejando la variable "ACTINGOU" (actingOutHostility) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.877	8	.868

Dado que sig = 0.868 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	65	13	83.3
		malignant	19	40	67.8
	Overall Percentage				76.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CRITICIS	.087	.099	.767	1	.381	1.091
	PARANOID	.119	.191	.392	1	.531	1.127
	SELFCRIT	-.075	.102	.537	1	.463	.928
	GUILT	-.344	.193	3.189	1	.074	.709
	DIRECTIO	.118	.064	3.456	1	.063	1.125
	THYROID(1)	1.920	1.361	1.990	1	.158	6.818
	ALLERGY(1)	-1.849	.595	9.644	1	.002	.157
	TEMPER1(1)	2.169	.484	20.122	1	.000	8.750
	FEELING1(1)	.781	.458	2.912	1	.088	2.184
	Constant	-1.737	.559	9.668	1	.002	.176

a. Variable(s) entered on step 1: CRITICIS, PARANOID, SELFCRIT, GUILT, DIRECTIO, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "SELFCRIT" (selfCriticism) tiene nivel de significación: sig = 0.463 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	10	126	136.220
Reducido	9	127	136.426

$\chi^2_{1,\alpha}$	α
0.206	0.64992

0.64992 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	G. L.	Deviance
Completo	126	Ver cuadro siguiente
Reducido	127	136.426

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.960	136.424	0.002	0.96433
FEELING2	0.825	136.378	0.048	0.82658
AGEOFMEN	0.613	136.181	0.245	0.62062

En cada prueba, el Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2 y AGEOFMEN no forman parte del modelo.

Prueba dejando la variable "SELFCRIPT" (selfCriticism) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.338	8	.825

Dado que sig = 0.825 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	65	13	83.3
		malignant	18	41	69.5
	Overall Percentage				77.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CRITICIS	.075	.098	.587	1	.443	1.078
	PARANOID	.087	.187	.218	1	.640	1.091
	GUILT	-.381	.187	4.152	1	.042	.683
	DIRECTIO	.092	.053	3.025	1	.082	1.097
	THYROID(1)	1.927	1.365	1.993	1	.158	6.870
	ALLERGY(1)	-1.837	.597	9.473	1	.002	.159
	TEMPER1(1)	2.173	.483	20.281	1	.000	8.786
	FEELING1(1)	.721	.449	2.585	1	.108	2.057
	Constant	-1.805	.556	10.531	1	.001	.165

a. Variable(s) entered on step 1: CRITICIS, PARANOID, GUILT, DIRECTIO, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "PARANOID" (paranoidHostility) tiene nivel de significación: sig = 0.640 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	127	136.426
Reducido	8	128	136.965

$\chi^2_{1,\alpha}$	α
0.539	0.46285

0.46285 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	127	Ver cuadro siguiente
Reducido	9	128	136.965

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.943	136.960	0.005	0.94363
FEELING2	0.822	136.915	0.050	0.82306
AGEOFMEN	0.602	136.704	0.261	0.60943
ACTINGOU	0.651	136.761	0.204	0.65151

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN y ACTINGOU no forman parte del modelo.

Prueba dejando la variable "PARANOID" (paranoidHostility) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.887	8	.770

Dado que sig = 0.770 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted		
		TUMOR		Percentage Correct
		benign	malignant	
Step 1 TUMOR	benign	65	13	83.3
	malignant	19	40	67.8
Overall Percentage				76.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CRITICIS	.098	.085	1.313	1	.252	1.103
	GUILT	-.351	.173	4.122	1	.042	.704
	DIRECTIO	.097	.051	3.656	1	.056	1.102
	THYROID(1)	2.038	1.375	2.197	1	.138	7.673
	ALLERGY(1)	-1.841	.597	9.500	1	.002	.159
	TEMPER1(1)	2.152	.478	20.258	1	.000	8.606
	FEELING1(1)	.716	.448	2.555	1	.110	2.047
	Constant	-1.887	.526	12.891	1	.000	.151

a. Variable(s) entered on step 1: CRITICIS, GUILT, DIRECTIO, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "CRITICIS" (criticismOfOthers) tiene nivel de significación: sig = 0.252 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	128	136.965
Reducido	7	129	137.186

$\chi^2_{1,\alpha}$	α
0.221	0.63828

0.63828 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	129	Ver cuadro siguiente
Reducido	7	128	137.186

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.889	137.167	0.019	0.89037
FEELING2	0.898	137.170	0.016	0.89934
AGEOFMEN	0.687	137.029	0.157	0.69193
ACTINGOU	0.708	137.047	0.139	0.70928
SELFCRIT	0.546	136.822	0.364	0.54629

En cada prueba. el Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU y SELFCRIT no forman parte del modelo.

Prueba dejando la variable "CRITICIS" (criticismOfOthers) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.766	8	.878

Dado que sig = 0.878 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	64	14	82.1
		malignant	20	39	66.1
	Overall Percentage				75.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GUILT	-.268	.154	3.055	1	.080	.765
	DIRECTIO	.091	.050	3.348	1	.067	1.095
	THYROID(1)	2.202	1.412	2.433	1	.119	9.044
	ALLERGY(1)	-1.796	.593	9.161	1	.002	.166
	TEMPER1(1)	2.158	.476	20.522	1	.000	8.658
	FEELING1(1)	.696	.443	2.462	1	.117	2.005
	Constant	-1.527	.398	14.734	1	.000	.217

a. Variable(s) entered on step 1: GUILT, DIRECTIO, THYROID, ALLERGY, TEMPER1, FEELING1.

La variable "THYROID" tiene nivel de significación: sig = 0.119 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	129	137.186
Reducido	6	130	138.511

$\chi^2_{1,\alpha}$	α
1.325	0.24970

0.24970 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	129	Ver cuadro siguiente
Reducido	6	130	138.511

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.825	138.462	0.049	0.82481
FEELING2	0.909	138.498	0.013	0.90922
AGEOFMEN	0.958	138.508	0.003	0.95632
ACTINGOU	0.877	138.487	0.024	0.87688
SELFCRIT	0.821	138.460	0.051	0.82133
PARANOID	0.337	137.558	0.953	0.32896

En cada prueba, El Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU, SELFCRIT y PARANOID no forman parte del modelo.

Prueba dejando la variable "THYROID" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.999	8	.934

Dado que sig = 0.934 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted		
		TUMOR		Percentage Correct
		benign	malignant	
Step 1 TUMOR	benign	63	15	80.8
	malignant	21	38	64.4
Overall Percentage				73.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GUILT	-.276	.149	3.424	1	.064	.759
	DIRECTIO	.091	.048	3.603	1	.058	1.095
	ALLERGY(1)	-1.697	.577	8.647	1	.003	.183
	TEMPER1(1)	2.118	.465	20.786	1	.000	8.317
	FEELING1(1)	.818	.434	3.549	1	.060	2.267
	Constant	-1.512	.392	14.910	1	.000	.220

a. Variable(s) entered on step 1: GUILT, DIRECTIO, ALLERGY, TEMPER1, FEELING1.

La variable "GUILT" tiene nivel de significación: sig = 0.064 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	130	138.511
Reducido	5	131	141.525

$\chi^2_{1,\alpha}$	α
3.014	0.08255

0.08255 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	130	Ver cuadro siguiente
Reducido	5	131	141.525

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.943	141.520	0.005	0.94369
FEELING2	0.700	141.376	0.149	0.69949
AGEOFMEN	0.872	141.499	0.026	0.87190
ACTINGOU	0.814	141.469	0.056	0.81293
SELFCRIT	0.897	141.508	0.017	0.89626
PARANOID	0.231	140.027	1.498	0.22098
CRITICIS	0.220	139.997	1.528	0.21641

En cada prueba, el Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU, SELFCRIT, PARANOID y CRITICIS no forman parte del modelo.

Prueba dejando la variable "GUILT" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.989	8	.858

Dado que sig = 0.858 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	63	15	80.8
		malignant	24	35	59.3
	Overall Percentage				71.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DIRECTIO	.043	.040	1.107	1	.293	1.044
	ALLERGY(1)	-1.531	.549	7.784	1	.005	.216
	TEMPER1(1)	1.885	.429	19.306	1	.000	6.589
	FEELING1(1)	.798	.426	3.502	1	.061	2.221
	Constant	-1.573	.405	15.113	1	.000	.207

a. Variable(s) entered on step 1: DIRECTIO, ALLERGY, TEMPER1, FEELING1.

La variable "DIRECTIO" (directionOfHostility) tiene nivel de significación: sig = 0.293 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	5	131	141.525
Reducido	4	132	145.169

$\chi^2_{1,\alpha}$	α
3.644	0.05627

0.05627 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	5	131	Ver cuadro siguiente
Reducido	4	132	145.169

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.868	145.141	0.028	0.86711
FEELING2	0.755	145.072	0.097	0.75546
AGEOFMEN	0.618	144.931	0.238	0.62565
ACTINGOU	0.392	144.448	0.721	0.39582
SELFCRIT	0.346	144.294	0.875	0.34957
PARANOID	0.822	145.117	0.052	0.81962
CRITICIS	0.732	145.049	0.120	0.72903
THYROID	0.104	141.775	3.394	0.06543

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU, SELFCRIT, PARANOID, CRITICIS y THYROID no forman parte del modelo.

Prueba dejando la variable "DIRECTIO" (directionOfHostility) fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.580	5	.611

Dado que sig = 0.611 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted		
		TUMOR		Percentage Correct
		benign	malignant	
Step 1 TUMOR	benign	65	13	83.3
	malignant	25	34	57.6
Overall Percentage				72.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ALLERGY(1)	-1.398	.527	7.041	1	.008	.247
	TEMPER1(1)	1.904	.426	19.985	1	.000	6.715
	FEELING1(1)	.835	.424	3.867	1	.049	2.304
	Constant	-1.376	.346	15.842	1	.000	.252

a. Variable(s) entered on step 1: ALLERGY, TEMPER1, FEELING1.

Todas las variables tienen nivel de significación : sig < 0.05, las estimaciones de los coeficientes son significativas.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	132	145.169
Reducido	3	133	147.318

$\chi^2_{1,\alpha}$	α
2.149	0.14266

0.14266 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	132	Ver cuadro siguiente
Reducido	3	133	147.318

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.921	147.309	0.009	0.92442
FEELING2	0.840	147.278	0.040	0.84148
AGEOFMEN	0.869	147.290	0.028	0.86711
ACTINGOU	0.804	147.258	0.060	0.80650
SELFCRIT	0.433	146.520	0.798	0.37169
PARANOID	0.398	146.453	0.865	0.35234
CRITICIS	0.320	146.136	1.182	0.27695
THYROID	0.071	142.938	4.380	0.03636
GUILT	0.705	147.180	0.138	0.71028

Con la variable THYROID dentro del modelo el Nivel de Significación = 0.071 > 0.05 pero se puede considerar significativa, para decidirlo vemos el aporte a reducir la deviance, α < 0.05 indica que la variable aporta a reducir la deviance una vez que las otras variables han sido incluidas en el modelo. En otras pruebas la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco al ajuste del modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU, SELFCRIT, PARANOID, CRITICIS y GUILT no forman parte del modelo. La variable THYROID forma parte del modelo.

Modelo cuando ingresa la variable "THYROID" dentro del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.603	4	.626

Dado que sig = 0.626 > 0.05 no se rechaza la hipótesis nula de buen ajuste

Classification Table[a]

			Predicted		
			TUMOR		Percentage Correct
	Observed		benign	malignant	
Step 1	TUMOR	benign	64	14	82.1
		malignant	22	37	62.7
	Overall Percentage				73.7

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ALLERGY(1)	-1.608	.564	8.130	1	.004	.200
	TEMPER1(1)	1.934	.435	19.739	1	.000	6.917
	FEELING1(1)	.685	.436	2.472	1	.116	1.984
	THYROID(1)	2.292	1.269	3.263	1	.071	9.893
	Constant	-1.363	.347	15.393	1	.000	.256

a. Variable(s) entered on step 1: ALLERGY, TEMPER1, FEELING1, THYROID.

La variable "FEELING1" tiene el Nivel de Sig = 0.116 > 0.05 por lo tanto se considera que la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	132	142.938
Reducido	3	133	147.318

$\chi^2_{1,\alpha}$	α
4.38	0.03636

0.03636 < 0.05 indica que la variable que no se incluyó en el modelo aporta a reducir la Deviance una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variable "THYROID" forma parte del modelo.

Prueba dejando la variable "FEELING1" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.367	2	.505

Dado que sig = 0.505 > 0.05 no se rechaza la hipótesis nula de buen ajuste

Classification Table[a]

			Predicted		
			TUMOR		Percentage
	Observed		benign	malignant	Correct
Step 1	TUMOR	benign	65	13	83.3
		malignant	22	37	62.7
	Overall Percentage				74.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ALLERGY(1)	-1.564	.561	7.776	1	.005	.209
	TEMPER1(1)	2.132	.421	25.670	1	.000	8.431
	THYROID(1)	2.596	1.270	4.179	1	.041	13.408
	Constant	-1.096	.290	14.232	1	.000	.334

a. Variable(s) entered on step 1: ALLERGY, TEMPER1, THYROID.

Todas las variables tiene el Nivel de Sig. < 0.05, se considera que las estimaciones de los coeficientes es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	132	142.938
Reducido	3	133	145.400

$\chi^2_{1,\alpha}$	α
2.462	0.11663

0.11663 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	132	Ver cuadro siguiente
Reducido	3	133	145.400

Nombre de variable	Nivel de Significación (Sig)	Deviance	$\chi^2_{1,\alpha}$	α
TEMPER2	0.784	145.325	0.075	0.78419
FEELING2	0.242	144.042	1.358	0.24388
AGEOFMEN	0.920	145.389	0.011	0.91647
ACTINGOU	0.772	145.317	0.083	0.77327
SELFCRIT	0.669	145.182	0.218	0.64057
PARANOID	0.623	145.120	0.280	0.59670
CRITICIS	0.558	145.028	0.372	0.54192
GUILT	0.627	145.170	0.230	0.63152

El Nivel de Significación > 0.05 indica que en cada prueba la estimación del coeficiente no es significativa.
En cada una de las pruebas, α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variables TEMPER2, FEELING2, AGEOFMEN, ACTINGOU, SELFCRIT, PARANOID, CRITICIS y GUILT no forman parte del modelo.

Pruebas con las variables que son individualmente significativas

Modelo	Cantidad de variables	G. L.	Deviance
Completo	3	133	145.400
Reducido	2	134	Ver cuadro siguiente

Nombre de variable fuera del modelo	Deviance	$\chi^2_{1,\alpha}$	α
ALLERGY	154.368	8.968	0.00275
TEMPER1	175.331	29.931	0.00000
THYROID	151.191	5.791	0.01611

En cada prueba la variable que sale del modelo tiene $\alpha < 0.05$, indicando que en cada prueba la variable aporta al modelo una vez que las otras fueron incluidas en el modelo.

Decisión: las variables ALLERGY, TEMPER1 y THYROID forman parte del modelo.

7.1.3 Ejemplo 3 "Intensive Care Unit (ICU)" [HOS/89]

Estadísticas Descriptivas

Tablas de Distribución de Frecuencias de las variables categóricas

Bicarbonate from Initial Blood Gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	>=18	185	92.5	92.5	92.5
	<18	15	7.5	7.5	100.0
	Total	200	100.0	100.0	

Cancer Part of Present Problem

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	180	90.0	90.0	90.0
	Yes	20	10.0	10.0	100.0
	Total	200	100.0	100.0	

CPR Prior to ICU Admission

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	187	93.5	93.5	93.5
	Yes	13	6.5	6.5	100.0
	Total	200	100.0	100.0	

Creatinine from Initial Blood Gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<=2.0	190	95.0	95.0	95.0
	>2.0	10	5.0	5.0	100.0
	Total	200	100.0	100.0	

History of Chronic Renal Failure

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	181	90.5	90.5	90.5
	Yes	19	9.5	9.5	100.0
	Total	200	100.0	100.0	

Long Bone, Multiple, Neck, Single Area, or Hip Fracture

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	185	92.5	92.5	92.5
	Yes	15	7.5	7.5	100.0
	Total	200	100.0	100.0	

Infection Probable at ICU Admission

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	116	58.0	58.0	58.0
	Yes	84	42.0	42.0	100.0
	Total	200	100.0	100.0	

La variable "Level of Consciousness at ICU Admission (LOC)" fue transformada en dos variables dummy "LOC1 y "LOC2".

Level of Consciousness at ICU Admission

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Coma or Stupor	185	92.5	92.5	92.5
	Deep Stupor	5	2.5	2.5	95.0
	Coma	10	5.0	5.0	100.0
	Total	200	100.0	100.0	

PCO2 from initial Blood Gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<=45	180	90.0	90.0	90.0
	>45	20	10.0	10.0	100.0
	Total	200	100.0	100.0	

PH from Initial Blood Gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	>=7.25	187	93.5	93.5	93.5
	<7.25	13	6.5	6.5	100.0
	Total	200	100.0	100.0	

PO2 from Initial Blood Gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	>60	184	92.0	92.0	92.0
	<=60	16	8.0	8.0	100.0
	Total	200	100.0	100.0	

Previous Admission to an ICU within 6 months

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	170	85.0	85.0	85.0
	Yes	30	15.0	15.0	100.0
	Total	200	100.0	100.0	

La variable RACE fue transformada en dos variables dummy "RACE1" y "RACE2".

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	175	87.5	87.5	87.5
	Black	15	7.5	7.5	95.0
	Other	10	5.0	5.0	100.0
	Total	200	100.0	100.0	

Service at ICU Admission

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Medical	93	46.5	46.5	46.5
	Surgical	107	53.5	53.5	100.0
	Total	200	100.0	100.0	

SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	124	62.0	62.0	62.0
	Female	76	38.0	38.0	100.0
	Total	200	100.0	100.0	

Type of Admission

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Elective	53	26.5	26.5	26.5
	Emergency	147	73.5	73.5	100.0
	Total	200	100.0	100.0	

Tabla de Distribución de Frecuencias de la variable de respuesta "STA"

Vital Status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Lived	160	80.0	80.0	80.0
	Died	40	20.0	20.0	100.0
	Total	200	100.0	100.0	

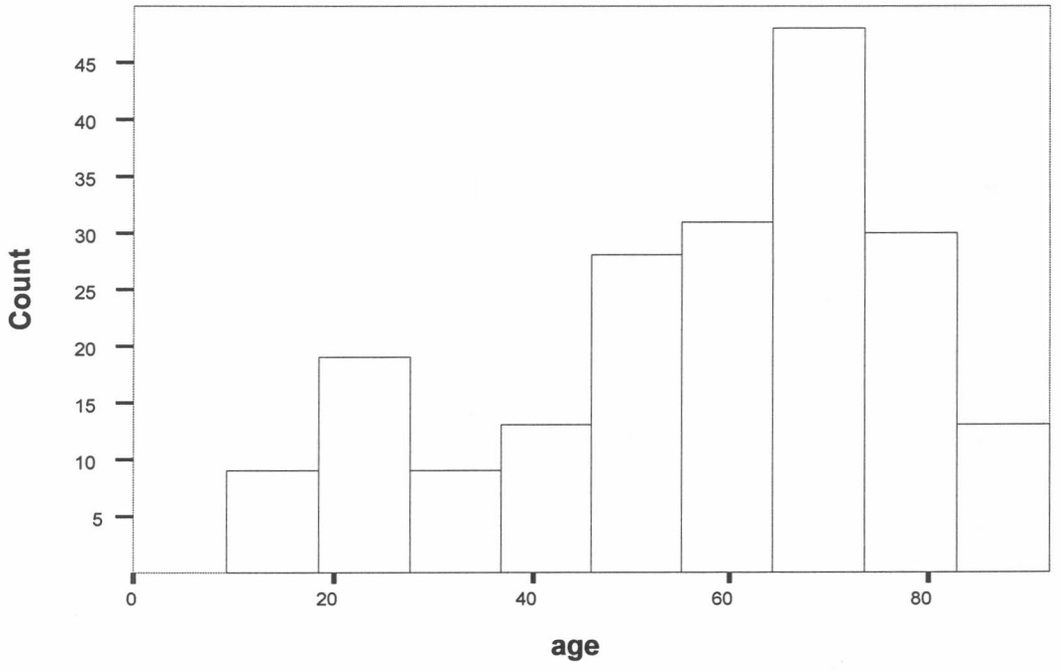
Estadísticas Descriptivas e Histogramas de las variables continuas

Descriptive Statistics

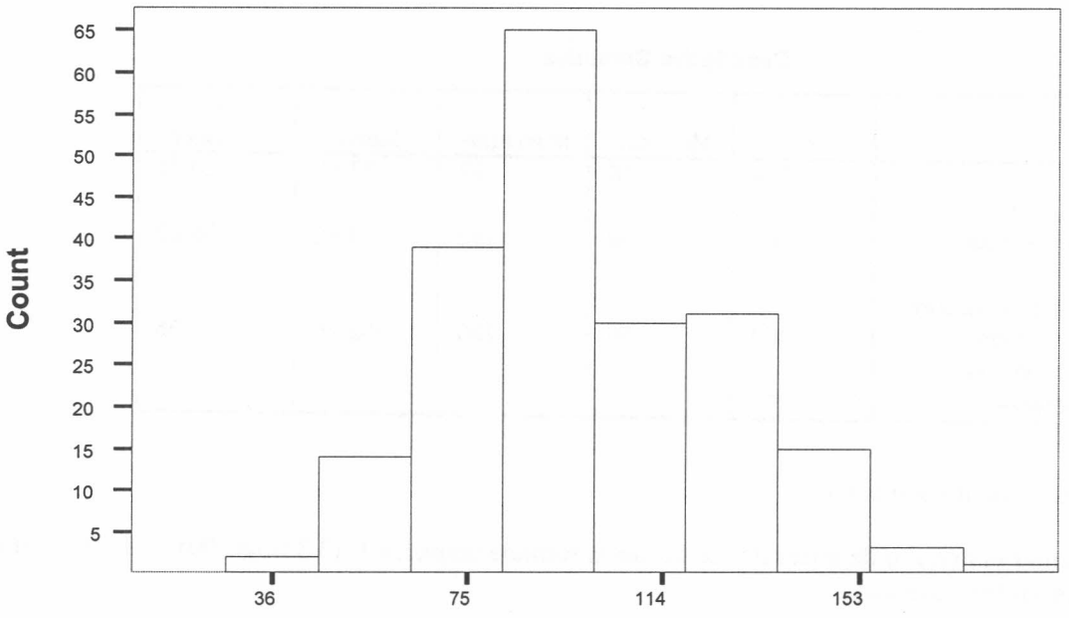
	N	Minimum	Maximum	Mean	Std. Deviation
AGE	200	16	92	57.55	20.05
Heart Rate at ICU Admission medida en Beats/min	200	39	192	98.92	26.83
Systolic Blood Pressure at ICU Admission medida en mm Hg	200	36	256	132.28	32.95
Valid N (listwise)	200				

Todas las variables son enteras.

Para calcular el número de intervalos se utiliza la fórmula empírica 1+(3.3*log(200)) = 8.59 se utilizan 9 intervalos en los histogramas.

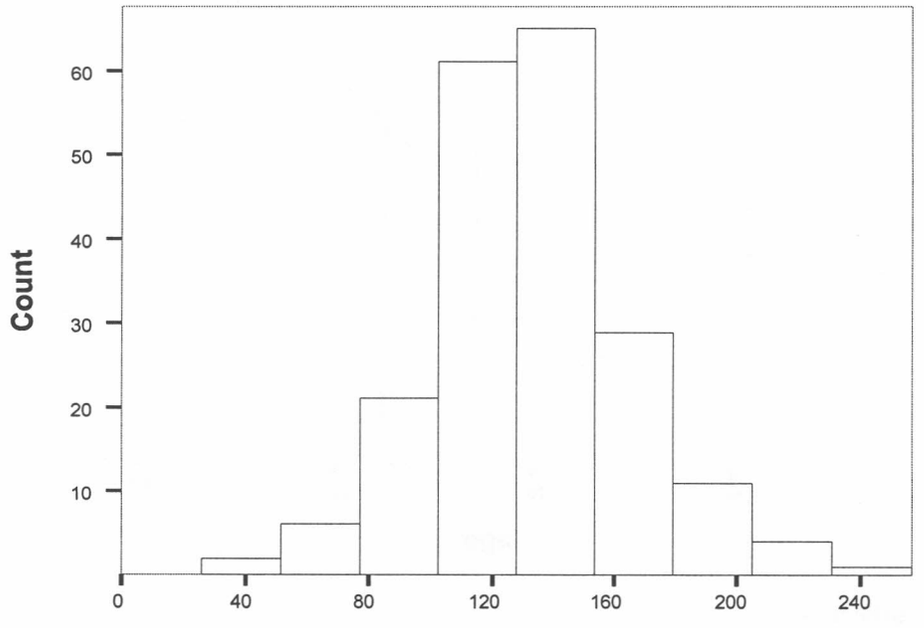


No se encuentran outliers



Heart Rate at ICU Admission medida en Beats/min

No se encuentran outliers.



Systolic Blood Pressure at ICU Admission medida en mm Hg

No se encuentran outliers.

Análisis Discriminante de Fisher

Vectores de Medias y Matriz de Varianzas-Covarianzas

Vectores de Medias	Media de la clase 0	Media de la clase 1
AGE	55.6500	65.1250
BIC	0.0625	0.1250
CAN	0.1000	0.1000
CPR	0.0375	0.1750
CRE	0.0313	0.1250
CRN	0.0688	0.2000
FRA	0.0750	0.0750
HRA	98.5000	100.6250
INF	0.3750	0.6000
LOC1	0.0000	0.1250
LOC2	0.0125	0.2000
PCO	0.1000	0.1000
PH	0.0563	0.1000
PO2	0.0688	0.1250
PRE	0.1438	0.1750
RACE1	0.8625	0.9250
RACE2	0.0875	0.0250
SER	0.5813	0.3500
SEX	0.3750	0.4000
SYS	135.6438	118.8250
TYP	0.6813	0.9500

Matriz de Varianzas-Covarianzas Combinada

389.711	0.242	-0.146	-0.155	0.339	0.727	-1.602	16.954	1.187	-0.099	0.110	0.990	0.170	0.764	0.450	1.205	-0.683	0.398	0.904	54.045	-2.079
0.242	0.069	-0.008	-0.001	0.015	0.012	-0.006	1.095	0.021	-0.003	-0.001	-0.003	0.020	0.014	-0.007	0.004	0.000	-0.038	0.001	-1.139	0.017
-0.146	-0.008	0.091	-0.002	0.000	-0.005	-0.008	0.134	-0.007	-0.003	-0.005	0.010	-0.002	-0.003	-0.015	0.003	0.003	0.032	0.007	0.623	-0.049
-0.155	-0.001	-0.002	0.058	0.005	0.011	0.000	0.756	0.013	-0.004	0.018	0.009	0.005	-0.007	-0.011	-0.003	0.002	-0.025	0.015	-0.539	0.011
0.339	0.015	0.000	0.005	0.047	0.018	0.001	0.477	0.006	-0.003	0.005	-0.005	0.006	0.000	-0.008	-0.005	0.002	-0.013	0.011	-0.406	0.009
0.727	0.012	-0.005	0.011	0.018	0.084	-0.002	-0.154	0.010	0.000	0.006	-0.005	0.013	0.001	0.005	-0.010	0.004	-0.016	0.008	1.128	0.010
-1.602	-0.006	-0.008	0.000	0.001	-0.002	0.070	-0.479	-0.002	-0.002	0.001	-0.003	0.000	-0.001	-0.001	0.004	-0.001	0.025	-0.004	-0.940	0.015
16.954	1.095	0.134	0.756	0.477	-0.154	-0.479	722.734	4.066	-0.314	-0.394	1.619	1.571	1.547	0.314	0.126	0.209	-4.345	0.510	-44.501	1.999
1.187	0.021	-0.007	0.013	0.006	0.010	-0.002	4.066	0.238	0.000	0.002	0.023	0.021	0.025	0.031	0.015	-0.009	-0.047	0.005	-3.133	0.027
-0.099	-0.003	-0.003	-0.004	-0.003	0.000	-0.002	-0.314	0.000	0.022	-0.005	-0.003	-0.003	0.002	0.001	-0.008	0.004	0.006	0.000	0.929	-0.004
0.110	-0.001	-0.005	0.018	0.005	0.006	0.001	-0.394	0.002	-0.005	0.042	0.010	-0.005	-0.001	-0.009	0.004	-0.002	-0.005	0.005	-1.151	0.005
0.990	-0.003	0.010	0.009	-0.005	-0.005	-0.003	1.619	0.023	-0.003	0.010	0.091	0.034	0.032	-0.010	0.008	-0.003	-0.024	0.017	0.073	0.002
0.170	0.020	-0.002	0.005	0.006	0.013	0.000	1.571	0.021	-0.003	-0.005	0.034	0.061	0.020	-0.005	0.008	-0.004	-0.028	0.010	0.080	0.010
0.764	0.014	-0.003	-0.007	0.000	0.001	-0.001	1.547	0.025	0.002	-0.001	0.032	0.020	0.074	-0.002	0.004	-0.005	-0.031	-0.006	-0.749	0.009
0.450	-0.007	-0.015	-0.011	-0.008	0.005	-0.001	0.314	0.031	0.001	-0.009	-0.010	-0.005	-0.002	0.129	-0.017	0.009	0.016	0.013	-0.170	-0.012
1.205	0.004	0.003	-0.003	-0.005	-0.010	0.004	0.126	0.015	-0.008	0.004	0.008	0.008	0.004	-0.017	0.110	-0.066	0.014	0.002	-0.966	-0.011
-0.683	0.000	0.003	0.002	0.002	0.004	-0.001	0.209	-0.009	0.004	-0.002	-0.003	-0.004	-0.005	0.009	-0.066	0.069	-0.018	-0.003	0.496	0.008
0.398	-0.038	0.032	-0.025	-0.013	-0.016	0.025	-4.345	-0.047	0.006	-0.005	-0.024	-0.028	-0.031	0.016	0.014	-0.018	0.243	-0.018	0.897	-0.109
0.904	0.001	0.007	0.015	0.011	0.008	-0.004	0.510	0.005	0.000	0.005	0.017	0.010	-0.006	0.013	0.002	-0.003	-0.018	0.238	0.865	0.025
54.045	-1.139	0.623	-0.539	-0.406	1.128	-0.940	-44.501	-3.133	0.929	-1.151	0.073	0.080	-0.749	-0.170	-0.966	0.496	0.897	0.865	1045.608	-1.967
-2.079	0.017	-0.049	0.011	0.009	0.010	0.015	1.999	0.027	-0.004	0.005	0.002	0.010	0.009	-0.012	-0.011	0.008	-0.109	0.025	-1.967	0.185

Criterio de ajuste del modelo en cada paso

F de la tabla = 3.8883

número de paso	variable sale/entra	SCD MC	SCD MC SCD 0	SCD MC SCD 1	SCD MR	SCD MR SCD 0	SCD MR SCD 1	SCE MC	SCE MR	SCD MC-MR	SCE MC-MR	razón
1	sale: AGE NO	4158	2884.2915	1273.7085	3960	2724.7544	1235.2456	127.39218	109.575	198	17.8172	17.9972
1	sale: BIC SI	4158	2884.2915	1273.7085	3960	2760.1629	1199.8371	127.39218	127.3607	198	0.03151	0.03183
2	sale:AGE NO	3960	2760.1629	1199.8371	3762	2600.8112	1161.1888	127.360667	109.5621	198	17.7986	17.9784
2	sale: CAN NO	3960	2760.1629	1199.8371	3762	2605.2242	1156.7758	127.360667	111.5882	198	15.7725	15.9318
2	sale: CPR SI	3960	2760.1629	1199.8371	3762	2641.9997	1120.0003	127.360667	124.9447	198	2.41594	2.44035
3	sale:AGE NO	3762	2641.9997	1120.0003	3564	2482.767	1081.233	124.944723	107.5298	198	17.415	17.5909
3	entra: BIC NO	3960	2768.5553	1191.4447	3762	2641.9997	1120.0003	125.055399	124.9447	198	0.11068	0.11179
3	sale:CAN NO	3762	2641.9997	1120.0003	3564	2487.0283	1076.9717	124.944723	109.1377	198	15.807	15.9667
3	sale: CRE SI	3762	2641.9997	1120.0003	3564	2525.9062	1038.0938	124.944723	124.0623	198	0.88238	0.89129
4	sale: AGE NO	3564	2525.9062	1038.0938	3366	2365.3596	1000.6404	124.062346	105.4099	198	18.6525	18.8409
4	entra: BIC NO	3762	2659.7068	1102.2932	3564	2525.9062	1038.0938	124.086199	124.0623	198	0.02385	0.02409
4	sale: CAN NO	3564	2525.9062	1038.0938	3366	2371.5248	994.47524	124.062346	107.7889	198	16.2735	16.4379
4	entra: CPR NO	3762	2644.8119	1117.1881	3564	2525.9062	1038.0938	126.366972	124.0623	198	2.30463	2.32791
4	sale: CRN SI	3564	2525.9062	1038.0938	3366	2400.2259	965.77408	124.062346	123.7048	198	0.35759	0.3612
5	sale: AGE NO	3366	2400.2259	965.77408	3168	2238.7695	929.23053	123.704757	103.4347	198	20.2701	20.4748
5	entra: BIC NO	3564	2535.3944	1028.6056	3366	2400.2259	965.77408	123.718109	123.7048	198	0.01335	0.01349
5	sale: CAN NO	3366	2400.2259	965.77408	3168	2245.7064	922.29363	123.704757	107.0726	198	16.6321	16.8001
5	entra: CPR NO	3564	2520.06	1043.94	3366	2400.2259	965.77408	126.181423	123.7048	198	2.47667	2.50168
5	entra: CRE NO	3564	2511.7114	1052.2886	3366	2400.2259	965.77408	124.806818	123.7048	198	1.10206	1.11319
5	sale: FRA SI	3366	2400.2259	965.77408	3168	2238.5552	929.44478	123.704757	121.7092	198	1.99553	2.01568
6	sale: AGE NO	3168	2238.5552	929.44478	2970	2073.3116	896.68839	121.70923	103.4331	198	18.2761	18.4607
6	entra: BIC NO	3366	2373.366	992.634	3168	2238.5552	929.44478	121.746718	121.7092	198	0.03749	0.03787
6	sale: CAN NO	3168	2238.5552	929.44478	2970	2084.65	885.34995	121.70923	106.2355	198	15.4737	15.63
6	entra: CPR NO	3366	2357.7705	1008.2295	3168	2238.5552	929.44478	124.388367	121.7092	198	2.67914	2.7062
6	entra: CRE NO	3366	2349.2695	1016.7305	3168	2238.5552	929.44478	123.058748	121.7092	198	1.34952	1.36315
6	entra: CRN NO	3366	2364.0071	1001.9929	3168	2238.5552	929.44478	122.153632	121.7092	198	0.4444	0.44889

6	sale: HRA SI	3168	2238.5552	929.44478	2970	2083.1485	886.85145	121.70923	121.6183	198	0.0909	0.09181
7	sale: AGE NO	2970	2083.1485	886.85145	2772	1917.9085	854.09147	121.618334	103.3435	198	18.2748	18.4594
7	entra: BIC NO	3168	2218.0353	949.96468	2970	2083.1485	886.85145	121.657419	121.6183	198	0.03908	0.03948
7	sale: CAN NO	2970	2083.1485	886.85145	2772	1928.987	843.01303	121.618334	106.2354	198	15.383	15.5383
7	entra: CPR NO	3168	2200.9604	967.03959	2970	2083.1485	886.85145	124.200285	121.6183	198	2.58195	2.60803
7	entra: CRE NO	3168	2194.2534	973.74656	2970	2083.1485	886.85145	122.930714	121.6183	198	1.31238	1.32564
7	entra: CRN NO	3168	2207.0074	960.99259	2970	2083.1485	886.85145	122.092759	121.6183	198	0.47442	0.47922
7	entra: FRA NO	3168	2244.6977	923.30227	2970	2083.1485	886.85145	123.630697	121.6183	198	2.01236	2.03269
7	sale: INF SI	2970	2083.1485	886.85145	2772	1921.5607	850.43932	121.618334	121.6158	198	0.00257	0.00259
8	sale: AGE NO	2772	1921.5607	850.43932	2574	1756.2407	817.75933	121.615767	103.0007	198	18.615	18.8031
8	entra: BIC NO	2970	2054.4345	915.56552	2772	1921.5607	850.43932	121.65298	121.6158	198	0.03721	0.03759
8	sale: CAN NO	2772	1921.5607	850.43932	2574	1766.4756	807.5244	121.615767	106.1225	198	15.4933	15.6498
8	entra: CPR NO	2970	2036.5254	933.47457	2772	1921.5607	850.43932	124.182601	121.6158	198	2.56683	2.59276
8	entra: CRE NO	2970	2031.0556	938.94439	2772	1921.5607	850.43932	122.930713	121.6158	198	1.31495	1.32823
8	entra: CRN NO	2970	2045.0385	924.96153	2772	1921.5607	850.43932	122.092517	121.6158	198	0.47675	0.48157
8	entra: FRA NO	2970	2083.1704	886.8296	2772	1921.5607	850.43932	123.630661	121.6158	198	2.01489	2.03525
8	entra: HRA NO	2970	2080.1454	889.85458	2772	1921.5607	850.43932	121.695636	121.6158	198	0.07987	0.08068
8	sale: LOC1 NO	2772	1921.5607	850.43932	2574	1908.4128	665.58723	121.615767	70.93572	198	50.6801	51.192
8	sale: LOC2 NO	2772	1921.5607	850.43932	2574	1870.1684	703.83161	121.615767	82.65729	198	38.9585	39.352
8	sale: PCO NO	2772	1921.5607	850.43932	2574	1778.7845	795.21549	121.615767	115.684	198	5.93175	5.99167
8	sale: PH NO	2772	1921.5607	850.43932	2574	1786.1951	787.80485	121.615767	116.1065	198	5.50929	5.56494
8	sale: PO2 SI	2772	1921.5607	850.43932	2574	1801.9458	772.05416	121.615767	121.4695	198	0.14632	0.14779
9	sale: AGE NO	2574	1801.9458	772.05416	2376	1635.7511	740.24893	121.469452	102.986	198	18.4835	18.6702
9	entra: BIC NO	2772	1940.7203	831.27969	2574	1801.9458	772.05416	121.529171	121.4695	198	0.05972	0.06032
9	sale: CAN NO	2574	1801.9458	772.05416	2376	1646.7325	729.2675	121.469452	105.9607	198	15.5088	15.6654
9	entra: CPR NO	2772	1919.8352	852.16477	2574	1801.9458	772.05416	124.173906	121.4695	198	2.70445	2.73177
9	entra: CRE NO	2772	1911.196	860.80404	2574	1801.9458	772.05416	122.775009	121.4695	198	1.30556	1.31874
9	entra: CRN NO	2772	1925.2572	846.74283	2574	1801.9458	772.05416	121.943605	121.4695	198	0.47415	0.47894
9	entra: FRA NO	2772	1962.1452	809.85479	2574	1801.9458	772.05416	123.390423	121.4695	198	1.92097	1.94037
9	entra: HRA NO	2772	1962.1843	809.81571	2574	1801.9458	772.05416	121.566934	121.4695	198	0.09748	0.09847
9	entra: INF NO	2772	1964.637	807.36299	2574	1801.9458	772.05416	121.470456	121.4695	198	0.001	0.00101
9	sale: LOC1 NO	2574	1801.9458	772.05416	2376	1789.4222	586.57783	121.469452	70.74236	198	50.7271	51.2395
9	sale: LOC2 NO	2574	1801.9458	772.05416	2376	1751.2366	624.76341	121.469452	82.01472	198	39.4547	39.8533
9	sale: PCO NO	2574	1801.9458	772.05416	2376	1652.5724	723.42755	121.469452	114.2793	198	7.19013	7.26276
9	sale: PH NO	2574	1801.9458	772.05416	2376	1665.9352	710.06482	121.469452	116.0855	198	5.38398	5.43837

9	sale: PRE NO	2574	1801.9458	772.05416	2376	1648.6824	727.31764	121.469452	117.1394	198	4.33002	4.37375
9	sale: RACE1 SI	2574	1801.9458	772.05416	2376	1641.3892	734.61082	121.469452	121.4665	198	0.00295	0.00298
10	sale: AGE NO	2376	1641.3892	734.61082	2178	1473.6787	704.32128	121.466501	102.7206	198	18.7459	18.9353
10	entra: BIC NO	2574	1780.1044	793.89557	2376	1641.3892	734.61082	121.526966	121.4665	198	0.06047	0.06108
10	sale: CAN NO	2376	1641.3892	734.61082	2178	1485.9935	692.00649	121.466501	105.9368	198	15.5297	15.6866
10	entra: CPR NO	2574	1760.3144	813.68563	2376	1641.3892	734.61082	124.167375	121.4665	198	2.70087	2.72816
10	entra: CRE NO	2574	1752.0587	821.94126	2376	1641.3892	734.61082	122.770171	121.4665	198	1.30367	1.31684
10	entra: CRN NO	2574	1760.3531	813.64693	2376	1641.3892	734.61082	121.94302	121.4665	198	0.47652	0.48133
10	entra: FRA NO	2574	1802.906	771.094	2376	1641.3892	734.61082	123.341944	121.4665	198	1.87544	1.89439
10	entra: HRA NO	2574	1800.8952	773.10482	2376	1641.3892	734.61082	121.564996	121.4665	198	0.09849	0.09949
10	entra: INF NO	2574	1803.9028	770.09717	2376	1641.3892	734.61082	121.467336	121.4665	198	0.00083	0.00084
10	sale: LOC1 NO	2376	1641.3892	734.61082	2178	1630.3936	547.60636	121.466501	70.18859	198	51.2779	51.7959
10	sale: LOC2 NO	2376	1641.3892	734.61082	2178	1590.7695	587.23052	121.466501	82.00395	198	39.4625	39.8612
10	sale: PCO NO	2376	1641.3892	734.61082	2178	1492.0166	685.98339	121.466501	114.2738	198	7.19267	7.26533
10	sale: PH NO	2376	1641.3892	734.61082	2178	1505.4566	672.54336	121.466501	116.0827	198	5.38384	5.43822
10	entra: PO2 NO	2574	1759.9896	814.01035	2376	1641.3892	734.61082	121.609902	121.4665	198	0.1434	0.14485
10	sale: PRE NO	2376	1641.3892	734.61082	2178	1488.6293	689.37066	121.466501	117.0722	198	4.39429	4.43868
10	sale: RACE2 NO	2376	1641.3892	734.61082	2178	1456.397	721.60295	121.466501	117.5456	198	3.92094	3.96055
10	sale:SER SI	2376	1641.3892	734.61082	2178	1491.9594	686.04057	121.466501	119.6903	198	1.77625	1.79419
11	sale: AGE NO	2178	1491.9594	686.04057	1980	1325.9051	654.09489	119.69025	99.74636	198	19.9439	20.1453
11	entra: BIC NO	2376	1630.2041	745.79595	2178	1491.9594	686.04057	119.692518	119.6903	198	0.00227	0.00229
11	sale: CAN NO	2178	1491.9594	686.04057	1980	1336.0209	643.97905	119.69025	104.6839	198	15.0063	15.1579
11	entra: CPR NO	2376	1610.1532	765.84676	2178	1491.9594	686.04057	123.054017	119.6903	198	3.36377	3.39774
11	entra: CRE NO	2376	1602.6242	773.37577	2178	1491.9594	686.04057	121.161518	119.6903	198	1.47127	1.48613
11	entra: CRN NO	2376	1611.5087	764.49135	2178	1491.9594	686.04057	120.249405	119.6903	198	0.55915	0.5648
11	entra: FRA NO	2376	1648.6294	727.3706	2178	1491.9594	686.04057	120.485004	119.6903	198	0.79475	0.80278
11	entra: HRA NO	2376	1649.0537	726.9463	2178	1491.9594	686.04057	119.690638	119.6903	198	0.00039	0.00039
11	entra: INF NO	2376	1655.4915	720.50855	2178	1491.9594	686.04057	119.737311	119.6903	198	0.04706	0.04754
11	sale: LOC1 NO	2178	1491.9594	686.04057	1980	1481.5619	498.43805	119.69025	69.48749	198	50.2028	50.7099
11	sale: LOC2 NO	2178	1491.9594	686.04057	1980	1441.4085	538.59152	119.69025	80.13729	198	39.553	39.9525
11	sale: PCO NO	2178	1491.9594	686.04057	1980	1341.8427	638.15728	119.69025	112.9826	198	6.70765	6.77541
11	sale: PH NO	2178	1491.9594	686.04057	1980	1354.3143	625.68574	119.69025	113.1191	198	6.57111	6.63749
11	entra: PO2 NO	2376	1612.6344	763.36564	2178	1491.9594	686.04057	119.712018	119.6903	198	0.02177	0.02199
11	sale: PRE NO	2178	1491.9594	686.04057	1980	1339.1927	640.80732	119.69025	115.6747	198	4.01552	4.05609
11	entra: RACE1 NO	2376	1652.5221	723.4779	2178	1491.9594	686.04057	119.692336	119.6903	198	0.00209	0.00211

11	sale: RACE2 SI	2178	1491.9594	686.04057	1980	1306.7731	673.22688	119.69025	116.5029	198	3.18737	3.21957
12	sale: SEX SI	1980	1306.7731	673.22688	1782	1146.3263	635.67374	116.502876	114.8452	198	1.65771	1.67446
13	sale: AGE NO	1782	1146.3263	635.67374	1584	980.48069	603.51931	114.845163	94.06629	198	20.7789	20.9888
13	entra: BIC NO	1980	1284.3988	695.60124	1782	1146.3263	635.67374	114.846171	114.8452	198	0.00101	0.00102
13	sale: CAN NO	1782	1146.3263	635.67374	1584	991.86389	592.13611	114.845163	101.8516	198	12.9936	13.1249
13	entra: CPR NO	1980	1264.4779	715.52209	1782	1146.3263	635.67374	117.35216	114.8452	198	2.507	2.53232
13	entra: CRE NO	1980	1256.6444	723.35557	1782	1146.3263	635.67374	115.72927	114.8452	198	0.88411	0.89304
13	entra: CRN NO	1980	1267.4162	712.58378	1782	1146.3263	635.67374	115.20896	114.8452	198	0.3638	0.36747
13	entra: FRA NO	1980	1303.1484	676.85161	1782	1146.3263	635.67374	115.747918	114.8452	198	0.90276	0.91187
13	entra: HRA NO	1980	1303.8038	676.19625	1782	1146.3263	635.67374	114.845379	114.8452	198	0.00022	0.00022
13	entra: INF NO	1980	1309.047	670.95298	1782	1146.3263	635.67374	115.021138	114.8452	198	0.17597	0.17775
13	sale: LOC1 NO	1782	1146.3263	635.67374	1584	1137.7403	446.25966	114.845163	67.1716	198	47.6736	48.1551
13	sale: LOC2 NO	1782	1146.3263	635.67374	1584	1096.9927	487.00732	114.845163	76.32087	198	38.5243	38.9134
13	sale: PCO NO	1782	1146.3263	635.67374	1584	996.74579	587.25421	114.845163	107.5686	198	7.27661	7.35011
13	sale: PH NO	1782	1146.3263	635.67374	1584	1008.3642	575.63578	114.845163	107.9807	198	6.8645	6.93384
13	entra: PO2 NO	1980	1267.6272	712.37275	1782	1146.3263	635.67374	114.858976	114.8452	198	0.01381	0.01395
13	sale: PRE SI	1782	1146.3263	635.67374	1584	992.36477	591.63523	114.845163	112.114	198	2.73114	2.75873
14	entra: RACE1 NO	1782	1170.2852	611.71482	1584	992.36477	591.63523	113.085939	112.114	198	0.97192	0.98174
14	entra: RACE2 NO	1782	1179.2471	602.75288	1584	992.36477	591.63523	114.419801	112.114	198	2.30578	2.32907
14	entra: SER NO	1782	1141.5947	640.40527	1584	992.36477	591.63523	113.004695	112.114	198	0.89068	0.89967
14	sale: SYS NO	1584	992.36477	591.63523	1386	852.02741	533.97259	112.11402	103.3223	198	8.79176	8.88057
14	sale: TYP NO	1584	992.36477	591.63523	1386	809.95476	576.04524	112.11402	88.87852	198	23.2355	23.4702

Funciónes de Fisher y Tablas de Clasificaciones de cada paso

En cada paso se muestran las variables dentro del modelo con las estimaciones de los coeficientes de la función. Se usa p = 0.95, y corresponde F de la tabla = 3.8883.

Paso número 0 (con todas las variables)

AGE	-0.0451	CRE	-0.9311	INF	0.0484	PH	-1.7230	RACE2	1.4509	TYP	-2.1381
BIC	0.1399	CRN	-0.1370	LOC1	-9.2506	PO2	0.1060	SER	0.7072		
CAN	-2.7207	FRA	-0.9320	LOC2	-5.0987	PRE	-1.2511	SEX	0.6313		
CPR	-1.2960	HRA	0.0032	PCO	1.6750	RACE1	-0.0700	SYS	0.0153		

Tabla de Clasificaciones

	predicho 0	1	
original 0	146	14	160
original 1	15	25	40
	161	39	200

Cantidad de casos bien clasificados 171
Cantidad de casos mal clasificados 29

Porcentajes

	predicho 0	1
original 0	91.25	8.75
original 1	37.50	62.50

Porcentaje de casos bien clasificados 85.50
Porcentaje de casos mal clasificados 14.50

Paso número 1

La variable BIC queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0451	CRE	-0.9016	INF	0.0544	PH	-1.6779	RACE2	1.4616	TYP	-2.1400
		CRN	-0.1329	LOC1	-9.2596	PO2	0.1226	SER	0.6936		
CAN	-2.7249	FRA	-0.9413	LOC2	-5.0936	PRE	-1.2584	SEX	0.6319		
CPR	-1.3111	HRA	0.0032	PCO	1.6442	RACE1	-0.0613	SYS	0.0152		

Tabla de Clasificaciones

	predicho 0	1	
original 0	146	14	160
original 1	15	25	40
	161	39	200

Cantidad de casos bien clasificados 171
Cantidad de casos mal clasificados 29

Porcentajes

	predicho 0	1
original 0	91.25	8.75
original 1	37.50	62.50

Porcentaje de casos bien clasificados 85.50
Porcentaje de casos mal clasificados 14.50

Paso número 2

La variable CPR queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0445	CRE	-0.8490	INF	-5.4E-4	PH	-1.7342	RACE2	1.5626	TYP	-2.0889
		CRN	-0.2547	LOC1	-9.1629	PO2	0.3538	SER	0.8136		
CAN	-2.7278	FRA	-0.9924	LOC2	-5.5852	PRE	-1.1595	SEX	0.5732		
		HRA	0.0021	PCO	1.5844	RACE1	0.0686	SYS	0.0154		

Tabla de Clasificaciones

	predicho 0	1	
original 0	148	12	160
original 1	14	26	40
	162	38	200

Cantidad de casos bien clasificados 174
Cantidad de casos mal clasificados 26

Porcentajes

	predicho 0	1
original 0	92.50	7.50
original 1	35.00	65.00

Porcentaje de casos bien clasificados 87.00
Porcentaje de casos mal clasificados 13.00

Paso número 3

La variable CPRE queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0458			INF	-0.0064	PH	-1.8300	RACE2	1.5905	TYP	-2.0999
		CRN	-0.3996	LOC1	-9.0810	PO2	0.3555	SER	0.8347		
CAN	-2.7635	FRA	-1.0440	LOC2	-5.6674	PRE	-1.0899	SEX	0.5402		
		HRA	0.0016	PCO	1.7106	RACE1	0.1452	SYS	0.0158		

Tabla de Clasificaciones

	predicho 0	1	
original 0	146	14	160
original 1	13	27	40
	159	41	200

Cantidad de casos bien clasificados 173
Cantidad de casos mal clasificados 27

Porcentajes

	predicho 0	1
original 0	91.25	8.75
original 1	32.50	67.50

Porcentaje de casos bien clasificados 86.50
Porcentaje de casos mal clasificados 13.50

Paso número 4

La variable CRN queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0469			INF	-0.0225	PH	-1.9725	RACE2	1.6140	TYP	-2.1223
				LOC1	-9.0847	PO2	0.3536	SER	0.8535		
CAN	-2.7886	FRA	-1.0662	LOC2	-5.7805	PRE	-1.1060	SEX	0.5381		
		HRA	0.0020	PCO	1.8092	RACE1	0.2056	SYS	0.0152		

Tabla de Clasificaciones

	predicho 0	1	
original 0	146	14	160
original 1	14	26	40
	160	40	200

Cantidad de casos bien clasificados 172
Cantidad de casos mal clasificados 28

Porcentajes

	predicho 0	1
original 0	91.25	8.75
original 1	35.00	65.00

Porcentaje de casos bien clasificados 86.00
Porcentaje de casos mal clasificados 14.00

Paso número 5

La variable FRA queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0424			INF	-0.0478	PH	-2.0025	RACE2	1.4571	TYP	-2.2389
				LOC1	-8.9518	PO2	0.2726	SER	0.6404		
CAN	-2.6691			LOC2	-5.7757	PRE	-1.0873	SEX	0.5281		
		HRA	0.0022	PCO	1.7800	RACE1	0.0595	SYS	0.0156		

Tabla de Clasificaciones

	predicho 0	1	
original 0	145	15	160
original 1	12	28	40
	157	43	200

Cantidad de casos bien clasificados 173
Cantidad de casos mal clasificados 27

Porcentajes

	predicho 0	1
original 0	90.63	9.37
original 1	30.00	70.00

Porcentaje de casos bien clasificados 86.50
Porcentaje de casos mal clasificados 13.50

Paso número 6

La variable HRA queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0424			INF	-0.0202	PH	-1.9837	RACE2	1.4684	TYP	-2.2323
				LOC1	-8.9771	PO2	0.2886	SER	0.6122		
CAN	-2.6528			LOC2	-5.8027	PRE	-1.0820	SEX	0.5281		
				PCO	1.7935	RACE1	0.0661	SYS	0.0156		

Tabla de Clasificaciones

	predicho 0	1	
original 0	144	16	160
original 1	12	28	40
	156	44	200

Cantidad de casos bien clasificados 172
Cantidad de casos mal clasificados 28

Porcentajes

	predicho 0	1
original 0	90.00	10.00
original 1	30.00	70.00

Porcentaje de casos bien clasificados 86.00
Porcentaje de casos mal clasificados 14.00

Paso número 7

La variable INF queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0425					PH	-1.9878	RACE2	1.4698	TYP	-2.2347
				LOC1	-8.9830	PO2	0.2868	SER	0.6152		
CAN	-2.6552			LOC2	-5.8036	PRE	-1.0883	SEX	0.5289		
				PCO	1.7915	RACE1	0.0638	SYS	0.0157		

Tabla de Clasificaciones

	predicho 0	1	
original 0	144	16	160
original 1	12	28	40
	156	44	200

Cantidad de casos bien clasificados 172
Cantidad de casos mal clasificados 28

Porcentajes

	predicho 0	1
original 0	90.00	10.00
original 1	30.00	70.00

Porcentaje de casos bien clasificados 86.00
Porcentaje de casos mal clasificados 14.00

Paso número 8

La variable PO2 queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0420					PH	-1.9549	RACE2	1.4279	TYP	-2.2338
				LOC1	-8.9274			SER	0.5862		
CAN	-2.6565			LOC2	-5.8273	PRE	-1.0833	SEX	0.5114		
				PCO	1.8777	RACE1	0.0452	SYS	0.0154		

Tabla de Clasificaciones

	predicho 0	1	
original 0	143	17	160
original 1	12	28	40
	155	45	200

Cantidad de casos bien clasificados 171
Cantidad de casos mal clasificados 29

Porcentajes

	predicho 0	1
original 0	89.37	10.63
original 1	30.00	70.00

Porcentaje de casos bien clasificados 85.50
Porcentaje de casos mal clasificados 14.50

Paso número 9

La variable RACE1 queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0420					PH	-1.9528	RACE2	1.3868	TYP	-2.2342
				LOC1	-8.9340			SER	0.5860		
CAN	-2.6546			LOC2	-5.8261	PRE	-1.0860	SEX	0.5111		
				PCO	1.8780			SYS	0.0154		

Tabla de Clasificaciones

	predicho 0	1	
original 0	143	17	160
original 1	12	28	40
	155	45	200

Cantidad de casos bien clasificados 171
Cantidad de casos mal clasificados 29

Porcentajes

	predicho 0	1
original 0	89.37	10.63
original 1	30.00	70.00

Porcentaje de casos bien clasificados 85.50
Porcentaje de casos mal clasificados 14.50

Paso número 10

La variable SER queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0431					PH	-2.1303	RACE2	1.2334	TYP	-2.5591
				LOC1	-8.8185						
CAN	-2.6056			LOC2	-5.8326	PRE	-1.0354	SEX	0.5129		
				PCO	1.8085			SYS	0.0153		

Tabla de Clasificaciones

	predicho 0	1	
original 0	142	18	160
original 1	14	26	40
	156	44	200

Cantidad de casos bien clasificados 168
Cantidad de casos mal clasificados 32

Porcentajes

	predicho 0	1
original 0	88.75	11.25
original 1	35.00	65.00

Porcentaje de casos bien clasificados 84.00
Porcentaje de casos mal clasificados 16.00

Paso número 11

La variable RACE2 queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0449					PH	-2.2217			TYP	-2.4764
				LOC1	-8.5906						
CAN	-2.5111			LOC2	-5.8286	PRE	-0.9208	SEX	0.4844		
				PCO	1.8398			SYS	0.0159		

Tabla de Clasificaciones

	predicho 0	1	
original 0	141	19	160
original 1	15	25	40
	156	44	200

Cantidad de casos bien clasificados 166
Cantidad de casos mal clasificados 34

Porcentajes

	predicho 0	1
original 0	88.13	11.87
original 1	37.50	62.50

Porcentaje de casos bien clasificados 83.00
Porcentaje de casos mal clasificados 17.00

Paso número 12

La variable SEX queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0434					PH	-2.1707			TYP	-2.3578
				LOC1	-8.5446						
CAN	-2.3951			LOC2	-5.7406	PRE	-0.8411				
				PCO	1.8811			SYS	0.0164		

Tabla de Clasificaciones

	predicho 0	1	
original 0	148	12	160
original 1	15	25	40
	163	37	200

Cantidad de casos bien clasificados 173
Cantidad de casos mal clasificados 27

Porcentajes

	predicho 0	1
original 0	92.50	7.50
original 1	37.50	62.50

Porcentaje de casos bien clasificados 86.50
Porcentaje de casos mal clasificados 13.50

Paso número 13

La variable PRE queda fuera del modelo y se muestra en blanco su posición:

AGE	-0.0440					PH	-2.1138			TYP	-2.2630
				LOC1	-8.4935						
CAN	-2.1980			LOC2	-5.5430						
				PCO	1.9156			SYS	0.0168		

Tabla de Clasificaciones

	predicho 0	1	
original 0	143	17	160
original 1	14	26	40
	157	43	200

Cantidad de casos bien clasificados 169
Cantidad de casos mal clasificados 31

Porcentajes

	predicho 0	1
original 0	89.37	10.63
original 1	35.00	65.00

Porcentaje de casos bien clasificados 84.50
Porcentaje de casos mal clasificados 15.50

Regresión Logística

Prueba con todas las variables

Deviance = -2 Log likelihood = 112.174.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.127	8	.332

Dado que sig = 0.332 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			STA		Percentage Correct
			0	1	
Step 1	STA	0	156	4	97.5
		1	19	21	52.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.056	.018	9.331	1	.002	1.058
	BIC(1)	.262	.897	.086	1	.770	1.300
	CAN(1)	-3.483	1.121	9.650	1	.002	.031
	CPR(1)	-1.032	.990	1.087	1	.297	.356
	CRE(1)	-.100	1.131	.008	1	.929	.904
	CRN(1)	-.119	.845	.020	1	.888	.888
	FRA(1)	-1.649	1.093	2.277	1	.131	.192
	HRA	-.003	.010	.080	1	.778	.997
	INF(1)	.108	.556	.038	1	.846	1.114
	LOC1(1)	-19.118	39.474	.235	1	.628	.000
	LOC2(1)	-3.458	1.341	6.646	1	.010	.031
	PCO(1)	2.084	1.165	3.201	1	.074	8.033
	PH(1)	-1.771	1.212	2.134	1	.144	.170
	PO2(1)	.677	.940	.518	1	.472	1.967
	PRE(1)	-1.279	.702	3.321	1	.068	.278
	RACE1(1)	.583	1.313	.197	1	.657	1.791
	RACE2(1)	7.438	20.543	.131	1	.717	1698.667
	SER(1)	.674	.629	1.148	1	.284	1.962
	SEX(1)	.721	.546	1.746	1	.186	2.057
	SYS	-.021	.009	4.871	1	.027	.979
	TYP(1)	-3.748	1.342	7.798	1	.005	.024
	Constant	18.247	33.993	.288	1	.591	8.4E+07

a. Variable(s) entered on step 1: AGE, BIC, CAN, CPR, CRE, CRN, FRA, HRA, INF, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "CRE" tiene nivel de significación sig = 0.929 > 0.05, entonces se considera no significativa para el modelo.

Prueba dejando la variable "CRE" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.213	8	.325

Dado que sig = 0.325 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			STA		Percentage
	Observed		0	1	Correct
Step 1	STA	0	156	4	97.5
		1	19	21	52.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.057	.018	9.733	1	.002	1.058
	BIC(1)	.249	.884	.080	1	.778	1.283
	CAN(1)	-3.489	1.117	9.758	1	.002	.031
	CPR(1)	-1.022	.982	1.082	1	.298	.360
	CRN(1)	-.139	.815	.029	1	.865	.871
	FRA(1)	-1.657	1.091	2.307	1	.129	.191
	HRA	-.003	.010	.074	1	.785	.997
	INF(1)	.112	.554	.041	1	.839	1.119
	LOC1(1)	-19.132	39.408	.236	1	.627	.000
	LOC2(1)	-3.476	1.328	6.849	1	.009	.031
	PCO(1)	2.103	1.145	3.373	1	.066	8.191
	PH(1)	-1.790	1.195	2.245	1	.134	.167
	PO2(1)	.694	.918	.572	1	.450	2.002
	PRE(1)	-1.272	.698	3.324	1	.068	.280
	RACE1(1)	.615	1.261	.238	1	.625	1.850
	RACE2(1)	7.466	20.528	.132	1	.716	1748.131
	SER(1)	.678	.627	1.168	1	.280	1.970
	SEX(1)	.719	.545	1.739	1	.187	2.052
	SYS	-.021	.009	5.099	1	.024	.979
	TYP(1)	-3.754	1.341	7.837	1	.005	.023
	Constant	18.148	33.907	.286	1	.592	7.6E+07

a. Variable(s) entered on step 1: AGE, BIC, CAN, CPR, CRN, FRA, HRA, INF, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "CRN" tiene nivel de significación sig = 0.865 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	21	178	112.174
Reducido	20	179	112.182

$\chi^2_{1,\alpha}$	α
0.008	0.92873

0.92873 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba dejando la variable "CRN" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.299	8	.318

Dado que sig = 0.318 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1 STA	0	156	4	97.5
	1	19	21	52.5
Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.057	.018	9.975	1	.002	1.059
	BIC(1)	.238	.879	.073	1	.787	1.269
	CAN(1)	-3.511	1.113	9.957	1	.002	.030
	CPR(1)	-1.024	.979	1.093	1	.296	.359
	FRA(1)	-1.656	1.092	2.302	1	.129	.191
	HRA	-.003	.010	.083	1	.773	.997
	INF(1)	.101	.549	.034	1	.854	1.106
	LOC1(1)	-19.086	39.357	.235	1	.628	.000
	LOC2(1)	-3.493	1.320	6.998	1	.008	.030
	PCO(1)	2.135	1.129	3.576	1	.059	8.461
	PH(1)	-1.853	1.136	2.661	1	.103	.157
	PO2(1)	.688	.917	.563	1	.453	1.990
	PRE(1)	-1.289	.692	3.471	1	.062	.276
	RACE1(1)	.632	1.254	.254	1	.614	1.882
	RACE2(1)	7.478	20.505	.133	1	.715	1768.418
	SER(1)	.691	.624	1.229	1	.268	1.996
	SEX(1)	.723	.545	1.761	1	.185	2.060
	SYS	-.021	.009	5.129	1	.024	.979
	TYP(1)	-3.771	1.339	7.926	1	.005	.023
	Constant	18.021	33.857	.283	1	.595	6.7E+07

a. Variable(s) entered on step 1: AGE, BIC, CAN, CPR, FRA, HRA, INF, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "INF" tiene nivel de significación sig = 0.854 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	20	179	112.182
Reducido	19	180	112.210

$\chi^2_{1,\alpha}$	α
0.028	0.86711

0.86711 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	20	179	Ver cuadro siguiente
Reducido	19	180	112.210

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.896>0.05	112.194	0.016	0.89934

α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variable CRE sigue fuera del modelo.

Prueba dejando la variable "INF" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.312	8	.317

Dado que sig = 0.317 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	156	4	97.5
		1	18	22	55.0
	Overall Percentage				89.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.056	.018	10.205	1	.001	1.058
	BIC(1)	.254	.875	.084	1	.772	1.289
	CAN(1)	-3.504	1.111	9.942	1	.002	.030
	CPR(1)	-.993	.964	1.059	1	.303	.371
	FRA(1)	-1.651	1.091	2.287	1	.130	.192
	HRA	-.003	.010	.102	1	.750	.997
	LOC1(1)	-19.053	39.235	.236	1	.627	.000
	LOC2(1)	-3.454	1.295	7.114	1	.008	.032
	PCO(1)	2.105	1.113	3.576	1	.059	8.205
	PH(1)	-1.819	1.117	2.649	1	.104	.162
	PO2(1)	.696	.917	.576	1	.448	2.005
	PRE(1)	-1.265	.678	3.480	1	.062	.282
	RACE1(1)	.638	1.252	.260	1	.610	1.893
	RACE2(1)	7.490	20.440	.134	1	.714	1789.730
	SER(1)	.688	.622	1.224	1	.269	1.991
	SEX(1)	.711	.540	1.732	1	.188	2.037
	SYS	-.020	.009	5.267	1	.022	.980
	TYP(1)	-3.762	1.341	7.865	1	.005	.023
	Constant	17.930	33.754	.282	1	.595	6.1E+07

a. Variable(s) entered on step 1: AGE, BIC, CAN, CPR, FRA, HRA, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "BIC" tiene nivel de significación sig = 0.772, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	19	180	112.210
Reducido	18	181	112.244

$\chi^2_{1,\alpha}$	α
0.034	0.85371

0.85371 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	19	180	Ver cuadro siguiente
Reducido	18	181	112.244

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.889>0.05	112.225	0.019	0.89037
CRN	0.884>0.05	112.223	0.021	0.88478

En todos los casos α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : las variables CRE y CRN siguen fuera del modelo.

Prueba dejando la variable "BIC" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.164	8	.329

Dado que sig = 0.329 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	156	4	97.5
		1	19	21	52.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.056	.017	10.216	1	.001	1.057
	CAN(1)	-3.519	1.107	10.107	1	.001	.030
	CPR(1)	-1.018	.951	1.147	1	.284	.361
	FRA(1)	-1.671	1.085	2.369	1	.124	.188
	HRA	-.003	.010	.094	1	.759	.997
	LOC1(1)	-19.061	39.190	.237	1	.627	.000
	LOC2(1)	-3.375	1.249	7.303	1	.007	.034
	PCO(1)	2.001	1.046	3.657	1	.056	7.398
	PH(1)	-1.683	1.009	2.785	1	.095	.186
	PO2(1)	.740	.900	.676	1	.411	2.096
	PRE(1)	-1.291	.673	3.674	1	.055	.275
	RACE1(1)	.649	1.250	.270	1	.603	1.914
	RACE2(1)	7.506	20.444	.135	1	.714	1818.162
	SER(1)	.646	.605	1.138	1	.286	1.907
	SEX(1)	.703	.539	1.699	1	.192	2.019
	SYS	-.020	.009	5.162	1	.023	.980
	TYP(1)	-3.769	1.341	7.902	1	.005	.023
	Constant	18.102	33.697	.289	1	.591	7.3E+07

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, HRA, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "HRA" tiene nivel de significación sig = 0.759, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	18	181	112.244
Reducido	17	182	112.329

$\chi^2_{1,\alpha}$	α
0.085	0.77063

0.77063 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	18	181	Ver cuadro siguiente
Reducido	17	182	112.329

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.929>0.05	112.322	0.007	0.93332
CRN	0.904>0.05	112.315	0.014	0.90581
INF	0.832>0.05	112.284	0.045	0.83200

En todos los casos α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN e INF siguen fuera del modelo.

Prueba dejando la variable "HRA" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.070	8	.529

Dado que sig = 0.529 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	156	4	97.5
		1	19	21	52.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.055	.017	10.246	1	.001	1.057
	CAN(1)	-3.529	1.110	10.098	1	.001	.029
	CPR(1)	-.983	.946	1.080	1	.299	.374
	FRA(1)	-1.688	1.081	2.435	1	.119	.185
	LOC1(1)	-21.225	64.839	.107	1	.743	.000
	LOC2(1)	-3.454	1.221	8.004	1	.005	.032
	PCO(1)	2.037	1.035	3.871	1	.049	7.666
	PH(1)	-1.683	1.007	2.794	1	.095	.186
	PO2(1)	.737	.892	.683	1	.408	2.090
	PRE(1)	-1.289	.674	3.662	1	.056	.275
	RACE1(1)	.669	1.246	.288	1	.591	1.952
	RACE2(1)	8.593	33.544	.066	1	.798	5391.192
	SER(1)	.610	.594	1.054	1	.305	1.840
	SEX(1)	.712	.539	1.746	1	.186	2.038
	SYS	-.020	.009	5.140	1	.023	.980
	TYP(1)	-3.764	1.349	7.781	1	.005	.023
	Constant	18.920	55.633	.116	1	.734	1.6E+08

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, RACE2, SER, SEX, SYS, TYP.

La variable "RACE2" tiene nivel de significación sig = 0.798, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	17	182	112.329
Reducido	16	183	112.420

$\chi^2_{1,\alpha}$	α
0.091	0.76291

0.76291 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	17	182	Ver cuadro siguiente
Reducido	16	183	112.420

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.953	112.421		
CRN	0.885	112.399	0.021	0.88478
INF	0.800	112.360	0.060	0.80650
BIC	0.782	112.343	0.077	0.78140

En la variable CRE: M. C. Deviance = 112.421 > 112.420 al ingresar la variable crece la deviance entonces la variable no ayuda al ajuste del modelo, por lo tanto queda fuera del mismo. En las variables CRN, INF y BIC α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo, quedan fuera.

Prueba dejando la variable "RACE2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.084	8	.335

Dado que sig = 0.335 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	157	3	98.1
		1	19	21	52.5
	Overall Percentage				89.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.054	.017	9.827	1	.002	1.056
	CAN(1)	-3.387	1.090	9.650	1	.002	.034
	CPR(1)	-1.063	.935	1.292	1	.256	.346
	FRA(1)	-1.420	1.018	1.947	1	.163	.242
	LOC1(1)	-13.254	22.799	.338	1	.561	.000
	LOC2(1)	-3.336	1.190	7.866	1	.005	.036
	PCO(1)	2.023	1.025	3.892	1	.049	7.560
	PH(1)	-1.633	.990	2.718	1	.099	.195
	PO2(1)	.527	.889	.352	1	.553	1.694
	PRE(1)	-1.172	.658	3.177	1	.075	.310
	RACE1(1)	-.955	1.149	.692	1	.406	.385
	SER(1)	.474	.570	.693	1	.405	1.607
	SEX(1)	.647	.533	1.473	1	.225	1.909
	SYS	-.019	.008	4.889	1	.027	.981
	TYP(1)	-3.769	1.345	7.854	1	.005	.023
	Constant	19.249	22.990	.701	1	.402	2.3E+08

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC1, LOC2, PCO, PH, PO2, PRE, RACE1, SER, SEX, SYS, TYP.

La variable "PO2" tiene nivel de significación sig = 0.553, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	16	183	112.420
Reducido	15	184	115.469

$\chi^2_{1,\alpha}$	α
3.049	0.08079

0.08079 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	16	183	Ver cuadro siguiente
Reducido	15	184	115.469

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.771>0.05	115.385	0.084	0.77195
CRN	0.853>0.05	115.435	0.034	0.85371
INF	0.732>0.05	115.351	0.118	0.73121
BIC	0.756>0.05	115.371	0.098	0.75424
HRA	0.621>0.05	115.221	0.248	0.61849

En todos las pruebas α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC y HRA siguen fuera del modelo.

Prueba dejando la variable "PO2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.570	8	.297

Dado que sig = 0.297 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	158	2	98.8
		1	18	22	55.0
	Overall Percentage				90.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.052	.017	9.559	1	.002	1.054
	CAN(1)	-3.334	1.092	9.328	1	.002	.036
	CPR(1)	-1.179	.910	1.675	1	.196	.308
	FRA(1)	-1.347	1.019	1.750	1	.186	.260
	LOC1(1)	-12.927	23.236	.310	1	.578	.000
	LOC2(1)	-3.289	1.181	7.762	1	.005	.037
	PCO(1)	2.121	1.017	4.349	1	.037	8.338
	PH(1)	-1.557	.974	2.554	1	.110	.211
	PRE(1)	-1.165	.656	3.153	1	.076	.312
	RACE1(1)	-.993	1.146	.751	1	.386	.370
	SER(1)	.409	.561	.532	1	.466	1.506
	SEX(1)	.605	.526	1.325	1	.250	1.832
	SYS	-.018	.008	4.599	1	.032	.983
	TYP(1)	-3.728	1.342	7.723	1	.005	.024
	Constant	19.211	23.431	.672	1	.412	2.2E+08

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC1, LOC2, PCO, PH, PRE, RACE1, SER, SEX, SYS, TYP.

La variable "LOC1" tiene nivel de significación sig = 0.578, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	15	184	115.469
Reducido	14	185	115.832

$\chi^2_{1,\alpha}$	α
0.363	0.54684

0.54684 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	15	184	Ver cuadro siguiente
Reducido	14	185	115.832

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.712>0.05	115.698	0.134	0.71432
CRN	0.860>0.05	115.801	0.031	0.86024
INF	0.700>0.05	115.683	0.149	0.69949
BIC	0.672>0.05	115.649	0.183	0.66881
HRA	0.609>0.05	115.566	0.266	0.60603
RACE2	0.723>0.05	113.139	2.693	0.10079

En todas las pruebas, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.
Decisión : Las variables CRE, CRN, INF, BIC, HRA y RACE2 siguen fuera del modelo.

Prueba dejando la variable "LOC1" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.109	8	.108

Dado que sig = 0.108 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted STA 0	Predicted STA 1	Percentage Correct
Step 1	STA	0	158	2	98.8
		1	26	14	35.0
	Overall Percentage				86.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.045	.014	9.873	1	.002	1.046
	CAN(1)	-2.364	.899	6.918	1	.009	.094
	CPR(1)	-.900	.823	1.194	1	.275	.407
	FRA(1)	-.795	.923	.742	1	.389	.451
	LOC2(1)	-2.985	1.060	7.922	1	.005	.051
	PCO(1)	1.821	.912	3.991	1	.046	6.178
	PH(1)	-1.351	.884	2.336	1	.126	.259
	PRE(1)	-.828	.571	2.104	1	.147	.437
	RACE1(1)	-.010	.736	.000	1	.989	.990
	SER(1)	.232	.494	.220	1	.639	1.261
	SEX(1)	.487	.459	1.125	1	.289	1.627
	SYS	-.009	.007	1.913	1	.167	.991
	TYP(1)	-2.739	.958	8.181	1	.004	.065
	Constant	3.867	2.096	3.405	1	.065	47.816

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC2, PCO, PH, PRE, RACE1, SER, SEX, SYS, TYP.

La variable "RACE1" tiene nivel de significación sig = 0.989 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	14	185	115.832
Reducido	13	186	144.955

$\chi^2_{1,\alpha}$	α
29.123	0.00000

0.00000 < 0.05 indica que las variables que no se incluyeron en el modelo aportan una vez que las otras variables han sido incluidas en el modelo.

El incremento notorio en la deviance influye en la clasificación de los casos de la clase 2, produciendo una baja en el porcentaje de casos bien clasificados para esa clase, pasando de 55% a 35%.

Decisión : que la variable "LOC1" siga formando parte del modelo y se vuelve al paso anterior, en donde la variable "SER" tiene un nivel de significación sig=0.466>0.05.

Prueba dejando la variable "SER" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.270	8	.140

Dado que sig = 0.140 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			Vital Status		Percentage Correct
	Observed		Lived	Died	
Step 1	Vital Status	Lived	156	4	97.5
		Died	19	21	52.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.052	.017	9.618	1	.002	1.053
	CAN(1)	-3.259	1.072	9.234	1	.002	.038
	CPR(1)	-1.322	.886	2.224	1	.136	.267
	FRA(1)	-1.147	.991	1.340	1	.247	.317
	LOC1(1)	-12.981	23.052	.317	1	.573	.000
	LOC2(1)	-3.198	1.152	7.702	1	.006	.041
	PCO(1)	2.084	1.017	4.196	1	.041	8.035
	PH(1)	-1.641	.974	2.841	1	.092	.194
	PRE(1)	-1.119	.653	2.936	1	.087	.327
	RACE1(1)	-.952	1.138	.700	1	.403	.386
	SEX(1)	.656	.523	1.571	1	.210	1.927
	SYS	-.018	.008	4.840	1	.028	.982
	TYP(1)	-3.931	1.303	9.107	1	.003	.020
	Constant	19.431	23.246	.699	1	.403	2.7E+08

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC1, LOC2, PCO, PH, PRE, RACE1, SEX, SYS, TYP.

La variable "RACE1" tiene nivel de significación sig = 0.403 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	14	185	115.832
Reducido	13	186	116.372

$\chi^2_{1,\alpha}$	α
0.54	0.46243

0.46243 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	14	185	Ver cuadro siguiente
Reducido	13	186	116.372

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.633>0.05	116.148	0.224	0.63601
CRN	0.800>0.05	116.309	0.063	0.80182
INF	0.758>0.05	116.277	0.095	0.75791
BIC	0.844>0.05	116.333	0.039	0.84345
HRA	0.737>0.05	116.259	0.113	0.73675
RACE2	0.729>0.05	113.878	2.494	0.11428
PO2	0.659>0.05	116.172	0.200	0.65472

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y En todos los casos α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2 y PO2 siguen fuera del modelo.

Prueba dejando la variable "RACE1" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.510	8	.301

Dado que sig = 0.301 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	157	3	98.1
		Died	19	21	52.5
	Overall Percentage				89.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.055	.016	11.171	1	.001	1.057
	CAN(1)	-3.378	1.075	9.878	1	.002	.034
	CPR(1)	-1.270	.870	2.131	1	.144	.281
	FRA(1)	-1.168	.994	1.382	1	.240	.311
	LOC1(1)	-12.718	23.068	.304	1	.581	.000
	LOC2(1)	-3.284	1.155	8.085	1	.004	.037
	PCO(1)	2.112	1.024	4.252	1	.039	8.263
	PH(1)	-1.732	.977	3.143	1	.076	.177
	PRE(1)	-.988	.629	2.469	1	.116	.372
	SEX(1)	.687	.520	1.745	1	.187	1.988
	SYS	-.019	.008	5.398	1	.020	.981
	TYP(1)	-4.003	1.312	9.307	1	.002	.018
	Constant	19.111	23.259	.675	1	.411	2.0E+08

a. Variable(s) entered on step 1: AGE, CAN, CPR, FRA, LOC1, LOC2, PCO, PH, PRE, SEX, SYS, TYP.

La variable "FRA" tiene nivel de significación sig = 0.240 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	13	186	116.372
Reducido	12	187	117.210

$\chi^2_{1,\alpha}$	α
0.838	0.35997

0.35997 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	13	186	Ver cuadro siguiente
Reducido	12	187	117.210

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.767>0.05	117.124	0.086	0.76933
CRN	0.867>0.05	117.182	0.028	0.86711
INF	0.757>0.05	117.114	0.096	0.75668
BIC	0.838>0.05	117.168	0.042	0.83762
HRA	0.668>0.05	117.024	0.186	0.66627
RACE2	0.741>0.05	113.969	3.514	0.06085
PO2	0.618>0.05	116.953	0.257	0.61219
SER	0.494>0.05	116.736	0.474	0.49115

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa.
En la variable RACE2, α = 0.06085 > 0.05 pero está cerca de 0.05. Dado que el Nivel de Significación = 0.741 > 0.05, se considera no significativa.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2 y SER siguen fuera del modelo.

Prueba dejando la variable "FRA" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.581	8	.093

Dado que sig = 0.093 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	156	4	97.5
		Died	20	20	50.0
	Overall Percentage				88.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.050	.015	10.373	1	.001	1.051
	CAN(1)	-3.192	1.044	9.350	1	.002	.041
	CPR(1)	-1.131	.838	1.822	1	.177	.323
	LOC1(1)	-12.570	23.048	.297	1	.585	.000
	LOC2(1)	-3.140	1.114	7.946	1	.005	.043
	PCO(1)	1.937	.968	4.002	1	.045	6.940
	PH(1)	-1.723	.949	3.297	1	.069	.178
	PRE(1)	-.919	.624	2.170	1	.141	.399
	SEX(1)	.654	.512	1.630	1	.202	1.922
	SYS	-.019	.008	5.527	1	.019	.981
	TYP(1)	-3.953	1.297	9.296	1	.002	.019
	Constant	17.868	23.204	.593	1	.441	5.8E+07

a. Variable(s) entered on step 1: AGE, CAN, CPR, LOC1, LOC2, PCO, PH, PRE, SEX, SYS, TYP.

La variable "SEX" tiene nivel de significación sig = 0.202, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	12	187	117.210
Reducido	11	188	118.468

$\chi^2_{1,\alpha}$	α
1.258	0.26203

0.26203 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	12	187	Ver cuadro siguiente
Reducido	11	188	118.468

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.799>0.05	118.404	0.064	0.80028
CRN	0.917>0.05	118.457	0.011	0.91647
INF	0.774>0.05	118.384	0.084	0.77195
BIC	0.743>0.05	118.359	0.109	0.74129
HRA	0.622>0.05	118.221	0.247	0.61920
RACE2	0.744>0.05	115.316	3.152	0.07583
PO2	0.677>0.05	118.290	0.178	0.67310
SER	0.679>0.05	118.295	0.173	0.67746
RACE1	0.395>0.05	117.596	0.872	0.35040

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa. En la variable RACE2, α = 0.07583> 0.05 cercano a 0.05, para las demás α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER y RACE1 siguen fuera del modelo.

Prueba dejando la variable "SEX" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.360	8	.313

Dado que sig = 0.313 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			Vital Status		Percentage
	Observed		Lived	Died	Correct
Step 1	Vital Status	Lived	155	5	96.9
		Died	20	20	50.0
	Overall Percentage				87.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.046	.015	9.741	1	.002	1.047
	CAN(1)	-2.853	.981	8.462	1	.004	.058
	CPR(1)	-1.055	.828	1.624	1	.203	.348
	LOC1(1)	-12.490	23.028	.294	1	.588	.000
	LOC2(1)	-3.088	1.095	7.950	1	.005	.046
	PCO(1)	1.947	.973	4.003	1	.045	7.008
	PH(1)	-1.630	.918	3.153	1	.076	.196
	PRE(1)	-.866	.620	1.953	1	.162	.421
	SYS	-.020	.008	6.209	1	.013	.981
	TYP(1)	-3.719	1.257	8.746	1	.003	.024
	Constant	17.936	23.179	.599	1	.439	6.2E+07

a. Variable(s) entered on step 1: AGE, CAN, CPR, LOC1, LOC2, PCO, PH, PRE, SYS, TYP.

La variable "CPR" tiene nivel de significación sig = 0.203 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	11	188	118.468
Reducido	10	189	120.162

$\chi^2_{1,\alpha}$	α
1.694	0.19307

0.19307 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	11	188	Ver cuadro siguiente
Reducido	10	189	120.162

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.898>0.05	120.146	0.016	0.89934
CRN	0.826>0.05	120.115	0.047	0.82837
INF	0.910>0.05	120.150	0.012	0.91277
BIC	0.829>0.05	120.115	0.047	0.82837
HRA	0.658>0.05	119.964	0.198	0.65634
RACE2	0.743>0.05	116.944	3.218	0.07283
PO2	0.808>0.05	120.103	0.059	0.80808
SER	0.585>0.05	119.861	0.301	0.58326
RACE1	0.347>0.05	119.076	1.086	0.29736
FRA	0.266>0.05	119.034	1.128	0.28820

En la variable RACE2, α = 0.07283 > 0.05 se puede considerar cerceno a 0.05, pero el Nivel de Significación =0.743 > 0.05 indica que la estimación del coeficiente no es significativa. En las otras pruebas, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER, RACE1 y FRA siguen fuera del modelo.

Prueba dejando la variable "CPR" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.075	8	.426

Dado que sig = 0.426 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	156	4	97.5
		Died	21	19	47.5
	Overall Percentage				87.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.044	.014	9.439	1	.002	1.045
	CAN(1)	-2.922	1.001	8.524	1	.004	.054
	LOC1(1)	-12.462	22.911	.296	1	.586	.000
	LOC2(1)	-3.320	1.025	10.491	1	.001	.036
	PCO(1)	1.780	.922	3.728	1	.054	5.928
	PH(1)	-1.689	.898	3.539	1	.060	.185
	PRE(1)	-.791	.613	1.661	1	.198	.454
	SYS	-.019	.008	6.038	1	.014	.981
	TYP(1)	-3.835	1.274	9.057	1	.003	.022
	Constant	17.437	23.059	.572	1	.450	3.7E+07

a. Variable(s) entered on step 1: AGE, CAN, LOC1, LOC2, PCO, PH, PRE, SYS, TYP.

La variable "PRE" tiene nivel de significación sig = 0.198 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	10	189	120.162
Reducido	9	190	121.706

$\chi^2_{1,\alpha}$	α
1.544	0.21402

0.21402 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	10	189	Ver cuadro siguiente
Reducido	9	190	121.706

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.912>0.05	121.694	0.012	0.91277
CRN	0.734>0.05	121.592	0.114	0.73564
INF	0.942>0.05	121.701	0.005	0.94363
BIC	0.724>0.05	121.580	0.126	0.72262
HRA	0.820>0.05	121.654	0.052	0.81962
RACE2	0.744>0.05	118.533	3.173	0.07486
PO2	0.645>0.05	121.489	0.217	0.64134
SER	0.421>0.05	121.046	0.660	0.41656
RACE1	0.376>0.05	120.753	0.953	0.32896
FRA	0.343>0.05	120.875	0.831	0.36198
SEX	0.230>0.05	120.215	1.491	0.22206

En la variable RACE2, α = 0.07486 > 0.05 se puede considerar cerceno a 0.05, pero el Nivel de Significación = 0.744 > 0.05 indica que la estimación del coeficiente no es significativa. En las otras pruebas, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER, RACE1, FRA y SEX siguen fuera del modelo.

Prueba dejando la variable "PRE" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.889	8	.116

Dado que sig = 0.116 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	158	2	98.8
		Died	21	19	47.5
	Overall Percentage				88.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.044	.014	10.004	1	.002	1.045
	CAN(1)	-2.751	.990	7.721	1	.005	.064
	LOC1(1)	-12.286	23.151	.282	1	.596	.000
	LOC2(1)	-3.157	1.012	9.736	1	.002	.043
	PCO(1)	1.777	.918	3.751	1	.053	5.913
	PH(1)	-1.592	.867	3.375	1	.066	.203
	SYS	-.019	.008	5.973	1	.015	.981
	TYP(1)	-3.798	1.285	8.732	1	.003	.022
	Constant	16.189	23.271	.484	1	.487	1.1E+07

a. Variable(s) entered on step 1: AGE, CAN, LOC1, LOC2, PCO, PH, SYS, TYP.

La variable "PH" tiene nivel de significación sig = 0.066 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	190	121.706
Reducido	8	191	123.288

$\chi^2_{1,\alpha}$	α
1.582	0.20847

0.20847 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	190	Ver cuadro siguiente
Reducido	8	191	123.288

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.973>0.05	123.287	0.001	0.97477
CRN	0.660>0.05	123.098	0.190	0.66292
INF	0.750>0.05	123.187	0.101	0.75063
BIC	0.620>0.05	123.036	0.252	0.61567
HRA	0.861>0.05	123.257	0.031	0.86024
RACE2	0.761>0.05	120.773	2.515	0.11277
PO2	0.660>0.05	123.089	0.199	0.65553
SER	0.491>0.05	122.806	0.482	0.48752
RACE1	0.517>0.05	122.799	0.489	0.48437
FRA	0.398>0.05	122.623	0.665	0.41480
SEX	0.255>0.05	121.951	1.337	0.24756
CPR	0.249>0.05	122.020	1.268	0.26014

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER, RACE1, FRA, SEX y CPR siguen fuera del modelo.

Prueba dejando la variable "PH" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	12.115	8	.146

Dado que sig = 0.146 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			Vital Status		Percentage
	Observed		Lived	Died	Correct
Step 1	Vital Status	Lived	156	4	97.5
		Died	23	17	42.5
	Overall Percentage				86.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.041	.013	9.684	1	.002	1.042
	CAN(1)	-2.768	1.027	7.264	1	.007	.063
	LOC1(1)	-12.270	22.921	.287	1	.592	.000
	LOC2(1)	-2.706	.927	8.527	1	.003	.067
	PCO(1)	1.030	.799	1.659	1	.198	2.800
	SYS	-.019	.008	6.081	1	.014	.982
	TYP(1)	-3.952	1.308	9.131	1	.003	.019
	Constant	15.158	23.038	.433	1	.511	3828063

a. Variable(s) entered on step 1: AGE, CAN, LOC1, LOC2, PCO, SYS, TYP.

La variable "PCO" tiene nivel de significación sig = 0.198 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	191	123.288
Reducido	7	192	126.575

$\chi^2_{1,\alpha}$	α
3.287	0.06983

0.06983 se puede considerar cercano a 0.05, pero dado Nivel de Significación > 0.05 se toma la decisión que la variable PH quede fuera del modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	191	Ver cuadro siguiente
Reducido	7	192	126.575

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.644>0.05	126.366	0.209	0.64755
CRN	0.296>0.05	125.518	1.057	0.30390
INF	0.512>0.05	126.146	0.429	0.51248
BIC	0.740>0.05	126.467	0.108	0.74243
HRA	0.999>0.05	126.571	0.004	0.94957
RACE2	0.756>0.05	123.683	2.892	0.08902
PO2	0.754>0.05	126.475	0.100	0.75183
SER	0.369>0.05	125.478	0.827	0.36314
RACE1	0.425>0.05	125.802	0.773	0.37929
FRA	0.376>0.05	125.852	0.723	0.39516
SEX	0.277>0.05	125.357	1.218	0.26975
CPR	0.195>0.05	124.982	1.593	0.20690
PRE	0.224>0.05	125.167	1.408	0.23539

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER, RACE1, FRA, SEX, CPR y PRE siguen fuera del modelo.

Prueba dejando la variable "PCO" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.985	8	.649

Dado que sig = 0.649 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			Vital Status		Percentage
	Observed		Lived	Died	Correct
Step 1	Vital Status	Lived	157	3	98.1
		Died	23	17	42.5
	Overall Percentage				87.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.038	.013	8.771	1	.003	1.039
	CAN(1)	-2.597	.963	7.283	1	.007	.074
	LOC1(1)	-12.240	22.984	.284	1	.594	.000
	LOC2(1)	-2.417	.874	7.642	1	.006	.089
	SYS	-.018	.007	5.674	1	.017	.982
	TYP(1)	-3.848	1.269	9.188	1	.002	.021
	Constant	15.669	23.091	.460	1	.497	6382065

a. Variable(s) entered on step 1: AGE, CAN, LOC1, LOC2, SYS, TYP.

La variable "LOC1" tiene nivel de significación sig = 0.594 > 0.05, entonces se considera no significativa para el modelo. Fue evaluada en un paso anterior.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	191	126.575
Reducido	6	192	128.444

$\chi^2_{1,\alpha}$	α
1.869	0.17159

0.17159 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	192	Ver cuadro siguiente
Reducido	6	193	128.444

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
CRE	0.509>0.05	128.018	0.426	0.51396
CRN	0.264>0.05	127.232	1.212	0.27094
INF	0.592>0.05	128.157	0.287	0.59215
BIC	0.638>0.05	128.228	0.216	0.64210
HRA	0.729>0.05	128.323	0.121	0.72795
RACE2	0.758>0.05	125.626	2.818	0.09321
PO2	0.498>0.05	127.960	0.484	0.48662
SER	0.464>0.05	127.898	0.546	0.45996
RACE1	0.448>0.05	127.747	0.697	0.40379
FRA	0.489>0.05	127.990	0.454	0.50044
SEX	0.257>0.05	127.113	1.331	0.24863
CPR	0.275>0.05	127.310	1.134	0.28692
PRE	0.208>0.05	126.937	1.507	0.21960
PH	0.352>0.05	127.620	0.824	0.36401

En cada prueba, el Nivel de Significación > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : Las variables CRE, CRN, INF, BIC, HRA, RACE2, PO2, SER, RACE1, FRA, SEX, CPR, PRE y PH siguen fuera del modelo.

Prueba dejando la variable "LOC1" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.655	8	.372

Dado que sig = 0.372 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Vital Status		Percentage Correct
			Lived	Died	
Step 1	Vital Status	Lived	158	2	98.8
		Died	27	13	32.5
	Overall Percentage				85.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.034	.012	8.836	1	.003	1.035
	CAN(1)	-1.804	.819	4.855	1	.028	.165
	LOC2(1)	-2.305	.851	7.343	1	.007	.100
	SYS	-.010	.006	2.298	1	.130	.990
	TYP(1)	-2.805	.897	9.781	1	.002	.060
	Constant	1.987	1.503	1.748	1	.186	7.291

a. Variable(s) entered on step 1: AGE, CAN, LOC2, SYS, TYP.

La variable "SYS" tiene nivel de significación sig = 0.130 > 0.05, entonces se considera no significativa para el modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	192	128.444
Reducido	5	193	154.352

$\chi^2_{1,\alpha}$	α
25.908	0.00000

0.00000 < 0.05 indica que las variables que no se incluyeron en el modelo aportan a reducir la deviance una vez que las otras variables han sido incluidas en el modelo.

Este aumento en la deviance coincide con el porcentaje más bajo de casos bien clasificados correspondiente a la clase de los enfermos.

La **decisión** es que la variable "LOC1" forma parte del modelo.

Pruebas con las variables individualmente significativas

Nombre de variable	Hosmer and Lemeshow Test		Deviance	$\chi^2_{1,\alpha}$	α
	χ^2	Sig			
AGE	9.757	0.282	139.210	10.766	0.00103
CAN	9.707	0.286	136.282	7.838	0.00512
LOC2	7.701	0.463	137.830	9.386	0.00219
SYS	4.108	0.847	134.598	6.154	0.01311
TYP	8.753	0.364	148.308	19.864	0.00001

En cada prueba, α < 0.05, indica que la variable aportan al ajuste del modelo reduciendo la deviance, por lo tanto las variables AGE, CAN, LOC2, SYS y TYP forman parte del modelo.

7.1.4 Ejemplo 4 "Low Birth Weight Data" [HOS/89]

Estadísticas Descriptivas

Tablas de Distribución de Frecuencia de las variables categóricas

History of Hypertension

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	177	93.7	93.7	93.7
	Yes	12	6.3	6.3	100.0
	Total	189	100.0	100.0	

La variable RACE se transforma en dos variables dummy RACE1 y RACE2.

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	White	96	50.8	50.8	50.8
	Black	26	13.8	13.8	64.6
	Other	67	35.4	35.4	100.0
	Total	189	100.0	100.0	

Smoking Status During Pregnancy

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	115	60.8	60.8	60.8
	Yes	74	39.2	39.2	100.0
	Total	189	100.0	100.0	

Presence of Uterine Irritability

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	161	85.2	85.2	85.2
	Yes	28	14.8	14.8	100.0
	Total	189	100.0	100.0	

Tabla de Distribución de frecuencia de la variable de respuesta "Low Birth Weight (LOW)"

Low Birth Weight

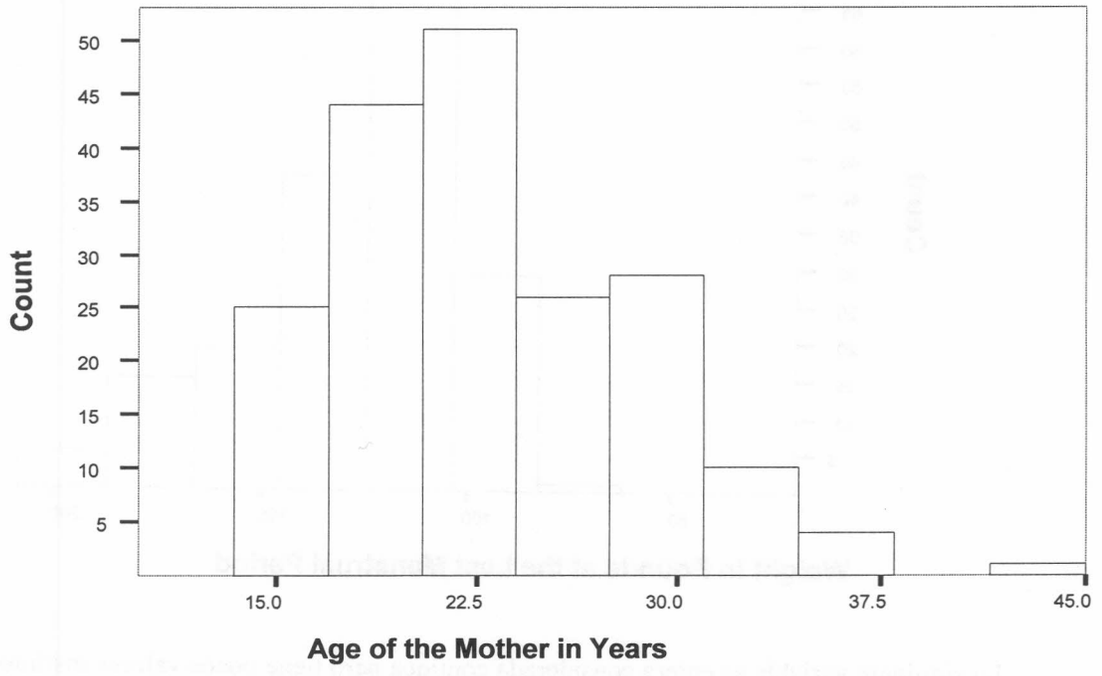
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Birth Weight >= 2500g	130	68.8	68.8	68.8
	Birth Weight < 2500g	59	31.2	31.2	100.0
	Total	189	100.0	100.0	

Estadística Descriptiva e histogramas de las variables continuas

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age of the Mother in Years	189	14	45	23.24	5.30
Number of Physician Visits During the First Trimester	189	0	6	.79	1.06
Weight in Pounds at the Last Menstrual Period	189	80	250	129.81	30.58
History of Premature Labor	189	0	3	.20	.49
Valid N (listwise)	189				

Para calcular el número de intervalos se utiliza la fórmula empírica1+(3.3*log(189)) = 8.51 se utilizan 9 intervalos en los histogramas.



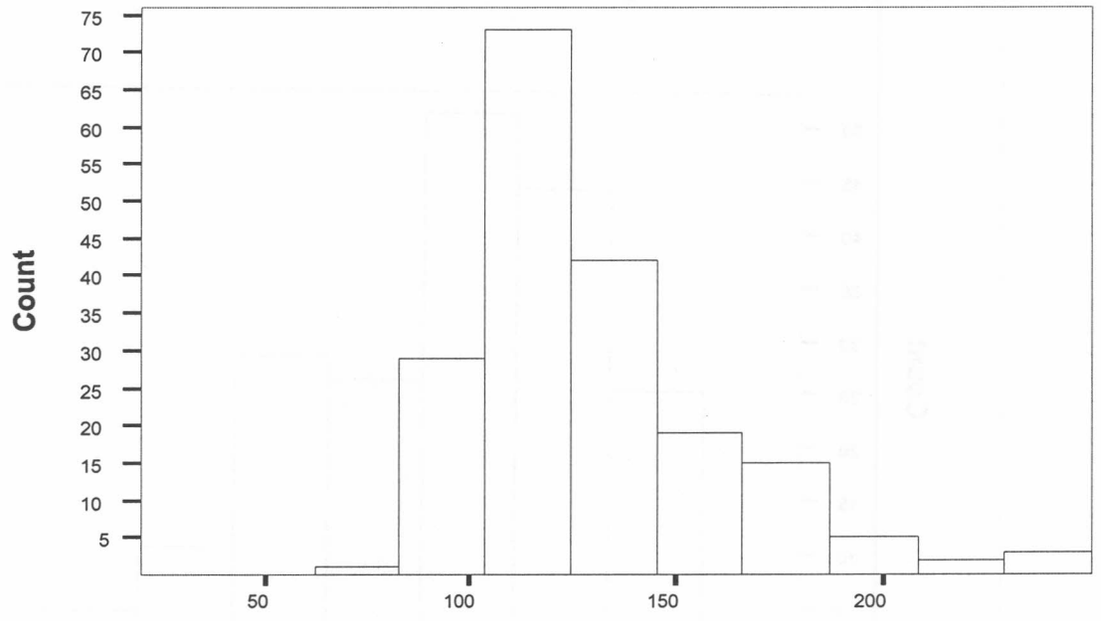
Age of the Mother in Years = 45 corresponde al caso 130, perteneciente a la clase 0.

La siguiente variable se muestra mejor con una tabla de frecuencias, es entera con pocos valores:

Number of Physician Visits During the First Trimester

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	100	52.9	52.9	52.9
	1	47	24.9	24.9	77.8
	2	30	15.9	15.9	93.7
	3	7	3.7	3.7	97.4
	4	4	2.1	2.1	99.5
	6	1	.5	.5	100.0
	Total	189	100.0	100.0	

Es una variable entera con significado numérico por lo tanto se la considera continua para el análisis y debido a que tiene 6 valores distintos no tiene sentido realizar un histograma con 9 intervalos, es preferible verlo en una tabla de distribución de frecuencias.

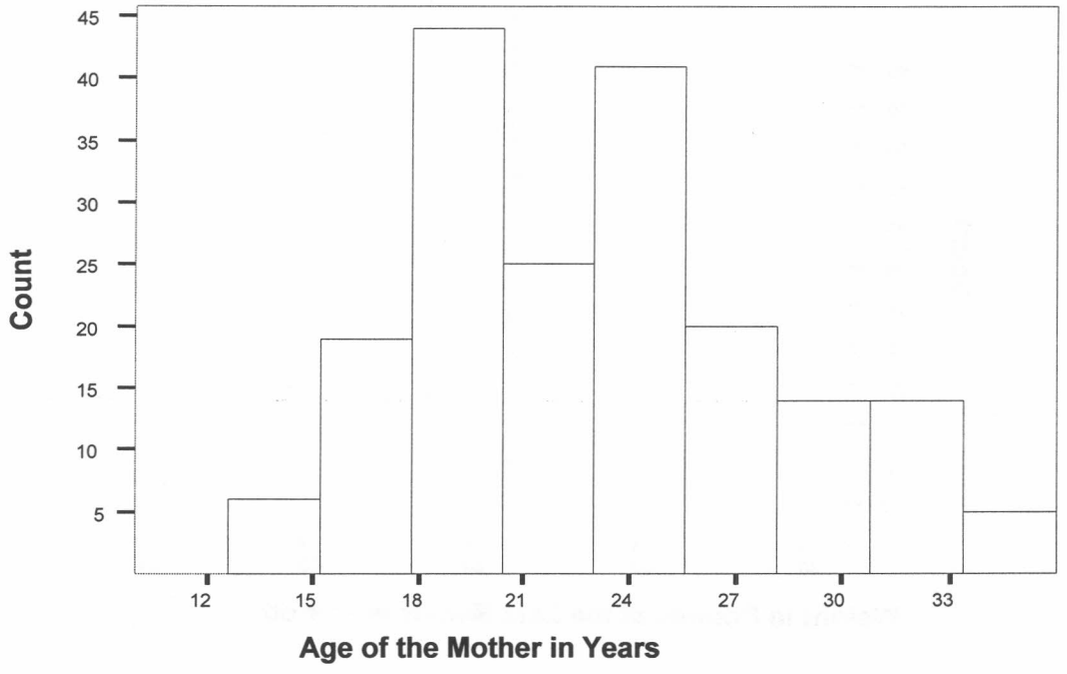


La siguiente variable es entera considerada continua pero tiene pocos valores distintos por lo tanto es preferible mostrarla con una tabla de frecuencias.

History of Premature Labor

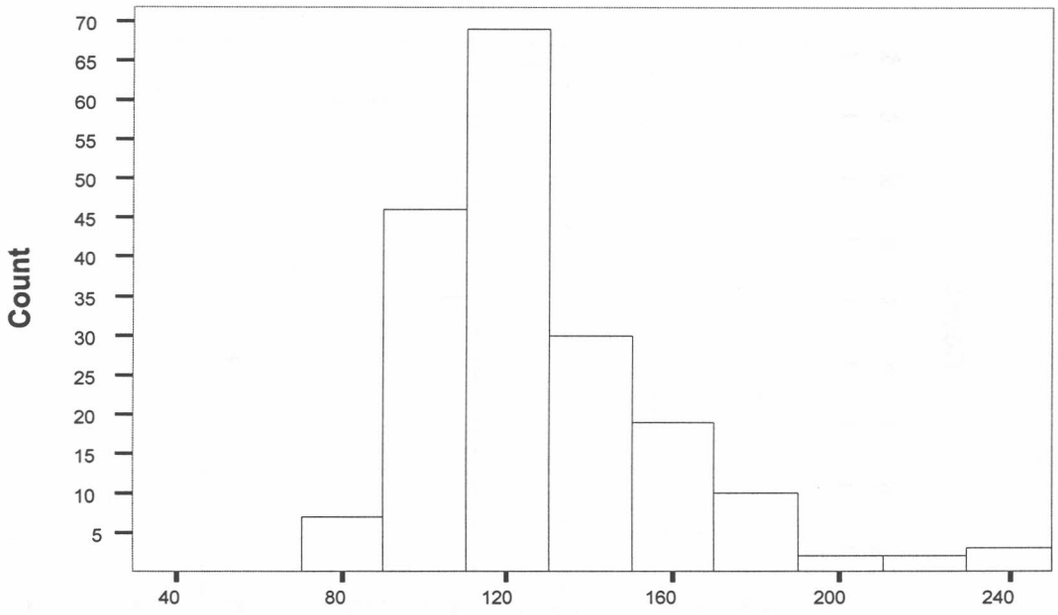
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	159	84.1	84.1	84.1
	1	24	12.7	12.7	96.8
	2	5	2.6	2.6	99.5
	3	1	.5	.5	100.0
	Total	189	100.0	100.0	

Sin outliers quedan los siguientes datos estadísticos:



Number of Physician Visits During the First Trimester

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	100	53.2	53.2	53.2
	1	46	24.5	24.5	77.7
	2	30	16.0	16.0	93.6
	3	7	3.7	3.7	97.3
	4	4	2.1	2.1	99.5
	6	1	.5	.5	100.0
	Total	188	100.0	100.0	



History of Premature Labor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	158	84.0	84.0	84.0
	1	24	12.8	12.8	96.8
	2	5	2.7	2.7	99.5
	3	1	.5	.5	100.0
	Total	188	100.0	100.0	

En el trabajo publicado se utilizan todos los casos del conjunto, nosotros adoptamos el ,mismo criterio y los utilizamos tanto para Regresión Logística como para Análisis Discriminante de Fisher.

Análisis Discriminante de Fisher

Vectores de Medias y Matriz de Varianzas-Covarianzas Combinada

Vectores de Medias

	Media de la clase 0	Media de la clase 1
AGE	23.6615	22.3051
FTV	0.8385	0.6949
HT	0.0385	0.1186
LWT	133.3000	122.1356
PTL	0.1308	0.3390
RACE1	0.5615	0.3898
RACE2	0.1154	0.1864
SMOKE	0.3385	0.5085
UI	0.1077	0.2373

Matriz de Varianzas-Covarianzas Combinada

	AGE	FTV	HT	LWT	PTL	RACE1	RACE2	SMOKE	UI
AGE	27.8268	1.1732	0.0030	26.0468	0.2495	0.4903	-0.2154	-0.0656	-0.1045
FTV	1.1732	1.1236	-0.0163	4.2286	-0.0169	0.0418	0.0042	-0.0093	-0.0185
HT	0.0030	-0.0163	0.0587	1.9709	-0.0055	-0.0029	0.0060	-0.0013	-0.0118
LWT	26.0468	4.2286	1.9709	913.0493	-1.6193	0.7325	2.5348	-0.2528	-1.3588
PTL	0.2495	-0.0169	-0.0055	-1.6193	0.2353	0.0089	-0.0090	0.0378	0.0344
RACE1	0.4903	0.0418	-0.0029	0.7325	0.0089	0.2462	-0.0680	0.0834	-0.0017
RACE2	-0.2154	0.0042	0.0060	2.5348	-0.0090	-0.0680	0.1188	-0.0036	-0.0066
SMOKE	-0.0656	-0.0093	-0.0013	-0.2528	0.0378	0.0834	-0.0036	0.2345	0.0061
UI	-0.1045	-0.0185	-0.0118	-1.3588	0.0344	-0.0017	-0.0066	0.0061	0.1239

Criterio para ajustar el modelo en cada paso

F de la tabla = 3.8911

número de paso	variable sale/entra	SCD MC	SCD MC SCD 0	SCD MC SCD 1	SCD MR	SCD MR SCD 0	SCD MR SCD 1	SCE MC	SCE MR	SCD MC-MR	SCE MC-MR	razón
1	sale: AGE SI	1683	1080.715	602.285	1496	946.6813	549.3187	36.6313	36.2202	187	0.41106	0.41546
2	sale: FTV SI	1496	946.68129	549.3187	1309	817.0587	491.9413	36.2202	36.2074	187	0.01288	0.01302
3	entra: AGE NO	1496	957.48714	538.5129	1309	817.0587	491.9413	36.5797	36.2074	187	0.37233	0.37631
3	sale: HT NO	1309	817.05874	491.9413	1122	732.1499	389.8501	36.2074	27.3015	187	8.90582	9.00107
3	sale: LWT NO	1309	817.05874	491.9413	1122	670.9434	451.0566	36.2074	29.6268	187	6.5806	6.65098
3	sale: PTL SI	1309	817.05874	491.9413	1122	704.8144	417.1856	36.2074	32.9376	187	3.26973	3.30471
4	entra: AGE NO	1309	845.43638	463.5636	1122	704.8144	417.1856	33.0668	32.9376	187	0.12917	0.13055
4	entra: FTV NO	1309	834.41676	474.5832	1122	704.8144	417.1856	32.9473	32.9376	187	0.00962	0.00973
4	sale: HT NO	1122	704.81443	417.1856	935	619.924	315.076	32.9376	24.0042	187	8.93339	9.02894
4	sale: LWT NO	1122	704.81443	417.1856	935	559.4454	375.5546	32.9376	25.6376	187	7.29999	7.37807
4	sale: RACE1 NO	1122	704.81443	417.1856	935	564.6234	370.3766	32.9376	27.9758	187	4.96187	5.01493
4	sale: RACE2 SI	1122	704.81443	417.1856	935	587.9678	347.0322	32.9376	32.2708	187	0.66681	0.67394
5	entra: AGE NO	1122	729.53149	392.4685	935	587.9678	347.0322	32.4558	32.2708	187	0.18497	0.18695
5	entra: FTV NO	1122	717.42079	404.5792	935	587.9678	347.0322	32.283	32.2708	187	0.01217	0.0123
5	sale: HT NO	935	587.96781	347.0322	748	503.1425	244.8575	32.2708	23.412	187	8.85883	8.95357
5	sale: LWT NO	935	587.96781	347.0322	748	446.2218	301.7782	32.2708	25.6348	187	6.63599	6.70696
5	entra: PTL NO	1122	699.86937	422.1306	935	587.9678	347.0322	35.4499	32.2708	187	3.17909	3.21309
5	sale: RACE1 NO	935	587.96781	347.0322	748	450.2467	297.7533	32.2708	23.9008	187	8.37001	8.45953
5	sale: SMOKE NO	935	587.96781	347.0322	748	456.8338	291.1662	32.2708	23.1923	187	9.07848	9.17558
5	sale: UI NO	935	587.96781	347.0322	748	486.7198	261.2802	32.2708	27.2034	187	5.06739	5.12158

Resumen de Clasificaciones

En cada paso se muestran las variables dentro del modelo con las estimaciones de los coeficientes de la función. Se usa p = 0.95, y corresponde F de la tabla = 3.8911.

Paso número 0 (con todas las variables)

AGE	0.0203	HT	-2.0189	PTL	-0.6381	RACE2	-0.4257	UI	-0.8627
FTV	-0.0348	LWT	0.0140	RACE1	0.7953	SMOKE	-0.8810		

Tabla de Clasificaciones

	predicho 0	1	
original 0	93	37	130
original 1	24	35	59
	117	72	189

Cantidad de casos bien clasificados 128
Cantidad de casos mal clasificados 61

Porcentajes

	predicho 0	1
original 0	71.54	28.46
original 1	40.68	59.32

Porcentaje de casos bien clasificados 67.72
Porcentaje de casos mal clasificados 32.28

Paso número 1

La variable AGE queda fuera del modelo y se muestra en blanco su posición:

		HT	-2.0295	PTL	-0.6078	RACE2	-0.4558	UI	-0.8801
FTV	-0.0171	LWT	0.0146	RACE1	0.8286	SMOKE	-0.9021		

Tabla de Clasificaciones

	predicho 0	1	
original 0	94	36	130
original 1	24	35	59
	118	71	189

Cantidad de casos bien clasificados 129
Cantidad de casos mal clasificados 60

Porcentajes

	predicho 0	1
original 0	72.31	27.69
original 1	40.68	59.32

Porcentaje de casos bien clasificados 68.25
Porcentaje de casos mal clasificados 31.75

Paso número 2

La variable FTV queda fuera del modelo y se muestra en blanco su posición:

		HT	-2.0214	PTL	-0.6075	RACE2	-0.4567	UI	-0.8780
		LWT	0.0145	RACE1	0.8252	SMOKE	-0.9004		

Tabla de Clasificaciones

	predicho 0	1	
original 0	94	36	130
original 1	24	35	59
	118	71	189

Cantidad de casos bien clasificados 129
Cantidad de casos mal clasificados 60

Porcentajes

	predicho 0	1
original 0	72.31	27.69
original 1	40.68	59.32

Porcentaje de casos bien clasificados 68.25
Porcentaje de casos mal clasificados 31.75

Paso número 3

La variable PTL queda fuera del modelo y se muestra en blanco su posición:

		HT	-2.0245			RACE2	-0.4283	UI	-1.0321
		LWT	0.0152	RACE1	0.8413	SMOKE	-0.9989		

Tabla de Clasificaciones

	predicho 0	1	
original 0	92	38	130
original 1	24	35	59
	116	73	189

Cantidad de casos bien clasificados 127
Cantidad de casos mal clasificados 62

Porcentajes

	predicho 0	1
original 0	70.77	29.23
original 1	40.68	59.32

Porcentaje de casos bien clasificados 67.20
Porcentaje de casos mal clasificados 32.80

Paso número 4

La variable RACE2 queda fuera del modelo y se muestra en blanco su posición:

		HT	-2.0158					UI	-1.0189
		LWT	0.0139	RACE1	0.9785	SMOKE	-1.0429		

Tabla de Clasificaciones

	predicho 0	1	
original 0	94	36	130
original 1	21	38	59
	115	74	189

Cantidad de casos bien clasificados 132
Cantidad de casos mal clasificados 57

Porcentajes

	predicho 0	1
original 0	72.31	27.69
original 1	35.59	64.41

Porcentaje de casos bien clasificados 69.84
Porcentaje de casos mal clasificados 30.16

Regresión Logística

Prueba con todas las variables

Deviance = -2 Log likelihood = 201.285.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.660	8	.685

Dado que sig = 0.685 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			LOW		Percentage Correct
			0	1	
Step 1	LOW	0	117	13	90.0
		1	36	23	39.0
	Overall Percentage				74.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.030	.037	.637	1	.425	.971
	FTV	.065	.172	.143	1	.705	1.067
	HT(1)	-1.863	.698	7.136	1	.008	.155
	LWT	-.015	.007	4.969	1	.026	.985
	PTL	.543	.345	2.474	1	.116	1.722
	RACE1(1)	.880	.441	3.990	1	.046	2.412
	RACE2(1)	-.392	.538	.531	1	.466	.676
	SMOKE(1)	-.939	.402	5.450	1	.020	.391
	UI(1)	-.768	.459	2.793	1	.095	.464
	Constant	4.442	1.673	7.050	1	.008	84.960

a. Variable(s) entered on step 1: AGE, FTV, HT, LWT, PTL, RACE1, RACE2, SMOKE, UI.

Dado que la variable "FTV" tiene nivel de significación sig = 0.705, la estimación del coeficiente no es significativa, es eliminada del modelo.

Tenemos entonces los resultados de la prueba dejando la variable "FTV" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.652	8	.290

Dado que sig = 0.290 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted LOW 0	Predicted LOW 1	Percentage Correct
Step 1	LOW	0	118	12	90.8
		1	38	21	35.6
	Overall Percentage				73.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.027	.036	.551	1	.458	.973
	HT(1)	-1.834	.692	7.026	1	.008	.160
	LWT	-.015	.007	4.803	1	.028	.985
	PTL	.542	.346	2.448	1	.118	1.719
	RACE1(1)	.862	.439	3.849	1	.050	2.367
	RACE2(1)	-.402	.539	.555	1	.456	.669
	SMOKE(1)	-.923	.401	5.306	1	.021	.397
	UI(1)	-.759	.459	2.727	1	.099	.468
	Constant	4.382	1.672	6.867	1	.009	79.968

a. Variable(s) entered on step 1: AGE, HT, LWT, PTL, RACE1, RACE2, SMOKE, UI.

Dado que la variable "AGE" tiene nivel de significación sig = 0.458, la estimación del coeficiente no es significativa, es eliminada del modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	179	201.285
Reducido	8	180	201.427

$\chi^2_{1,\alpha}$	α
0.142	0.70630

0.70630 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Tenemos entonces los resultados de la prueba dejando la variable "AGE" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.720	8	.461

Dado que sig = 0.461 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			LOW		Percentage Correct
			0	1	
Step 1	LOW	0	119	11	91.5
		1	37	22	37.3
	Overall Percentage				74.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	HT(1)	-1.855	.695	7.122	1	.008	.156
	LWT	-.016	.007	5.383	1	.020	.984
	PTL	.503	.341	2.175	1	.140	1.654
	RACE1(1)	.897	.434	4.275	1	.039	2.452
	RACE2(1)	-.429	.539	.633	1	.426	.651
	SMOKE(1)	-.939	.399	5.543	1	.019	.391
	UI(1)	-.786	.456	2.963	1	.085	.456
	Constant	3.922	1.538	6.504	1	.011	50.479

a. Variable(s) entered on step 1: HT, LWT, PTL, RACE1, RACE2, SMOKE, UI.

Dado que la variable "RACE2" tiene nivel de significación sig = 0.426, la estimación del coeficiente no es significativa, es eliminada del modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	180	201.427
Reducido	7	181	201.986

$\chi^2_{1,\alpha}$	α
0.559	0.45466

0.45466 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con la variable que fue evaluada anteriormente (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	180	Ver cuadro siguiente
Reducido	7	181	201.986

Nombre de variable que ingresa	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
FTV	0.813	210.930	0.056	0.81293

El nivel de significación sig = 0.813 > 0.05 indica que la estimación del coeficiente no es significativa.

α > 0.05, indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variable FTV no forma parte del modelo.

Prueba dejando la variable "RACE2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.374	8	.398

Dado que sig = 0.398 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted LOW 0	Predicted LOW 1	Percentage Correct
Step 1	LOW	0	118	12	90.8
		1	38	21	35.6
	Overall Percentage				73.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	HT(1)	-1.837	.693	7.028	1	.008	.159
	LWT	-.014	.006	4.863	1	.027	.986
	PTL	.494	.341	2.094	1	.148	1.639
	RACE1(1)	1.048	.390	7.208	1	.007	2.852
	SMOKE(1)	-.999	.393	6.474	1	.011	.368
	UI(1)	-.771	.459	2.822	1	.093	.462
	Constant	3.287	1.298	6.418	1	.011	26.771

a. Variable(s) entered on step 1: HT, LWT, PTL, RACE1, SMOKE, UI.

Dado que la variable "PTL" tiene nivel de significación sig = 0.148 > 0.05, la estimación del coeficiente no es significativa, es eliminada del modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	181	201.986
Reducido	6	182	202.616

$\chi^2_{1,\alpha}$	α
0.63	0.42736

0.42736 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables que fueron evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	181	Ver cuadro siguiente
Reducido	6	182	202.616

Nombre de variable que ingresa	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
FTV	0.799	202.551	0.065	0.79876
AGE	0.429	201.981	0.635	0.42553

En cada una de las pruebas, el Nivel de Significación (sig) > 0.05, indica que la estimación del coeficiente no es significativa.

En cada una de las pruebas, α > 0.05, indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: las variables FTV y AGE no forman parte del modelo.

Prueba dejando la variable "PTL" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.466	8	.234

Dado que sig = 0.234 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			LOW		Percentage Correct
			0	1	
Step 1	LOW	0	117	13	90.0
		1	38	21	35.6
	Overall Percentage				73.0

a. The cut value is .500

Todas las variables tienen nivel de significación sig<0.05, las estimaciones de los coeficientes se consideran significativas.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	HT(1)	-1.853	.688	7.251	1	.007	.157
	LWT	-.015	.006	5.608	1	.018	.985
	RACE1(1)	1.065	.388	7.527	1	.006	2.900
	SMOKE(1)	-1.091	.387	7.966	1	.005	.336
	UI(1)	-.893	.450	3.943	1	.047	.410
	Constant	3.677	1.266	8.432	1	.004	39.532

a. Variable(s) entered on step 1: HT, LWT, RACE1, SMOKE, UI.

Todas las variables tienen nivel de significación sig < 0.05, las estimaciones de los coeficientes se consideran significativas.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	182	202.616
Reducido	5	183	204.766

$\chi^2_{1,\alpha}$	α
2.15	0.14257

0.14257 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables que fueron evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	182	Ver cuadro siguiente
Reducido	5	183	204.766

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
FTV	0.797	204.700	0.066	0.79725
AGE	0.572	204.444	0.322	0.57041
RACE2	0.458	204.217	0.549	0.45873

En cada una de las pruebas, el Nivel de Significación (sig) > 0.05, indica que la estimación del coeficiente no es significativa, y α > 0.05, indicando que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: las variables FTV, AGE y RACE2 no forman parte del modelo.

Pruebas con las variables que son individualmente significativas

Modelo	Cantidad de variables	G. L.	Deviance
Completo	5	183	204.766
Reducido	4	184	Ver cuadro siguiente

Nombre de variable fuera del modelo	Hosmer and Lemeshow Test		Deviance	$\chi^2_{1,\alpha}$	α
	χ^2	sig			
HT	7.851	0.448	212.369	7.603	0.00583
LWT	2.132	0.831	211.167	6.401	0.01141
RACE1	12.137	0.145	212.826	8.060	0.00453
SMOKE	7.830	0.450	213.152	8.386	0.00378
UI	3.875	0.868	208.658	3.892	0.04587

En cada prueba $\alpha < 0.05$, indica que la variable que no fue incluida aporta significativamente al ajuste del modelo reduciendo la Deviance, por lo tanto las variables HT, LWT, RACE1, SMOKE y UI forman parte del modelo.

7.1.5 Ejemplo 5 "South African Heart Disease"

Estadísticas Descriptivas

Tablas de Distribuciones de Frecuencias de las variables categóricas

FAMHISTPRESENTE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Absent	270	58.4	58.4	58.4
	Present	192	41.6	41.6	100.0
	Total	462	100.0	100.0	

Tabla de distribuciones de Frecuencias de la variable de respuesta "CHD"

CHD

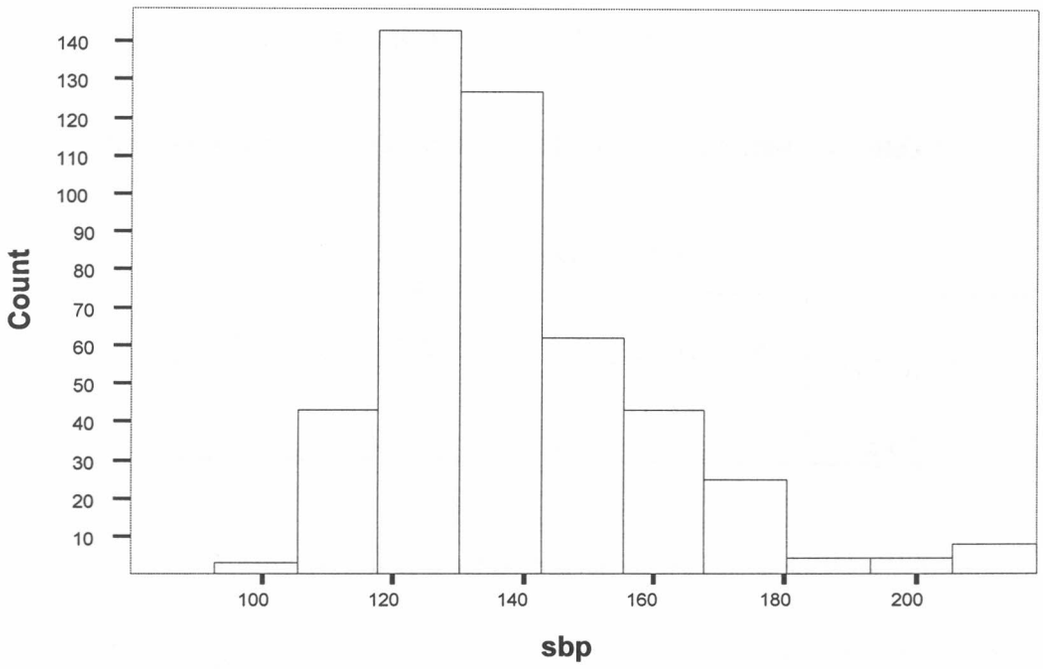
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	absence	302	65.4	65.4	65.4
	presence	160	34.6	34.6	100.0
	Total	462	100.0	100.0	

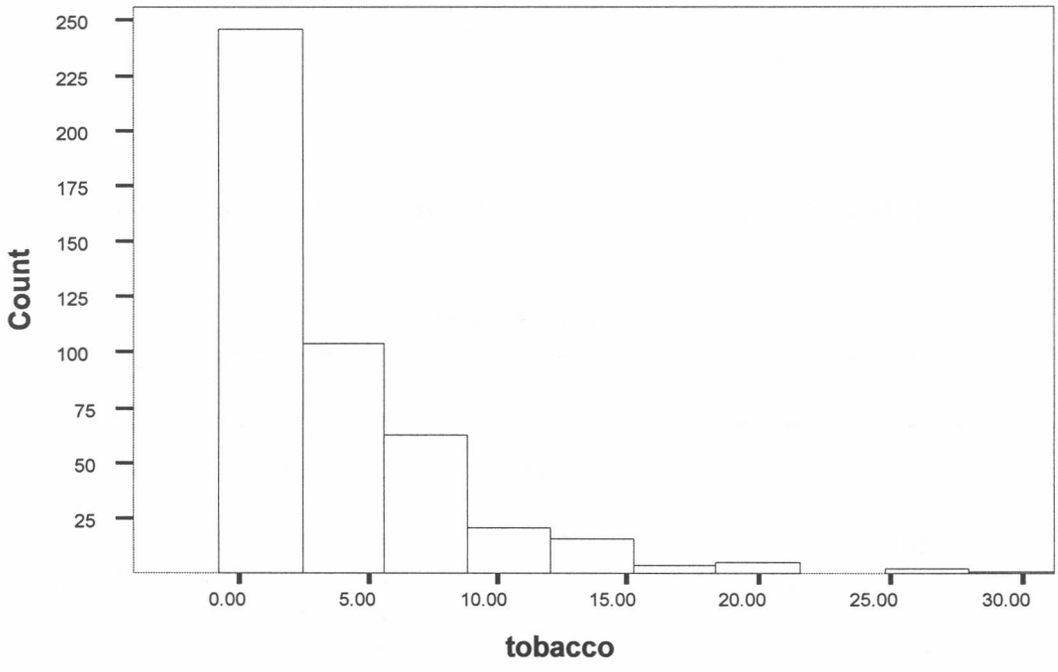
Estadísticas Descriptivas e Histogramas de las variables continuas

Descriptive Statistics

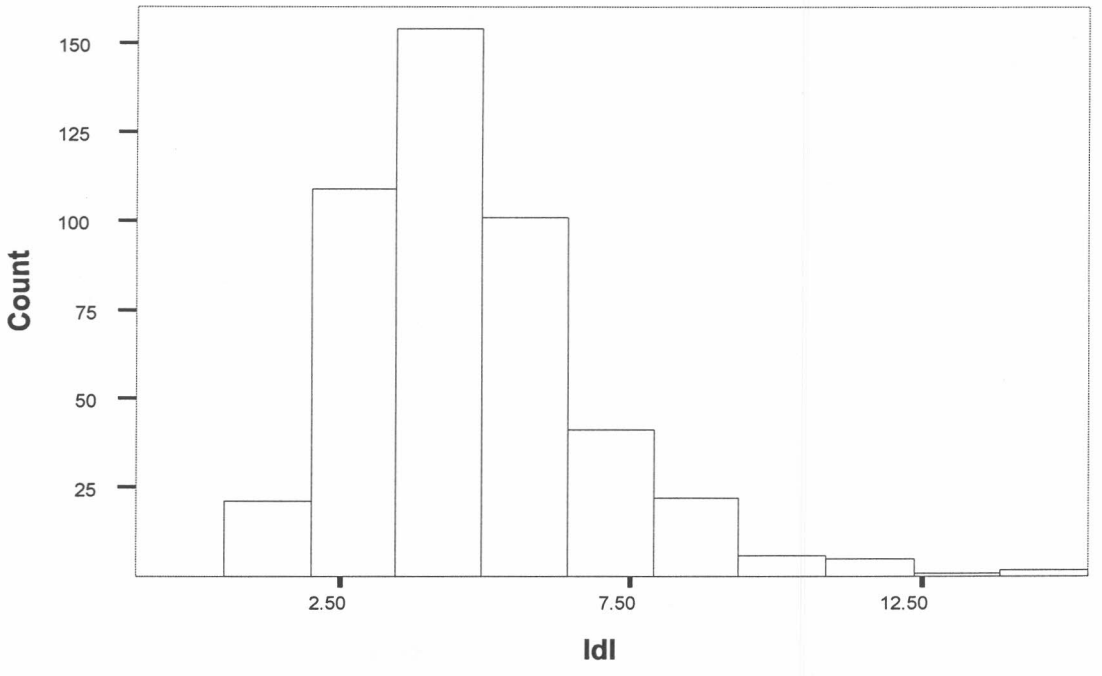
	N	Minimum	Maximum	Mean	Std. Deviation
SBP	462	101	218	138.33	20.50
TOBACCO	462	.00	31.20	3.6356	4.5930
LDL	462	.98	15.33	4.7403	2.0709
ADIPOSITY	462	6.74	42.49	25.4067	7.7807
TYPEA	462	13	78	53.10	9.82
OBESITY	462	14.70	46.58	26.0441	4.2137
ALCOHOL	462	.00	147.19	17.0444	24.4811
AGE	462	15	64	42.82	14.61
Valid N (listwise)	462				

Para calcular el número de intervalos se utiliza la fórmula empírica $1+(3.3*\log(462)) = 9.79$ se utilizan 10 intervalos.

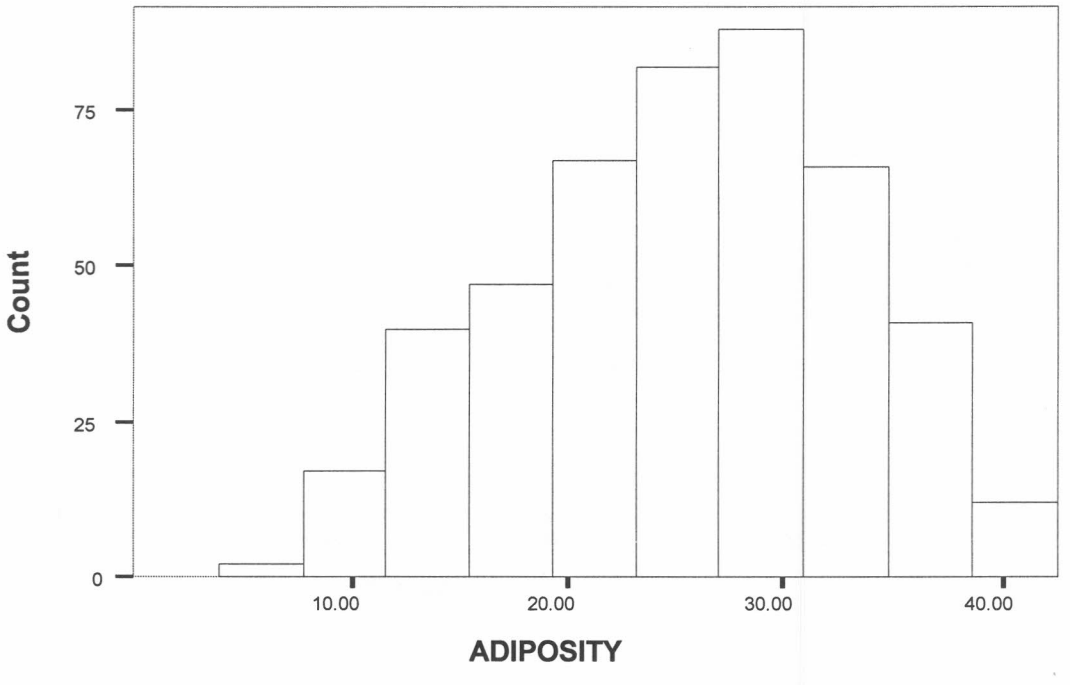


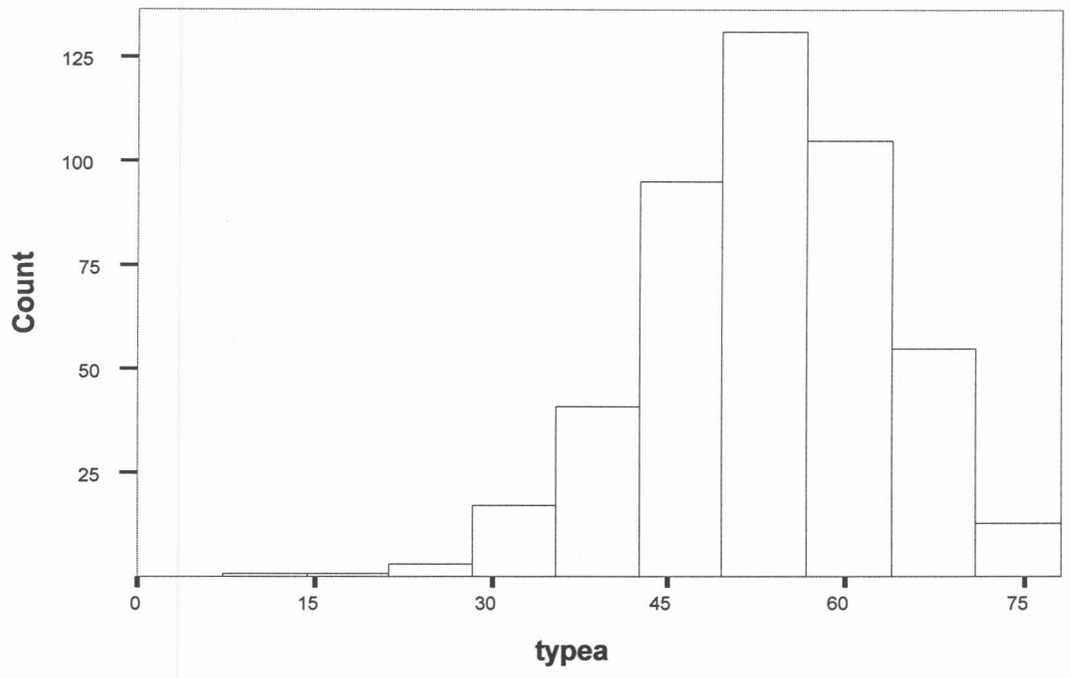


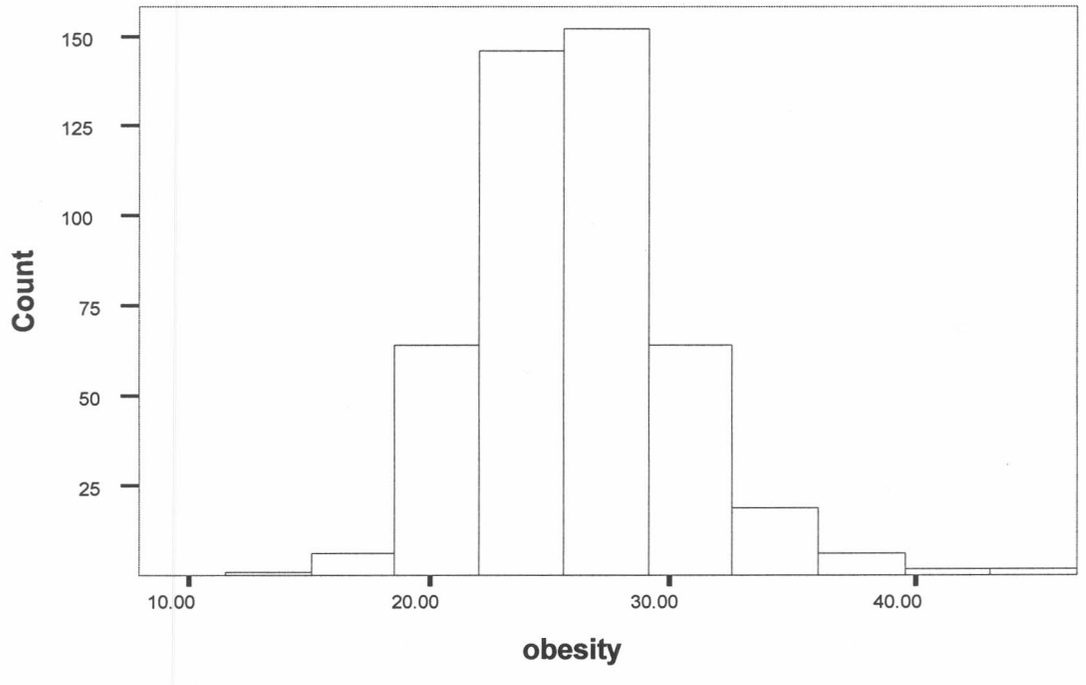
tobacco = 31.20 corresponde al caso 115, correspondiente a la clase 1.
tobacco = 27.40 corresponde al caso 162, correspondiente a la clase 1.
tobacco = 25.01 corresponde al caso 407, correspondiente a la clase 1.

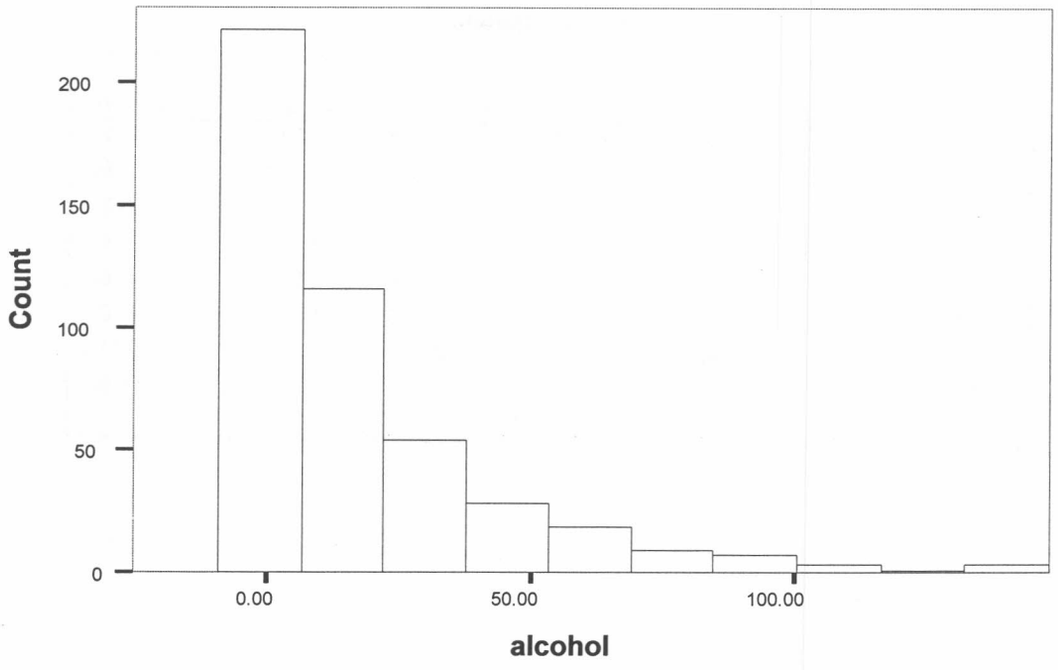


ldl=15.33 corresponde al caso 17, perteneciente a la clase 0.
ldl=14.16 corresponde al caso 413, perteneciente a la clase 1.

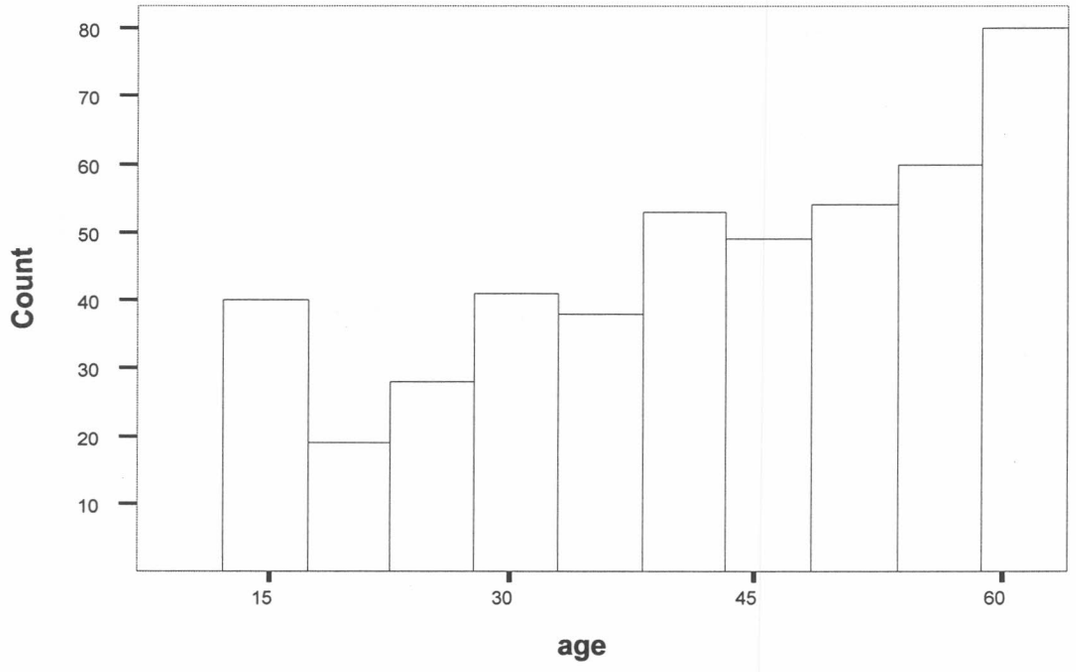



Count
125
100
75
50
25
0
15
30
45
60
75
typea


Count
150
125
100
75
50
25
10.00
20.00
30.00
40.00
obesity



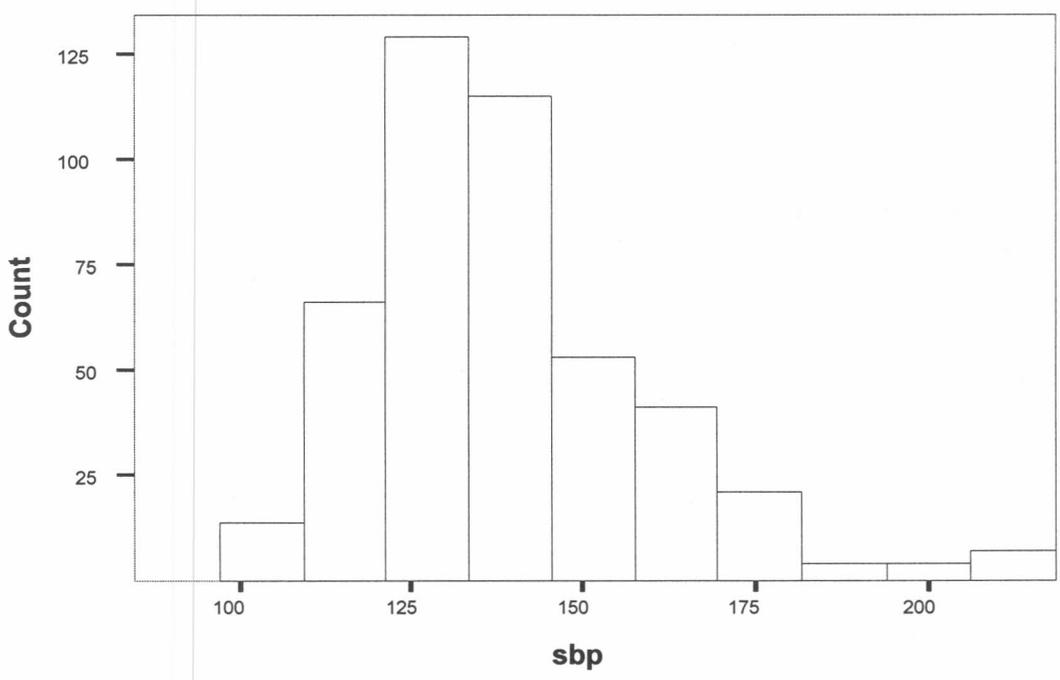
alcohol = 144.00 corresponde al caso 375
alcohol = 145.29 corresponde al caso 372
alcohol = 147.19 corresponde al caso 155

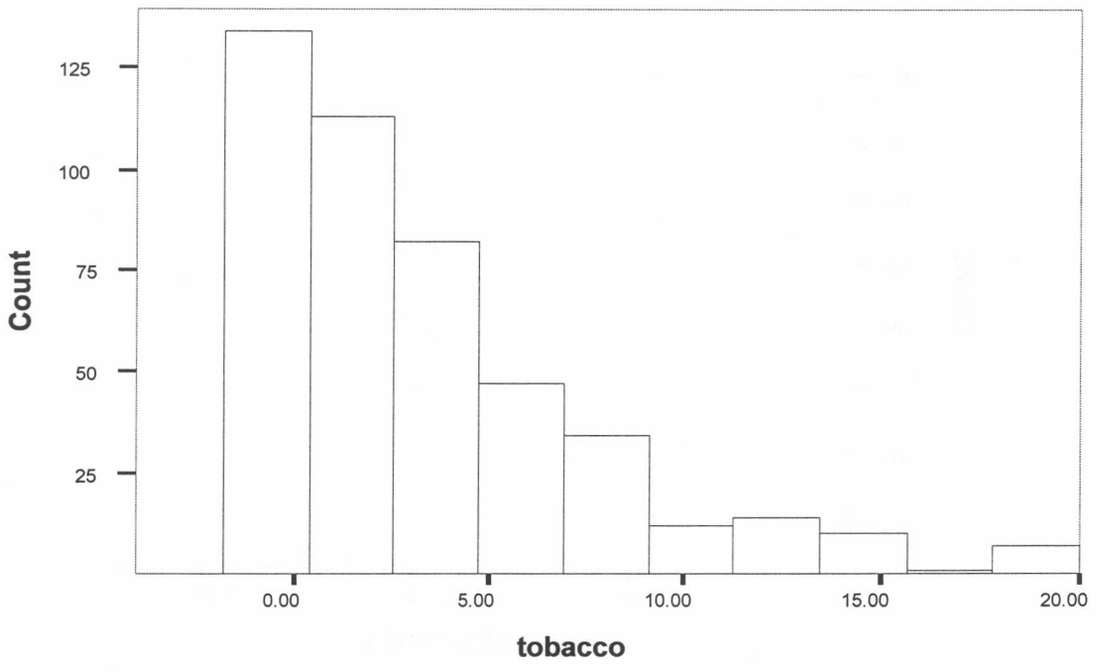


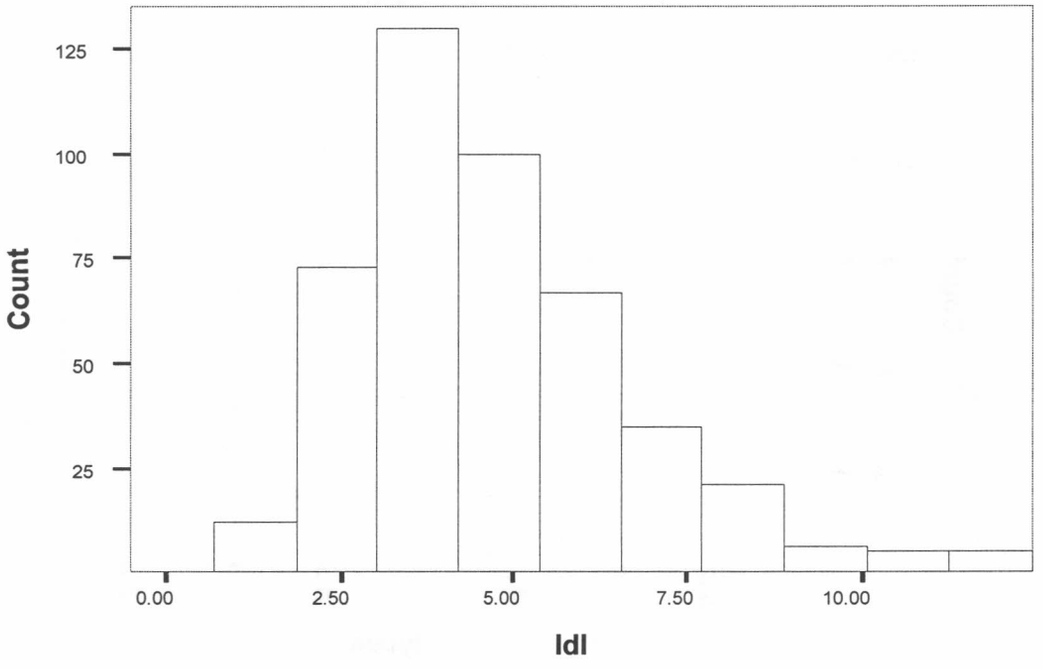
Sin los outliers quedan los siguientes histogramas:

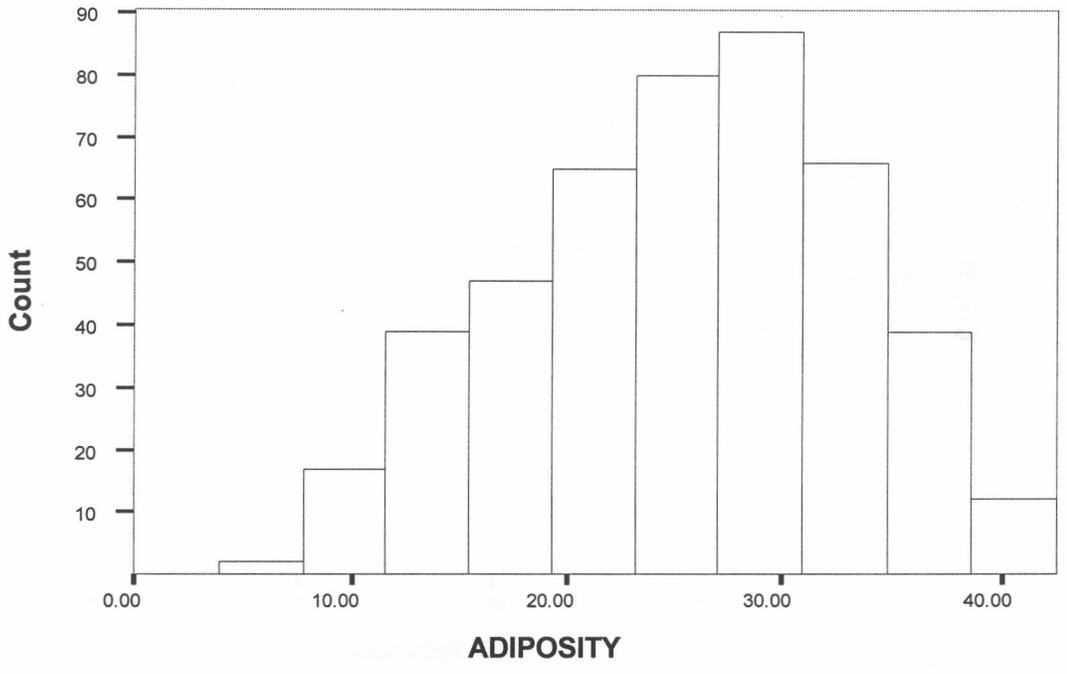
Descriptive Statistics

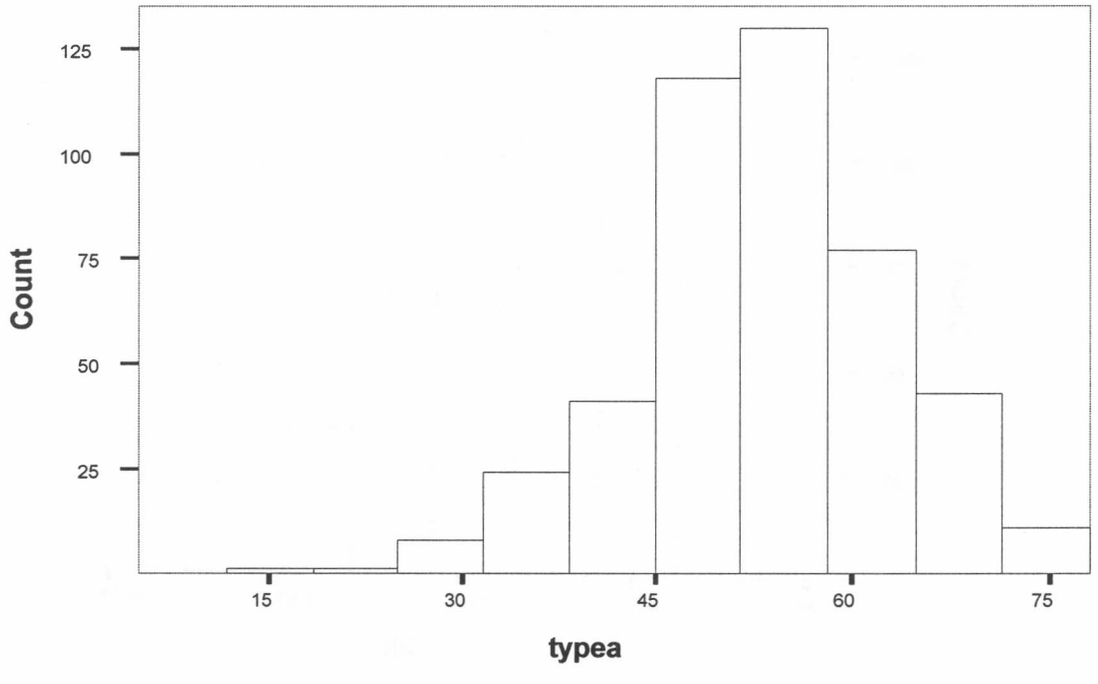
	N	Minimum	Maximum	Mean	Std. Deviation
SBP	454	101	218	138.16	20.23
TOBACCO	454	.00	20.00	3.4611	4.1694
LDL	454	.98	12.42	4.7079	1.9725
ADIPOSITY	454	6.74	42.49	25.3881	7.7849
TYPEA	454	13	78	52.99	9.82
OBESITY	454	14.70	46.58	26.0451	4.2203
ALCOHOL	454	.00	120.03	16.0890	22.2614
AGE	454	15	64	42.70	14.64
Valid N (listwise)	454				

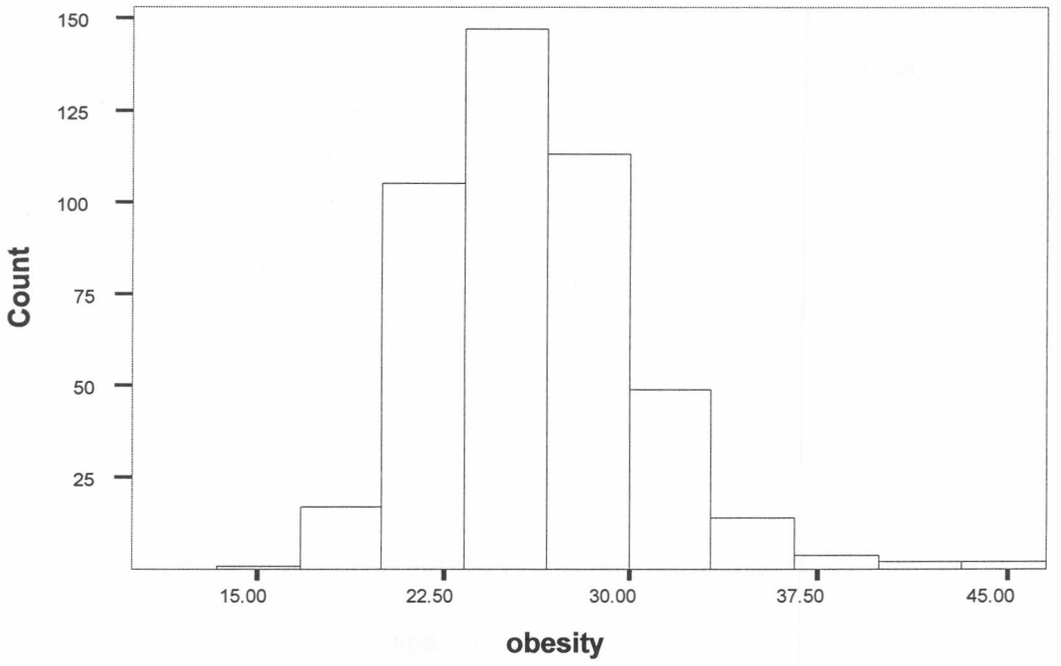


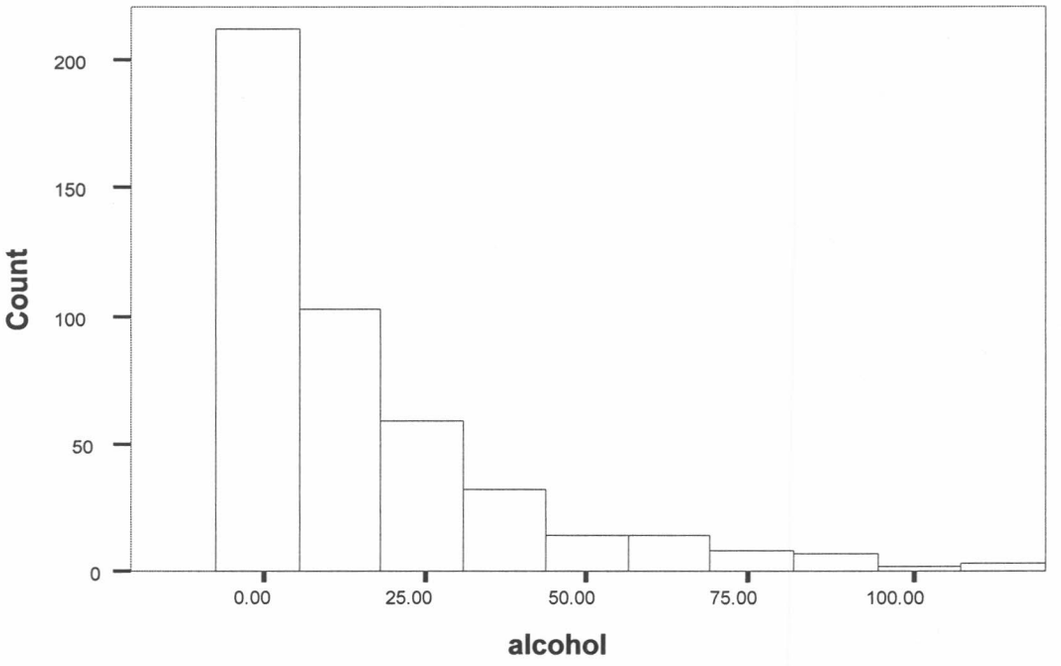

125
100
75
50
25
Count
0.00
5.00
10.00
15.00
20.00
tobacco

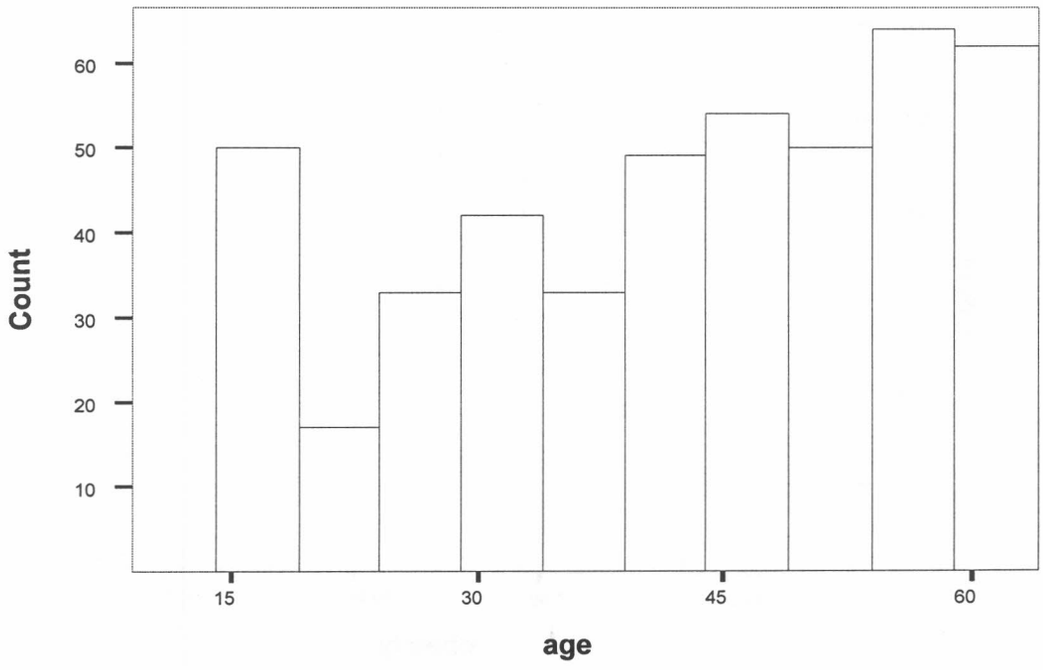

125
100
75
50
25
Count
0.00
2.50
5.00
7.50
10.00
ldl


Count
90
80
70
60
50
40
30
20
10
0.00
10.00
20.00
30.00
40.00
ADIPOSITY


Count
125
100
75
50
25
15
30
45
60
75
typea


150
125
100
75
50
25
Count
15.00
22.50
30.00
37.50
45.00
obesity


200
150
100
50
0
Count
0.00
25.00
50.00
75.00
100.00
alcohol


Count
60
50
40
30
20
10
15
30
45
60
age

Análisis Discriminante de Fisher

Vectores de Medias y Matriz de Varianzas-Covarianzas Combinada

Vectores de Medias

	Media de la clase 0	Media de la clase 1
SBP	135.4603	143.7375
TOBACCO	2.6347	5.5249
LDL	4.3442	5.4879
ADIPOSITY	23.9691	28.1203
FAMHISTPRESENTE	0.3179	0.6000
TYPEA	52.3675	54.4938
OBESITY	25.7375	26.6229
ALCOHOL	15.9314	19.1452
AGE	38.8543	50.2938

Matriz de Varianzas-Covarianzas Combinada

SBP	405.4348	14.5852	4.5812	49.1644	0.3370	-15.5877	18.9388	64.4004	95.1339
TOBACCO	14.5852	19.2426	0.7632	7.5381	0.0158	-2.0573	1.8335	20.5170	22.7654
LDL	4.5812	0.7632	4.0006	6.0327	0.0919	0.3446	2.6601	-2.5329	6.4789
ADIPOSITY	49.1644	7.5381	6.0327	56.7529	0.4328	-5.3096	22.7079	16.1190	60.5087
FAMHISTPRESENTE	0.3370	0.0158	0.0919	0.4328	0.2258	0.0811	0.1840	0.7685	0.9974
TYPEA	-15.5877	-2.0573	0.3446	-5.3096	0.0811	95.5656	2.6401	7.9600	-20.2783
OBESITY	18.9388	1.8335	2.6601	22.7079	0.1840	2.6401	17.6154	4.6894	15.6970
ALCOHOL	64.4004	20.5170	-2.5329	16.1190	0.7685	7.9600	4.6894	598.2766	27.8859
AGE	95.1339	22.7654	6.4789	60.5087	0.9974	-20.2783	15.6970	27.8859	184.1321

Criterio para ajustar el modelo en cada paso

p = 0.95, F de la tabla = 3.8616

número de paso	variable sale/entra	SCD						SCE		SCD	SCE	razón
		MC			MR			MC	MR	MC-MR	MC-MR	
			SCD 0	SCD 1		SCD 0	SCD 1					
1	sale: SBP SI	4140	2469.1701	1670.8299	3680	2243.3273	1436.6727	142.1058	139.9809	460	2.1249	2.13413
2	sale: TOBACCO NO	3680	2243.3273	1436.6727	3220	2051.2899	1168.7101	139.9809	124.6514	460	15.3295	15.3962
2	sale: LDL NO	3680	2243.3273	1436.6727	3220	1977.9252	1242.0748	139.9809	127.4649	460	12.516	12.5704
2	sale: ADIPOSITY SI	3680	2243.3273	1436.6727	3220	1941.8109	1278.1891	139.9809	139.5428	460	0.43809	0.44
3	entra: SBP NO	3680	2168.3546	1511.6454	3220	1941.8109	1278.1891	141.7955	139.5428	460	2.25269	2.26249
3	sale: TOBACCO NO	3220	1941.8109	1278.1891	2760	1749.7732	1010.2268	139.5428	124.2144	460	15.3284	15.3951
3	sale: LDL NO	3220	1941.8109	1278.1891	2760	1680.6923	1079.3077	139.5428	125.3527	460	14.1902	14.2519
3	sale: FAMHISTPRESENTE NO	3220	1941.8109	1278.1891	2760	1647.3829	1112.6171	139.5428	117.3788	460	22.164	22.2604
3	sale: TYPEA NO	3220	1941.8109	1278.1891	2760	1653.5877	1106.4123	139.5428	128.6092	460	10.9336	10.9811
3	sale: OBESITY SI	3220	1941.8109	1278.1891	2760	1663.9043	1096.0957	139.5428	136.3156	460	3.22721	3.24124
4	entra: SBP NO	3220	1894.1541	1325.8459	2760	1663.9043	1096.0957	137.841	136.3156	460	1.52542	1.53206
4	sale: TOBACCO NO	2760	1663.9043	1096.0957	2300	1471.7422	828.25775	136.3156	120.8185	460	15.4971	15.5645
4	sale: LDL NO	2760	1663.9043	1096.0957	2300	1405.3887	894.61125	136.3156	124.6671	460	11.6486	11.6992
4	entra: ADIPOSITY NO	3220	1947.7695	1272.2305	2760	1663.9043	1096.0957	136.9043	136.3156	460	0.58871	0.59127
4	sale: FAMHISTPRESENTE NO	2760	1663.9043	1096.0957	2300	1368.9161	931.08388	136.3156	114.6391	460	21.6765	21.7708
4	sale: TYPEA NO	2760	1663.9043	1096.0957	2300	1377.6635	922.33651	136.3156	126.417	460	9.8986	9.94164
4	sale: ALCOHOL SI	2760	1663.9043	1096.0957	2300	1384.3459	915.65405	136.3156	136.2708	460	0.04483	0.04503
5	entra: SBP NO	2760	1615.2294	1144.7706	2300	1384.3459	915.65405	137.7211	136.2708	460	1.45026	1.45656
5	sale: TOBACCO NO	2300	1384.3459	915.65405	1840	1195.3133	644.68665	136.2708	120.5782	460	15.6926	15.7608
5	sale: LDL NO	2300	1384.3459	915.65405	1840	711.52729	1128.4727	136.2708	124.4259	460	11.8449	11.8964
5	entra: ADIPOSITY NO	2760	1671.8928	1088.1072	2300	1384.3459	915.65405	136.8808	136.2708	460	0.61003	0.61269
5	sale: FAMHISTPRESENTE NO	2300	1384.3459	915.65405	1840	1091.2659	748.73405	136.2708	114.6281	460	21.6427	21.7368
5	sale: TYPEA NO	2300	1384.3459	915.65405	1840	1096.6459	743.35413	136.2708	126.4131	460	9.85771	9.90057
5	entra: OBESITY NO	2760	1663.6787	1096.3213	2300	1384.3459	915.65405	139.5233	136.2708	460	3.25253	3.26668
5	sale: AGE NO	2300	1384.3459	915.65405	1840	1042.3964	797.6036	136.2708	109.4716	460	26.7991	26.9157

Función de Fisher y tablas de Clasificaciones en cada paso

En cada paso se muestran las variables dentro del modelo con las estimaciones de los coeficientes de la función. Se usa p = 0.95, y corresponde F de la tabla = 3.8616.

Paso número 0 (con todas las variables)

SBP	-0.0077	ADIPOSITY	-0.0132	OBESITY	0.0643
TOBACCO	-0.0954	FAMHISTPRESENTE	-0.9984	ALCOHOL	0.0013
LDL	-0.1910	TYPEA	-0.0350	AGE	-0.0394

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	209	93	302
original 1	42	118	160
	251	211	462

Cantidad de casos bien clasificados 327
Cantidad de casos mal clasificados 135

Porcentajes

	predicho 0	predicho 1
original 0	69.21	30.79
original 1	26.25	73.75

Porcentaje de casos bien clasificados 70.78
Porcentaje de casos mal clasificados 29.22

Paso número 1

La variable SBP queda fuera del modelo y se muestra en blanco su posición:

		ADIPOSITY	-0.0157	OBESITY	0.0616
TOBACCO	-0.0961	FAMHISTPRESENTE	-0.9891	ALCOHOL	7.5286E-4
LDL	-0.1903	TYPEA	-0.0343	AGE	-0.0421

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	211	91	302
original 1	44	116	160
	255	207	462

Cantidad de casos bien clasificados 327
Cantidad de casos mal clasificados 135

Porcentajes

	predicho 0	predicho 1
original 0	69.87	30.13
original 1	27.50	72.50

Porcentaje de casos bien clasificados 70.78
Porcentaje de casos mal clasificados 29.22

Paso número 2

La variable ADIPOSITY queda fuera del modelo y se muestra en blanco su posición:

				OBESITY	0.0454
TOBACCO	-0.0961	FAMHISTPRESENTE	-0.9874	ALCOHOL	5.7294E-4
LDL	-0.1979	TYPEA	-0.0337	AGE	-0.0456

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	209	93	302
original 1	42	118	160
	251	211	462

Cantidad de casos bien clasificados 327
Cantidad de casos mal clasificados 135

Porcentajes

	predicho 0	predicho 1
original 0	69.21	30.79
original 1	26.25	73.75

Porcentaje de casos bien clasificados 70.78
Porcentaje de casos mal clasificados 29.22

Paso número 3

La variable OBESITY queda fuera del modelo y se muestra en blanco su posición:

TOBACCO	-0.0966	FAMHISTPRESENTE	-0.9760	ALCOHOL	8.6790E-4
LDL	-0.1729	TYPEA	-0.0319	AGE	-0.0424

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	210	92	302
original 1	44	116	160
	254	208	462

Cantidad de casos bien clasificados 326
Cantidad de casos mal clasificados 136

Porcentajes

	predicho 0	predicho 1
original 0	69.54	30.46
original 1	27.50	72.50

Porcentaje de casos bien clasificados 70.56
Porcentaje de casos mal clasificados 29.44

Paso número 4

La variable ALCOHOL queda fuera del modelo y se muestra en blanco su posición:

TOBACCO	-0.0957	FAMHISTPRESENTE	-0.9730		
LDL	-0.1738	TYPEA	-0.0318	AGE	-0.0424

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	208	94	302
original 1	45	115	160
	253	209	462

Cantidad de casos bien clasificados 323
Cantidad de casos mal clasificados 139

Porcentajes

	predicho 0	predicho 1
original 0	68.87	31.13
original 1	28.13	71.87

Porcentaje de casos bien clasificados 69.91
Porcentaje de casos mal clasificados 30.09

Regresión Logística

Prueba con todas las variables dentro del modelo

Deviance = -2 Log likelihood = 472.140.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.251	8	.510

Dado que sig = 0.510 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			CHD		Percentage
	Observed		0	1	Correct
Step 1	CHD	0	256	46	84.8
		1	77	83	51.9
	Overall Percentage				73.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SBP	.007	.006	1.288	1	.256	1.007
	TOBACCO	.079	.027	8.903	1	.003	1.083
	LDL	.174	.060	8.498	1	.004	1.190
	ADIPOSIT	.019	.029	.403	1	.526	1.019
	FAMHISTP(1)	-.925	.228	16.488	1	.000	.396
	TYPEA	.040	.012	10.329	1	.001	1.040
	OBESITY	-.063	.044	2.021	1	.155	.939
	ALCOHOL	.000	.004	.001	1	.978	1.000
	AGE	.045	.012	13.901	1	.000	1.046
	Constant	-5.225	1.315	15.782	1	.000	.005

a. Variable(s) entered on step 1: SBP, TOBACCO, LDL, ADIPOSIT, FAMHISTP, TYPEA, OBESITY, ALCOHOL, AGE.

La variable "ALCOHOL" tiene nivel de significación sig = 0.978 > 0.05, la estimación del coeficiente no es significativa.

Prueba dejando la variable "ALCOHOL" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.415	8	.601

Dado que sig = 0.601 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted CHD 0	Predicted CHD 1	Percentage Correct
Step 1	CHD	0	256	46	84.8
		1	77	83	51.9
	Overall Percentage				73.4

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SBP	.007	.006	1.317	1	.251	1.007
	TOBACCO	.080	.026	9.272	1	.002	1.083
	LDL	.174	.059	8.560	1	.003	1.190
	ADIPOSIT	.019	.029	.406	1	.524	1.019
	FAMHISTP(1)	-.926	.227	16.595	1	.000	.396
	TYPEA	.040	.012	10.340	1	.001	1.040
	OBESITY	-.063	.044	2.027	1	.154	.939
	AGE	.045	.012	14.038	1	.000	1.046
	Constant	-5.224	1.315	15.792	1	.000	.005

a. Variable(s) entered on step 1: SBP, TOBACCO, LDL, ADIPOSIT, FAMHISTP, TYPEA, OBESITY, AGE.

La variable "ADIPOSIT" tiene nivel de significación sig = 0.524 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	9	452	472.140
Reducido	8	453	472.141

$\chi^2_{1,\alpha}$	α
0.001	0.974773

0.974773 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba dejando la variable "ADIPOSITY" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.707	8	.883

Dado que sig = 0.883 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			CHD		Percentage
	Observed		0	1	Correct
Step 1	CHD	0	258	44	85.4
		1	75	85	53.1
	Overall Percentage				74.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SBP	.007	.006	1.423	1	.233	1.007
	TOBACCO	.080	.026	9.327	1	.002	1.083
	LDL	.182	.058	9.832	1	.002	1.200
	FAMHISTP(1)	-.924	.227	16.576	1	.000	.397
	TYPEA	.039	.012	10.091	1	.001	1.040
	OBESITY	-.042	.029	2.055	1	.152	.959
	AGE	.049	.011	21.482	1	.000	1.050
	Constant	-5.492	1.245	19.456	1	.000	.004

a. Variable(s) entered on step 1: SBP, TOBACCO, LDL, FAMHISTP, TYPEA, OBESITY, AGE.

La variable "SBP" tiene nivel de significación sig = 0.233 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	453	472.141
Reducido	7	454	472.549

$\chi^2_{1,\alpha}$	α
0.408	0.522986

0.522986 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Prueba con la variable que fue evaluada anteriormente (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	8	453	Ver cuadro siguiente
Reducido	7	454	472.549

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
ALCOHOL	0.950	472.545	0.004	0.94957

El nivel de significación sig = 0.950 > 0.05 indica que la estimación del coeficiente no es significativa, y α > 0.05, indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: la variable ALCOHOL no forma parte del modelo.

Prueba dejando la variable "SBP" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.251	8	.619

Dado que sig = 0.619 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted		
		CHD		Percentage Correct
		0	1	
Step 1 CHD	0	257	45	85.1
	1	73	87	54.4
Overall Percentage				74.5

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	TOBACCO	.080	.026	9.481	1	.002	1.083
	LDL	.184	.058	9.972	1	.002	1.202
	FAMHISTP(1)	-.916	.226	16.366	1	.000	.400
	TYPEA	.038	.012	9.812	1	.002	1.039
	OBESITY	-.038	.029	1.669	1	.196	.963
	AGE	.052	.010	25.882	1	.000	1.053
	Constant	-4.787	1.088	19.370	1	.000	.008

a. Variable(s) entered on step 1: TOBACCO, LDL, FAMHISTP, TYPEA, OBESITY, AGE.

La variable "OBESITY" tiene nivel de significación sig = 0.196 > 0.05, la estimación del coeficiente no es significativa.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	454	472.549
Reducido	6	455	473.980

$\chi^2_{1,\alpha}$	α
1.431	0.231601

0.231601 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables que fueron evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	7	454	Ver cuadro siguiente
Reducido	6	455	473.980

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
ALCOHOL	0.832	473.935	0.045	0.83200
ADIPOSIT	0.475	473.980	0.515	0.47298

En cada prueba, el Nivel de Significación (sig) > 0.05, indica que la estimación del coeficiente no es significativa, y α > 0.05, indicando que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: las variables ALCOHOL y ADIPOSIT no forman parte del modelo.

Prueba dejando la variable "OBESITY" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.531	8	.992

Dado que sig = 0.992 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			CHD		Percentage Correct
	Observed		0	1	
Step 1	CHD	0	256	46	84.8
		1	73	87	54.4
	Overall Percentage				74.2

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	TOBACCO	.080	.026	9.646	1	.002	1.084
	LDL	.162	.055	8.685	1	.003	1.176
	FAMHISTP(1)	-.908	.226	16.183	1	.000	.403
	TYPEA	.037	.012	9.306	1	.002	1.038
	AGE	.050	.010	24.444	1	.000	1.052
	Constant	-5.538	.928	35.630	1	.000	.004

a. Variable(s) entered on step 1: TOBACCO, LDL, FAMHISTP, TYPEA, AGE.

Todas las variables tienen nivel de significación < 0.05, las estimaciones de los coeficientes son significativas.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	455	473.980
Reducido	5	456	475.686

$\chi^2_{1,\alpha}$	α
1.706	0.191505

0.191505 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables que fueron evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo	6	455	Ver cuadro siguiente
Reducido	5	456	475.686

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
ALCOHOL	0.851	475.651	0.035	0.85160
ADIPOSIT	0.623	475.443	0.243	0.62205
SBP	0.311	475.655	0.031	0.86024

En cada prueba, el Nivel de Significación (sig) > 0.05, indica que la estimación del coeficiente no es significativa, y α > 0.05, indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión: las variables ALCOHOL, ADIPOSIT y SBP no forman parte del modelo.

Pruebas con las variables que son individualmente significativas

Modelo	Cantidad de variables	G. L.	Deviance
Completo	5	456	475.686
Reducido	4	457	Ver cuadro siguiente

Nombre de variable	Hosmer and Lemeshow Test		Deviance	$\chi^2_{1,\alpha}$	α
	χ^2	sig			
TOBACCO	3.438	0.904	486.032	10.346	0.00130
LDL	8.165	0.417	484.714	9.028	0.00266
FAMISHISTP	7.165	0.519	492.095	16.409	0.00005
TYPEA	5.714	0.679	485.444	9.758	0.00179
AGE	15.883	0.044	502.379	26.693	0.00000

En cada prueba, $\alpha < 0.05$, indica que la variable que no fue incluida aporta significativamente al ajuste del modelo reduciendo la Deviance, por lo tanto las variables TOBACCO, LDL, FAMISHISTP, TYPEA y AGE forman parte del modelo.

7.1.6 Ejemplo 6 "French Wine - Dementia Study"

Estadísticas Descriptivas

Tablas de distribuciones de Frecuencias de las variables categóricas

La variable Wine Consumption se transforma en dos variables dummy "wine Consuption1" y "Wine Consumption2".

Wine Consumption

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No Wine Consumption	131	48.2	48.2	48.2
	Up to 1/4 liter/day	108	39.7	39.7	87.9
	More than 1/4 liter/day	33	12.1	12.1	100.0
	Total	272	100.0	100.0	

Diastolic Blood Pressure

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No High Blood Pressure	235	86.4	86.4	86.4
	High Blood Pressure	37	13.6	13.6	100.0
	Total	272	100.0	100.0	

Tabla de Distribuciones de Frecuencias de la variable de respuesta "Incident Dementia"

Incident Dementia

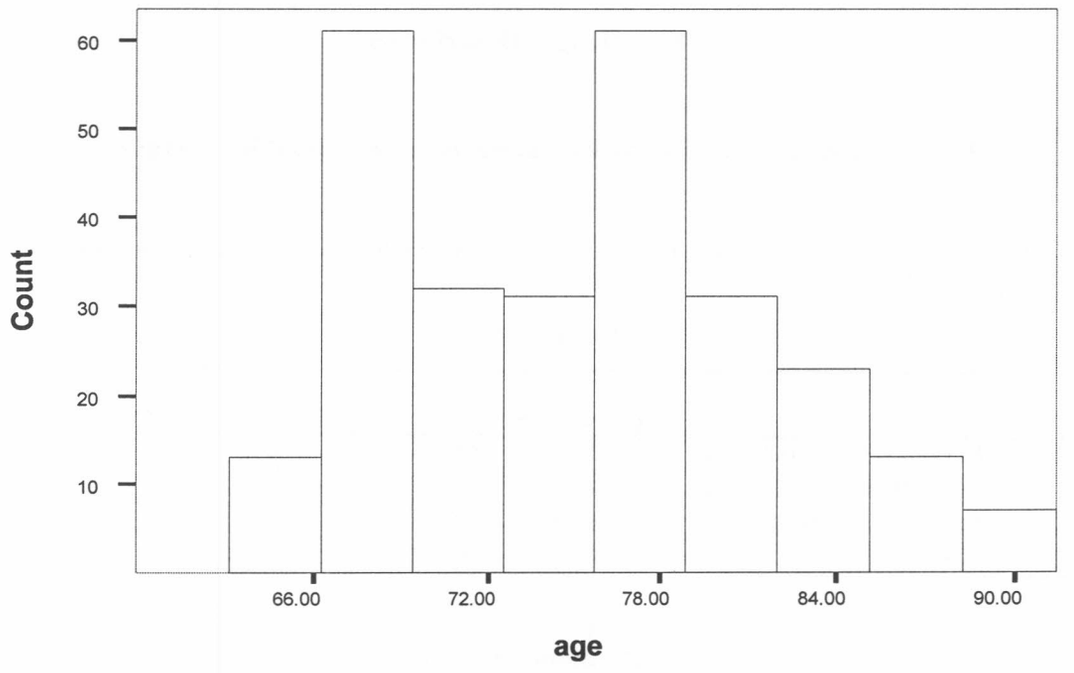
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	200	73.5	73.5	73.5
	Yes	72	26.5	26.5	100.0
	Total	272	100.0	100.0	

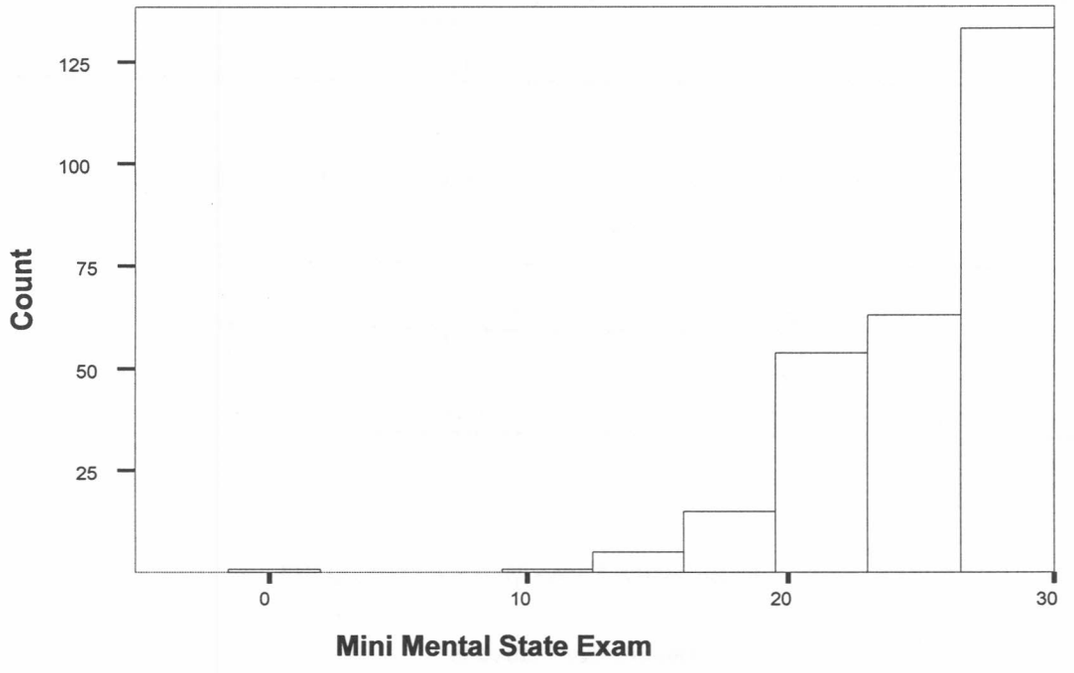
Estadística Descriptiva e Histograma

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	272	64.75	91.42	75.2213	6.4440
Mini Mental State Exam	272	0	30	25.30	3.99
Valid N (listwise)	272				

Para calcular el número de intervalos se utiliza la fórmula empírica $1+(3.3*\log(272)) = 9.03$ se utilizan 9 intervalos en el histograma.



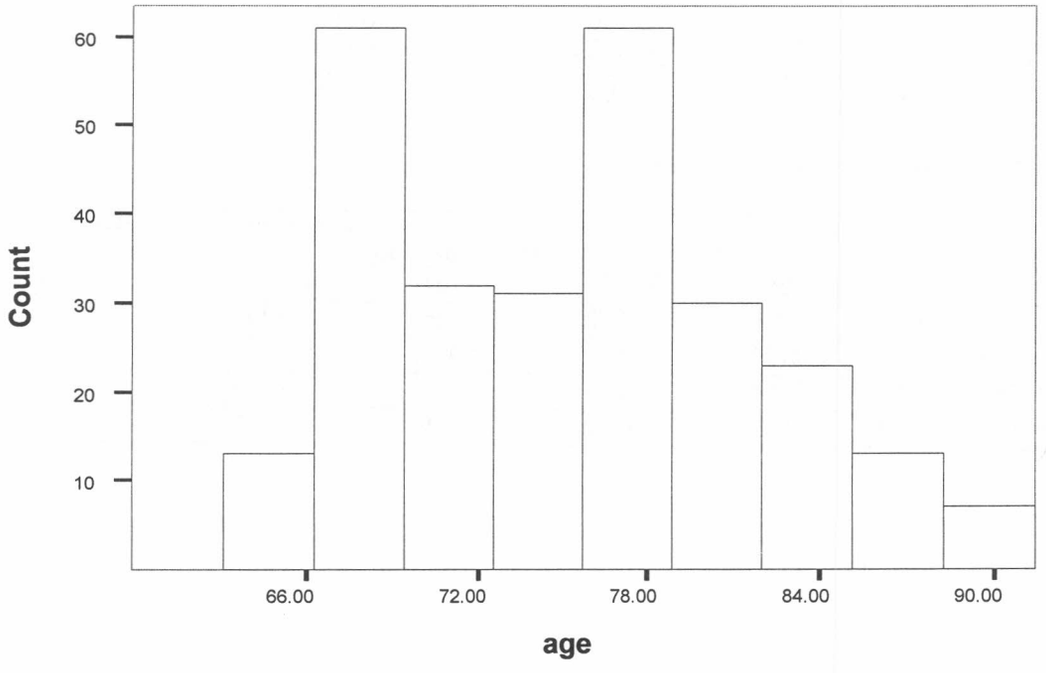


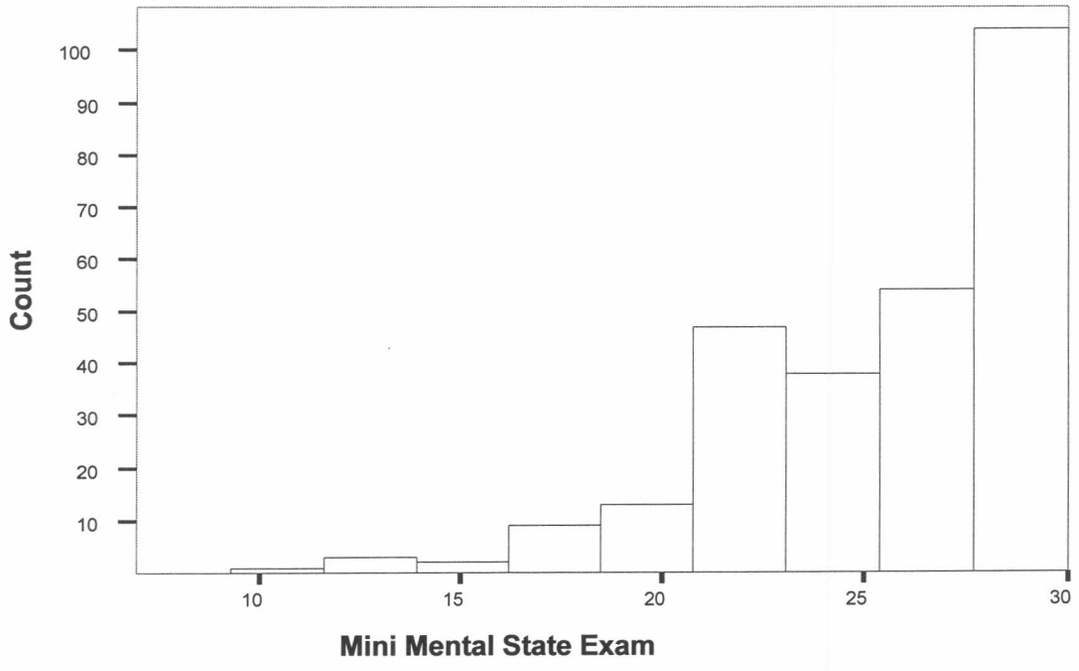
Mini Mental State Exam = 0 corresponde al caso 170, perteneciente a la clase 1.
El rango de la variable es [0, 30], por lo tanto el caso 170 es considerado dentro del análisis.

Sin outliers tenemos los siguientes histogramas:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	271	64.75	91.42	75.1972	6.4436
Mini Mental State Exam	271	11	30	25.39	3.69
Valid N (listwise)	271				





Análisis Discriminante de Fisher

Vectores de Medias y Matriz de Varianzas-Covarianzas

Vectores de Medias

	Media de la clase 0	Media de la clase 1
Age	73.6770	79.5111
WineConsumption1	0.3600	0.5000
WineConsumption2	0.1500	0.0417
MiniMentalStateExam	26.5300	21.8750
DiastolicBloodPressure	0.1600	0.0694

Matriz de Varianzas-Covarianzas Combinada

	Age	Wine Consumption1	Wine Consumption2	MiniMentalState Exam	DiastolicBlood Pressure
Age	35.0050	0.2808	0.0151	-4.3286	-0.1165
WineConsumption1	0.2808	0.2373	-0.0456	-0.3099	-0.0149
WineConsumption2	0.0151	-0.0456	0.1051	0.0425	-0.0148
MiniMentalStateExam	-4.3286	-0.3099	0.0425	11.7396	0.0765
DiastolicBloodPressure	-0.1165	-0.0149	-0.0148	0.0765	0.1168

Criterio ajustar el modelo en cada paso

F de la tabla = 3.8758

número de paso	Variable sale/entra	SCD						SCE		SCD MC-MR	SCE MC-MR	razón
		MC			MR			MC	MR			
			SCD 0	SCD 1		SCD 0	SCD 1					
1	sale: Age NO	1350	998.2711	351.7289	1080	794.5322	285.4678	131.7688	104.7354	270	27.0334	27.2336
1	sale: WineConsumption1 SI	1350	998.2711	351.7289	1080	801.6503	278.3497	131.7688	131.0123	270	0.75648	0.76208
2	sale: Age NO	1080	801.6503	278.3497	810	598.4519	211.5481	131.0123	104.44	270	26.5723	26.7691
2	sale: WineConsumption2 NO	1080	801.6503	278.3497	810	559.191	250.809	131.0123	125.6981	270	5.3142	5.35356
2	sale: MiniMentalStateExam NO	1080	801.6503	278.3497	810	682.2015	127.7985	131.0123	61.15173	270	69.8606	70.378
2	sale DiastolicBloodPressure SI	1080	801.6503	278.3497	810	571.1231	238.8769	131.0123	129.1687	270	1.84363	1.85728
3	sale: Age NO	810	571.1231	238.8769	540	368.272	171.728	129.1687	101.9411	270	27.2276	27.4293
3	Entra: WineConsumption1 NO	1080	766.0256	313.9744	810	571.1231	238.8769	129.6556	129.1687	270	0.48691	0.49052
3	sale: WineConsumption2 NO	810	571.1231	238.8769	540	327.9468	212.0532	129.1687	124.5885	270	4.5802	4.61413
3	sale: MiniMentalStateExam NO	810	571.1231	238.8769	540	452.2893	87.71072	129.1687	57.66761	270	71.5011	72.0307

Función de Fisher y Tabla de Clasificación en cada paso

En cada paso se muestran las variables dentro del modelo con las estimaciones de los coeficientes de la función. Se usa p = 0.95, F de la tabla = 3.8758.

Paso número 0 (con todas las variables)

Age	-0.1240	WineConsumption2	1.1054	DiastolicBloodPressure	0.5964
WineConsumption1	0.2632	MiniMentalStateExam	0.3498		

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	160	40	200
original 1	18	54	72
	178	94	272

Cantidad de casos bien clasificados 214
Cantidad de casos mal clasificados 58

Porcentajes

	predicho 0	predicho 1
original 0	80.00	20.00
original 1	25.00	75.00

Porcentaje de casos bien clasificados 78.68
Porcentaje de casos mal clasificados 21.32

Paso número 1

La variable `WineConsumption1` queda fuera del modelo y se muestra en blanco su posición:

Age	-0.1227	WineConsumption2	0.9873	DiastolicBloodPressure	0.5529
		MiniMentalStateExam	0.3440		

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	166	34	200
original 1	16	56	72
	182	90	272

Cantidad de casos bien clasificados 222
Cantidad de casos mal clasificados 50

Porcentajes

	predicho 0	predicho 1
original 0	83.00	17.00
original 1	22.22	77.78

Porcentaje de casos bien clasificados 81.62
Porcentaje de casos mal clasificados 18.38

Paso número 2

La variable `DiastolicBloodPressure` queda fuera del modelo y se muestra en blanco su posición:

Age	-0.1240	WineConsumption2	0.9081		
		MiniMentalStateExam	0.3474		

Tabla de Clasificaciones

	predicho 0	predicho 1	
original 0	164	36	200
original 1	19	53	72
	183	89	272

Cantidad de casos bien clasificados 217
Cantidad de casos mal clasificados 55

Porcentajes

	predicho 0	predicho 1
original 0	82.00	18.00
original 1	26.39	73.61

Porcentaje de casos bien clasificados 79.78
Porcentaje de casos mal clasificados 20.22

Regresión Logística

Prueba con todas las variables dentro del modelo

Deviance = -2 Log likelihood = 215.242.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.848	8	.944

Dado que sig = 0.944 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Incident Dementia		Percentage Correct
			No	Yes	
Step 1	Incident Dementia	No	184	16	92.0
		Yes	35	37	51.4
	Overall Percentage				81.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.115	.029	15.388	1	.000	1.122
	WINE1(1)	.330	.369	.804	1	.370	1.392
	WINE2(1)	1.545	.753	4.204	1	.040	4.688
	MINIMENT	-.304	.053	33.378	1	.000	.738
	DIASTOLI(1)	.746	.605	1.517	1	.218	2.108
	Constant	-4.653	2.958	2.476	1	.116	.010

a. Variable(s) entered on step 1: AGE, WINE1, WINE2, MINIMENT, DIASTOLI.

Dado que la variable "WINE1" tiene nivel de significación sig = 0.370 > 0.05, la estimación del coeficiente no es significativa, es eliminada del modelo

Prueba dejando la variable "WINE1" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.265	8	.618

Dado que sig = 0.618 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed		Predicted: Incident Dementia No	Predicted: Incident Dementia Yes	Percentage Correct
Step 1	Incident Dementia No	183	17	91.5
	Yes	34	38	52.8
	Overall Percentage			81.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.113	.029	14.893	1	.000	1.120
	WINE2(1)	1.377	.725	3.607	1	.058	3.963
	MINIMENT	-.296	.052	32.775	1	.000	.744
	DIASTOLI(1)	.682	.603	1.277	1	.258	1.977
	Constant	-4.272	2.938	2.115	1	.146	.014

a. Variable(s) entered on step 1: AGE, WINE2, MINIMENT, DIASTOLI.

Dado que la variable "DIASTOLI" tiene nivel de significación sig = 0.258 > 0.05, la estimación de los coeficientes no es significativa, es eliminada del modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	5	266	215.242
Reducido	4	267	216.058

$\chi^2_{1,\alpha}$	α
0.816	0.36635

0.36635 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Regresión Logística vs. Análisis Discriminante de Fisher
Departamento de Computación
Facultad de Ciencias Exactas y Naturales. U.B.A.

Prueba dejando la variable "DIASTOLI" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.185	8	.840

Dado que sig =0.840 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

			Predicted		
			Incident Dementia		Percentage
	Observed		No	Yes	Correct
Step 1	Incident Dementia	No	184	16	92.0
		Yes	34	38	52.8
	Overall Percentage				81.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.114	.029	15.202	1	.000	1.120
	WINE2(1)	1.309	.726	3.256	1	.071	3.704
	MINIMENT	-.302	.052	34.088	1	.000	.739
	Constant	-3.502	2.846	1.514	1	.218	.030

a. Variable(s) entered on step 1: AGE, WINE2, MINIMENT.

Dado que la variable "WINE2" tienen nivel de significación sig = 0.071 > 0.05, la estimación del coeficiente no es significativa, es eliminada del modelo.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	4	267	216.058
Reducido	3	268	217.447

$\chi^2_{1,\alpha}$	α
1.389	0.23857

0.23857 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Pruebas con las variables evaluadas en pasos anteriores (paso hacia atrás)

Modelo	Cantidad de variables	G. L.	Deviance
Completo		269	Ver cuadro siguiente
Reducido		270	217.447

Nombre de variable	Nivel de significación (sig)	Deviance	$\chi^2_{1,\alpha}$	α
WINE1	0.460>0.05	216.894	0.551	0.45791

0.45791 > 0.05 indica que las variables que no se incluyeron en el modelo aportan poco una vez que las otras variables han sido incluidas en el modelo.

Decisión : La variable WINE1 no forma parte del modelo.

Prueba dejando la variable "WINE2" fuera del modelo

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.858	8	.552

Dad que sig = 0.552 > 0.05 no se rechaza la hipótesis nula de buen ajuste.

Classification Table[a]

Observed			Predicted		
			Incident Dementia		Percentage Correct
			No	Yes	
Step 1	Incident Dementia	No	185	15	92.5
		Yes	35	37	51.4
	Overall Percentage				81.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.112	.029	15.287	1	.000	1.119
	MINIMENT	-.307	.051	35.666	1	.000	.736
	Constant	-2.094	2.706	.599	1	.439	.123

a. Variable(s) entered on step 1: AGE, MINIMENT.

Todas las variables tiene nivel de significación sig < 0.05, las estimaciones de los coeficientes son significativas.

Modelo	Cantidad de variables	G. L.	Deviance
Completo	3	268	217.447
Reducido	2	269	221.466

$\chi^2_{1,\alpha}$	α
4.019	0.04499

0.04499 < 0.05 indica que la variable que no se incluyó en el modelo aporta al ajuste del modelo una vez que las otras variables han sido incluidas en el modelo.

Decisión : la variable WINE2 forma parte del modelo.

Pruebas correspondientes a variables significativas individualmente

Modelo	Cantidad de variables	G. L.	Deviance
Completo	3	268	Ver cuadro siguiente
Reducido	2	269	217.447

Nombre de variable	Deviance	$\chi^2_{1,\alpha}$	α
AGE	233.930	16.483	0.00005
MINIMENT	262.794	45.347	0.00000

En cada prueba α < 0.05 , indica que la variable que quedó fuera del modelo en cada prueba aporta al ajuste del modelo reduciendo la Deviance.

Decisión: Las variables AGE y MINIMENT siguen formando parte del modelo.

7.2 Cálculos cuando las variables dummy no se codifican con 0 y 1

Matriz de Varianzas-Covarianzas Combinada

Llamamos nv1 al nuevo valor que toma la variable en los casos en que valía 1, y nv0 al nuevo valor que toma la variable en los casos en que valía 0. También, n_0=cantidad de casos en donde la variable dependiente (y) vale 0, n_1=cantidad de casos en donde la variable dependiente (y) vale 1, n es la cantidad total de casos, y p es la cantidad de variables independientes.

Se considera el juego de datos original y el juego de datos modificado, marcándolos con subíndice una "o" o una "m".

x_{o0j} = caso j-ésimo perteneciente a la clase 0, del juego de datos original.

x_{o1j} = caso j-ésimo perteneciente a la clase 1, del juego de datos original.

x_{m0j} =caso j-ésimo perteneciente a la clase 0, del juego de datos modificado.

x_{m1j} =caso j-ésimo perteneciente a la clase 1, del juego de datos modificado.

$(x_{o0j})_i$ =valor de la variable i-ésima, del caso j-ésimo perteneciente a la clase 0, del juego de datos original.

$(x_{o1j})_i$ =valor de la variable i-ésima del caso j-ésimo perteneciente a la clase 1, del juego de datos original.

$(x_{m0j})_i$ =valor de la variable i-ésima del caso j-ésimo perteneciente a la clase 0, del juego de datos modificado.

$(x_{m1j})_i$ =valor de la variable i-ésima del caso j-ésimo perteneciente a la clase 1, del juego de datos modificado.

Se definen los vectores de medias del juego de datos originales:

$$\bar{x}_{o0} = \left(\left(\bar{x}_{o0}\right)_1, \left(\bar{x}_{o0}\right)_2, \ldots\ldots, \left(\bar{x}_{o0}\right)_p\right) = \bar{x} \text{ de la clase 0 del juego de datos original.}$$

$$\bar{x}_{o0} = \frac{1}{n_0}\sum_{j=1}^{n_0} x_{o0j}$$

Donde la posición i-ésima se refiere a la media de la variable x_i para los casos en que y=0 en el juego de datos original.

$$\bar{x}_{o1} = \left(\left(\bar{x}_{o1}\right)_1, \left(\bar{x}_{o1}\right)_2, \ldots\ldots, \left(\bar{x}_{o1}\right)_p\right) = \bar{x} \text{ de la clase 1 del juego de datos original.}$$

$$\bar{x}_{o1} = \frac{1}{n_1} \times \sum_{j=1}^{n_1} x_{o1j}$$

Donde la posición i-ésima se refiere a la media de la variable x_i para los casos en que y=1 en el juego de datos original.

Para resumir la notación tenemos:

n_0-c10 = cantidad de casos en donde la variable en la posición i vale 0, pertenecientes a la clase 0 = cantidad de 0 en la clase 0.

n_1-c11 = cantidad de casos en donde la variable en la posición i vale 0, pertenecientes a la clase 1 = cantidad de 0 en la clase 1.

cantidad de casos en los que la variable en la posición i vale 1, pertenecientes a la clase 1 = cantidad de 1 en la clase 1 = c11.
cantidad de casos en los que la variable en la posición i vale 1, pertenecientes a la clase 0 = cantidad de 1 en la clase 0 = c10.

Debido a que en las variables que permanecen sin cambio no cambia su media, entonces las siguientes definiciones son sólo para las variables que presentan cambios en la codificación.

Por lo tanto tenemos las medias de las variables dentro de cada clase:

$$\left(\bar{x}_{o0}\right)_i = \frac{cantidad\ de\ 1\ en\ la\ clase\ 0}{n_0} = \frac{c10}{n_0}$$

$$\left(\bar{x}_{o1}\right)_i = \frac{cantidad\ de\ 1\ en\ la\ clase\ 1}{n_1} = \frac{c11}{n_1}$$

En el juego de datos modificado se definen las medias de las variables:

Cuando se refiere a cantidad de 1 en la clase 0 o cantidad de 1 en la clase 1 se refiere al valor que correspondía antes de la modificación, pues es lo mismo que referirse a la cantidad de nv1 en la clase 0 y la cantidad de nv1 en la clase 1, porque son las mismas posiciones en la matriz de datos.

$\bar{x}_{m0} = \bar{x}$ de los casos pertenecientes a la clase 0, del juego de datos modificado.

$$\left(\bar{x}_{m0}\right)_i = \frac{\left(cantidad\ de\ 0\ en\ la\ clase\ 0 \times nv0\right) + \left(cantidad\ de\ 1\ en\ la\ clase\ 0 \times nv1\right)}{n_0} =$$

$$= \frac{\left(n_0 - cantidad\ de\ 1\ en\ la\ clase\ 0\right) \times nv0 + \left(cantidad\ de\ 1\ en\ la\ clase\ 0\right) \times nv1}{n_0} =$$

$$= \frac{n_0}{n_0} \times nv0 - \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 0\right)}{n_0} \times nv0 + \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 0\right)}{n_0} \times nv1 =$$

$$= nv0 + \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 0\right)}{n_0} \times \left(nv1 - nv0\right) = nv0 + \left(\bar{x}_{o0}\right)_i \times \left(nv1 - nv0\right)$$

$\bar{x}_{m1} = \bar{x}$ de los casos pertenecientes a la clase 1, del juego de datos modificado

$$\left(\bar{x}_{m1}\right)_i = \frac{\left(cantidad\ de\ 0\ en\ la\ clase\ 1 \times nv0\right) + \left(cantidad\ de\ 1\ en\ la\ clase\ 1 \times nv1\right)}{n_1} =$$

$$= \frac{\left(n_1 - cantidad\ de\ 1\ en\ la\ clase\ 1\right) \times nv0 + \left(cantidad\ de\ 1\ en\ la\ clase\ 1\right) \times nv1}{n_1} =$$

$$= \frac{n_1}{n_1} \times nv0 - \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 1\right)}{n_1} \times nv0 + \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 1\right)}{n_1} \times nv1 =$$

$$= nv0 + \frac{\left(cantidad\ de\ 1\ en\ la\ clase\ 1\right)}{n_1} \times \left(nv1 - nv0\right) = nv0 + \left(\bar{x}_{o1}\right)_i \times \left(nv1 - nv0\right)$$

En caso en que se cambie solo los valores en los casos en los cuales vale 1 y permanezcan sin cambios los valores de los casos en los cuales vale 0, entonces queda:

$$\left(\bar{x}_{m0}\right)_i = nv0 + \left(\bar{x}_{o0}\right)_i \times \left(nv1 - nv0\right) = 0 + \left(\bar{x}_{o0}\right)_i \times \left(nv1 - 0\right) = \left(\bar{x}_{o0}\right)_i \times nv1$$
$$\left(\bar{x}_{m1}\right)_i = nv0 + \left(\bar{x}_{o1}\right)_i \times \left(nv1 - nv0\right) = 0 + \left(\bar{x}_{o1}\right)_i \times \left(nv1 - 0\right) = \left(\bar{x}_{o1}\right)_i \times nv1$$

Debido a que las medias de las variables que no fueron modificadas permanecen sin cambios, entonces tenemos los vectores de medias del juego de datos modificado:

$$\bar{x}_{m0} = \left(\left(\bar{x}_{o0}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1}, \left(\bar{x}_{m0}\right)_i, \left(\bar{x}_{o0}\right)_{i+1}, \ldots\ldots, \left(\bar{x}_{o0}\right)_p\right) = \bar{x}_m \text{ de la clase 0}$$
$$\bar{x}_{m0} = \left(\left(\bar{x}_{o0}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1},\ nv0 + \left(\bar{x}_{o0}\right)_i \times \left(nv1 - nv0\right),\ \left(\bar{x}_{o0}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p\right),$$

$$\bar{x}_{m1} = \left(\left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o1}\right)_{i-1}, \left(\bar{x}_{m1}\right)_i, \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o1}\right)_p\right) = \bar{x}_m \text{ de la clase 1}$$
$$\bar{x}_{m1} = \left(\left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o1}\right)_{i-1},\ nv0 + \left(\bar{x}_{o1}\right)_i \times \left(nv1 - nv0\right),\ \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o1}\right)_p\right)$$

Se definen las diferencias de vectores de medias:

Diferencia de medias del juego de datos originales:

$$\bar{x}_{o0} - \bar{x}_{o1} = \left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots\ldots, \left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i, \ldots\ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p\right)$$

En el vector de diferencias de medias del juego de datos modificado solo se modifica la posición de la variable que fue modificada:

$$\left(\bar{x}_{m0}\right)_i - \left(\bar{x}_{m1}\right)_i = nv0 + \left(\bar{x}_{o0}\right)_i \times \left(nv1 - nv0\right) - \left[nv0 + \left(\bar{x}_{o1}\right)_i \times \left(nv1 - nv0\right)\right] =$$
$$= \left(\bar{x}_{o0}\right)_i \times \left(nv1 - nv0\right) - \left[\left(\bar{x}_{o1}\right)_i \times \left(nv1 - nv0\right)\right] = \left(nv1 - nv0\right) \times \left[\left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i\right]$$

La diferencia de vectores de medias de las clases:

$$\bar{x}_{m0} - \bar{x}_{m1} = \left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} - \left(\bar{x}_{o1}\right)_{i-1}, \left(\bar{x}_{m0}\right)_i - \left(\bar{x}_{m1}\right)_i, \left(\bar{x}_{o0}\right)_{i+1} - \left(\bar{x}_{o1}\right)_{i+1}, \ldots \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p\right) =$$
$$= \left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} - \left(\bar{x}_{o1}\right)_{i-1}, (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i\right], \left(\bar{x}_{o0}\right)_{i+1} - \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p\right)$$

La suma de los vectores de medias de las clases es utilizado para calcular el punto de corte:

$$\left(\bar{x}_{m0}\right)_i + \left(\bar{x}_{m1}\right)_i = nv0 + \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0) + \left[nv0 + \left(\bar{x}_{o1}\right)_i \times (nv1 - nv0)\right] =$$
$$= 2 \times nv0 + \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0) + \left[\left(\bar{x}_{o1}\right)_i \times (nv1 - nv0)\right] = 2nv0 + (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right]$$

$$\bar{x}_{m0} + \bar{x}_{m1} = \left(\left(\bar{x}_{o0}\right)_1 + \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} + \left(\bar{x}_{o1}\right)_{i-1}, 2 \times nv0 + (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right], \left(\bar{x}_{o0}\right)_{i+1} + \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p + \left(\bar{x}_{o1}\right)_p\right)$$

Para ver que pasa con S_c del juego de datos modificado, llamamos S_{oc} al S_c del juego de datos originales y S_{mc} al S_c del juego de datos modificado.

Si la variable dummy modificada es la variable x_i, por un lado las posiciones en donde no interviene la variable x_i son iguales a S_c del juego de datos originales.

$S_{mc}(l, j) = S_{oc}(l, j)$ donde $l \neq i$, $j \neq i$, ni x_j, ni x_l fueron modificadas.

La varianza de la variable x_i de la clase 0:

$$S_{o0}(i,i) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left[\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i\right] \times \left[\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i\right] =$$
$$= \frac{1}{n_0 - 1} \times \left\{\left[(n_0 - c10) \times \left(0 - \left(\bar{x}_{o0}\right)_i\right)^2\right] + \left[c10 \times \left(1 - \left(\bar{x}_{o0}\right)_i\right)^2\right]\right\}$$

$$S_{m0}(i,i) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i\right] \times \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i\right]$$

$\left(x_{m0j}\right)_i = nv0$, ocurre en los mismos casos pertenecientes a la clase 0, donde la variable x_i vale 0, en el juego de datos original.

$\left(x_{m0j}\right)_i = nv1$, ocurre en los mismos casos pertenecientes a la clase 0, donde la variable x_i vale 1, en el juego de datos original.

$$Si\ \left(x_{m0j}\right)_i = nv0 \Rightarrow \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i\right] = \left[nv0 - nv0 - \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0)\right] =$$
$$= -\left(\bar{x}_{o0}\right)_i \times (nv1 - nv0)$$

$$Si\ \left(x_{m0j}\right)_i = nv1 \Rightarrow \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i\right] = \left[nv1 - nv0 - \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0)\right] =$$
$$= \left[1 - \left(\bar{x}_{o0}\right)_i\right] \times (nv1 - nv0)$$

$$S_{m0}(i,i) = \frac{1}{n_0 - 1} \times \left\{ \left[(n_0 - c10) \times \left(\bar{x}_{o0}\right)_i^2 \times (nv1 - nv0)^2 \right] + \left[c10 \times \left(1 - \left(\bar{x}_{o0}\right)_i\right)^2 \times (nv1 - nv0)^2 \right] \right\} =$$

$$= (nv1 - nv0)^2 \times \frac{1}{n_0 - 1} \times \left\{ \left[(n_0 - c10) \times \left(\bar{x}_{o0}\right)_i^2 \right] + \left[c10 \times \left(1 - \left(\bar{x}_{o0}\right)_i\right)^2 \right] \right\} = (nv1 - nv0)^2 \times S_{o0}(i,i)$$

Las covarianzas donde interviene la variable x_i de la clase 0:

$$S_{o0}(i,t) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left[\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right] \times \left[\left(x_{o0j}\right)_t - \left(\bar{x}_{o0}\right)_t \right]$$

$$S_{m0}(i,t) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i \right] \times \left[\left(x_{m0j}\right)_t - \left(\bar{x}_{m0}\right)_t \right]$$

$$Si \left(x_{m0j}\right)_i = nv0 \Rightarrow \left(x_{o0j}\right)_i = 0 \Rightarrow \left(\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right) = 0 - \left(\bar{x}_{o0}\right)_i$$
$$\left(\left(x_{m0j}\right)_i - \left(\bar{x}_{mo}\right)_i \right) = \left(nv0 - nv0 - \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0) \right) = -\left(\bar{x}_{o0}\right)_i \times (nv1 - nv0)$$

y

$$Si \left(x_{m0j}\right)_i = nv1 \Rightarrow \left(x_{o0j}\right)_i = 1 \Rightarrow \left(\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right) = 1 - \left(\bar{x}_{o0}\right)_i$$
$$\left(\left(x_{m0j}\right)_i - \left(\bar{x}_{mo}\right)_i \right) = \left(nv1 - nv0 - \left(\bar{x}_{o0}\right)_i \times (nv1 - nv0) \right) = \left(1 - \left(\bar{x}_{o0}\right)_i\right) \times (nv1 - nv0)$$

Entonces $\forall j : \left(\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i \right) = (nv1 - nv0) \times \left(\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right)$

Por lo tanto:

$$S_{m0}(i,t) = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left[\left(x_{m0j}\right)_i - \left(\bar{x}_{m0}\right)_i \right] \times \left[\left(x_{m0j}\right)_t - \left(\bar{x}_{m0}\right)_t \right] =$$

$$= \frac{1}{n_0 - 1} \times \sum_{j=1}^{n_0} (nv1 - nv0) \times \left\{ \left[\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right] \times \left[\left(x_{o0j}\right)_t - \left(\bar{x}_{o0}\right)_t \right] \right\} =$$

$$= (nv1 - nv0) \times \frac{1}{n_0 - 1} \times \sum_{j=1}^{n_0} \left\{ \left[\left(x_{o0j}\right)_i - \left(\bar{x}_{o0}\right)_i \right] \times \left[\left(x_{o0j}\right)_t - \left(\bar{x}_{o0}\right)_t \right] \right\} =$$

$$= (nv1 - nv0) \times S_{o0}(i,t)$$

La matriz de varianzas-covarianzas de la clase 1 tiene el mismo desarrollo que los vistos para la clase 0, quedando:

$$S_{m1}(i,i) = (nv1 - nv0)^2 \times S_{o1}(i,i)$$

$$S_{m1}(i,t) = (nv1 - nv0) \times S_{o1}(i,t)$$

La matriz de varianzas-covarianzas de la clase 0:

$$S_{m0} = \begin{bmatrix} (S_{o0})_{11} & \dots & (S_{o0})_{1i} \times (nv1 - nv0) & \dots & (S_{o0})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{i1} \times (nv1 - nv0) & \dots & (S_{o0})_{ii} \times (nv1 - nv0)^2 & \dots & (S_{o0})_{ip} \times (nv1 - nv0) \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{p1} & \dots & (S_{o0})_{pi} \times (nv1 - nv0) & \dots & (S_{o0})_{pp} \end{bmatrix}$$

La matriz de varianzas-covarianzas de la clase 1:

$$S_{m1} = \begin{bmatrix} (S_{o1})_{11} & \dots & (S_{o1})_{1i} \times (nv1 - nv0) & \dots & (S_{o1})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{i1} \times (nv1 - nv0) & \dots & (S_{o1})_{ii} \times (nv1 - nv0)^2 & \dots & (S_{o1})_{ip} \times (nv1 - nv0) \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{p1} & \dots & (S_{o1})_{pi} \times (nv1 - nv0) & \dots & (S_{o1})_{pp} \end{bmatrix}$$

La matriz de varianzas-covarianzas combinada del juego de datos original se define de la siguiente forma:

$$S_{oc} = \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right] S_{o0} + \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right] S_{o1}$$

Entonces S combinado del juego de datos modificado queda conformado como sigue:

$$S_{mc} = \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right] S_{m0} + \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right] S_{m1} =$$

$$\left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \begin{bmatrix} (S_{o0})_{11} & \dots & (S_{o0})_{1i} \times (nv1 - nv0) & \dots & (S_{o0})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{i1} \times (nv1 - nv0) & \dots & (S_{o0})_{ii} \times (nv1 - nv0)^2 & \dots & (S_{o0})_{ip} \times (nv1 - nv0) \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{p1} & \dots & (S_{o0})_{pi} \times (nv1 - nv0) & \dots & (S_{o0})_{pp} \end{bmatrix} +$$

$$+ \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \begin{bmatrix} (S_{o1})_{11} & \dots & (S_{o1})_{1i} \times (nv1 - nv0) & \dots & (S_{o1})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{i1} \times (nv1 - nv0) & \dots & (S_{o1})_{ii} \times (nv1 - nv0)^2 & \dots & (S_{o1})_{ip} \times (nv1 - nv0) \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{p1} & \dots & (S_{o1})_{pi} \times (nv1 - nv0) & \dots & (S_{o1})_{pp} \end{bmatrix} =$$

$$= \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (S_{o0})_{11} & \dots & (S_{o0})_{1i} & \dots & (S_{o0})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{i1} & \dots & (S_{o0})_{ii} & \dots & (S_{o0})_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{p1} & \dots & (S_{o0})_{pi} & \dots & (S_{o0})_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} +$$

$$+ \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (S_{o1})_{11} & \dots & (S_{o1})_{1i} & \dots & (S_{o1})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{i1} & \dots & (S_{o1})_{ii} & \dots & (S_{o1})_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{p1} & \dots & (S_{o1})_{pi} & \dots & (S_{o1})_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} =$$

$\left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right]$ y $\left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right]$ son constantes, entonces:

$$= \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \begin{bmatrix} (S_{o0})_{11} & \dots & (S_{o0})_{1i} & \dots & (S_{o0})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{i1} & \dots & (S_{o0})_{ii} & \dots & (S_{o0})_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o0})_{p1} & \dots & (S_{o0})_{pi} & \dots & (S_{o0})_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} +$$

$$+ \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)}\right] \times \begin{bmatrix} (S_{o1})_{11} & \dots & (S_{o1})_{1i} & \dots & (S_{o1})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{i1} & \dots & (S_{o1})_{ii} & \dots & (S_{o1})_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{o1})_{p1} & \dots & (S_{o1})_{pi} & \dots & (S_{o1})_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} =$$

$$= \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)} \right] \times S_{o0} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} +$$

$$+ \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)} \right] \times S_{o1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} =$$

$$= \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \left[\frac{n_0 - 1}{(n_0 - 1) + (n_1 - 1)} \right] \times S_{o0} + \times \left[\frac{n_1 - 1}{(n_0 - 1) + (n_1 - 1)} \right] \times S_{o1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} =$$

$$= \left\{ \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times S_{oc} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (nv1 - nv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \right\} =$$

$$= \begin{bmatrix} (S_{oc})_{11} & \dots & (S_{oc})_{1i} \times (nv1 - nv0) & \dots & (S_{oc})_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ (S_{oc})_{i1} \times (nv1 - nv0) & \dots & (S_{oc})_{ii} \times (nv1 - nv0)^2 & \dots & (S_{oc})_{ip} \times (nv1 - nv0) \\ \dots & \dots & \dots & \dots & \dots \\ (S_{oc})_{p1} & \dots & (S_{oc})_{pi} \times (nv1 - nv0) & \dots & (S_{oc})_{pp} \end{bmatrix}$$

7.2.2 Inversa de la Matriz de Varianzas-Covarianzas Combinada

Si en la variable x_i en lugar de usar 1 y 0 se usan dos números mv1 y mv0 (en lugar del 1 y del 0 respectivamente), entonces la matriz S combinado llamada S tiene la siguiente relación con la matriz S combinado original (usando 1 y0):

Se define la matriz S combinado de los datos originales como sigue:

$$S_o = \begin{bmatrix} (s_o)_{11} & ..... & (s_o)_{1i} & ..... & (s_o)_{1p} \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{i1} & ..... & (s_o)_{ii} & ..... & (s_o)_{ip} \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{p1} & ..... & (s_o)_{pi} & ..... & (s_o)_{pp} \end{bmatrix}$$

Entonces la matriz S combinado de los datos modificados queda como sigue:

$$S_m = \begin{bmatrix} (s_o)_{11} & ..... & (s_o)_{1i} \times (mv1 - mv0) & ..... & (s_o)_{1p} \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{i1} \times (mv1 - mv0) & ..... & (s_o)_{ii} \times (mv1 - mv0)^2 & ..... & (s_o)_{ip} \times (mv1 - mv0) \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{p1} & ..... & (s_o)_{pi} \times (mv1 - mv0) & ..... & (s_o)_{pp} \end{bmatrix}$$

Esta matriz es igual a la siguiente multiplicación de matrices:

$$S_m = \begin{bmatrix} 1 & ... & 0 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & 1 & 0 & 0 & ... & 0 \\ 0 & ... & 0 & (mv1 - mv0) & 0 & ... & 0 \\ 0 & ... & 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & 0 & 0 & 0 & ... & 1 \end{bmatrix} \times \begin{bmatrix} (s_o)_{11} & ..... & (s_o)_{1i} & ..... & (s_o)_{1p} \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{i1} & ..... & (s_o)_{ii} & ..... & (s_o)_{ip} \\ .... & ..... & ..... & ..... & ..... \\ (s_o)_{p1} & ..... & (s_o)_{pi} & ..... & (s_o)_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & ... & 0 & 0 & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & 1 & 0 & 0 & ... & 0 \\ 0 & ... & 0 & (mv1 - mv0) & 0 & ... & 0 \\ 0 & ... & 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ 0 & ... & 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

$$S_m^{-1} = \left[\begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (mv1 - mv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (s_o)_{11} & \dots\dots & (s_o)_{1i} & \dots\dots & (s_o)_{1p} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{i1} & \dots\dots & (s_o)_{ii} & \dots\dots & (s_o)_{ip} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{p1} & \dots\dots & (s_o)_{pi} & \dots\dots & (s_o)_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (mv1 - mv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}\right]^{-1} =$$

Son tres matrices de p×p, en donde p es la cantidad de variables y la posición que tiene el valor =mv1-mv0 es la posición (i,i). Para j≠i las posiciones (j,j) valen 1.

$$S_m^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (mv1 - mv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}^{-1} \times \left[\begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (mv1 - mv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (s_o)_{11} & \dots\dots & (s_o)_{1i} & \dots\dots & (s_o)_{1p} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{i1} & \dots\dots & (s_o)_{ii} & \dots\dots & (s_o)_{ip} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{p1} & \dots\dots & (s_o)_{pi} & \dots\dots & (s_o)_{pp} \end{bmatrix}\right]^{-1} =$$

$$S_m^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \dfrac{1}{(mv1 - mv0)} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (s_o)_{11} & \dots\dots & (s_o)_{1i} & \dots\dots & (s_o)_{1p} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{i1} & \dots\dots & (s_o)_{ii} & \dots\dots & (s_o)_{ip} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{p1} & \dots\dots & (s_o)_{pi} & \dots\dots & (s_o)_{pp} \end{bmatrix}^{-1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & (mv1 - mv0) & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}^{-1}$$

$$S_m^{-1} = \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \dfrac{1}{(mv1 - mv0)} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (s_o)_{11} & \dots\dots & (s_o)_{1i} & \dots\dots & (s_o)_{1p} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{i1} & \dots\dots & (s_o)_{ii} & \dots\dots & (s_o)_{ip} \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots\dots \\ (s_o)_{p1} & \dots\dots & (s_o)_{pi} & \dots\dots & (s_o)_{pp} \end{bmatrix}^{-1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \dfrac{1}{(mv1 - mv0)} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Entonces la matriz del medio es la inversa de S combinado del juego de datos originales. Si a cada posición (i,j) de S_o^{-1} se nombra como $(s_o^{-1})_{ij}$ entonces queda como sigue:

$$S_m^{-1} = \begin{bmatrix} (s_o^{-1})_{11} & \cdots & (s_0^{-1})_{1i} & \cdots & (s_o^{-1})_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{(s_o^{-1})_{i1}}{(mv1 - mv0)} & \cdots & \frac{(s_o^{-1})_{ii}}{(mv1 - mv0)} & \cdots & \frac{(s_o^{-1})_{ip}}{(mv1 - mv0)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (s_o^{-1})_{p1} & \cdots & (s_0^{-1})_{pi} & \cdots & (s_o^{-1})_{pp} \end{bmatrix} \times \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{(mv1 - mv0)} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} =$$

$$S_m^{-1} = \begin{bmatrix} (s_o^{-1})_{11} & \cdots & \frac{(s_o^{-1})_{1i}}{(mv1 - mv0)} & \cdots & (s_o^{-1})_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{(s_o^{-1})_{i1}}{(mv1 - mv0)} & \cdots & \frac{(s_o^{-1})_{ii}}{(mv1 - mv0)^2} & \cdots & \frac{(s_o^{-1})_{ip}}{(mv1 - mv0)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (s_o^{-1})_{p1} & \cdots & \frac{(s_o^{-1})_{pi}}{(mv1 - mv0)} & \cdots & (s_o^{-1})_{pp} \end{bmatrix}$$

7.2.3 Estimación del punto de corte y de los casos

Al Punto de Corte lo llamamos m.

m_o=punto de corte en el juego de datos original.

m_m=punto de corte el el juego de datos modificado.

Para las estimaciones de los casos se define:

x_{oj} =caso j-ésimo del juego de datos original, no se diferencia aquí a cual clase pertenece.

$(x_{oj})_i$ =valor de la variable i-ésima del caso j-ésimo del juego de datos original, no se diferencia aquí a cual clase pertenece.

x_{mj} =caso j-ésimo del juego de datos modificado, aquí no se diferencia a cual clase pertenece.

$(x_{mj})_i$ =valor de la variable i-ésima del caso j-ésimo del juego de datos modificado, aquí no se diferencia a cual clase pertenece.

$(\hat{y}_o)_j$ =estimación del caso j-ésimo del juego de datos original, aquí no se diferencia a cual clase pertenece.

$(\hat{y}_m)_j$ =estimación del caso j-ésimo del juego de datos modificado, aquí no se diferencia a cual clase pertenece.

$(S_{oc}^{-1})_i$ =columna i-ésima de la matriz inversa de varianzas-covarianzas combinada.

Estimación del punto de corte

$$\hat{m}_o = \frac{1}{2} \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \left(\bar{x}_{o0} + \bar{x}_{o1}\right)$$

$$\hat{m}_m = \frac{1}{2} \times \left(\bar{x}_{m0} - \bar{x}_{m1}\right)^T \times S_{mc}^{-1} \times \left(\bar{x}_{m0} + \bar{x}_{m1}\right) =$$

$$= \frac{1}{2} \times \left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} - \left(\bar{x}_{o1}\right)_{i-1},\ (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i\right],\ \left(\bar{x}_{o0}\right)_{i+1} - \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p \right) \times$$

$$\times \left\{ \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \frac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times S_{oc}^{-1} \times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \frac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \right\} \times$$

$$\times \left(\left(\bar{x}_{o0}\right)_1 + \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} + \left(\bar{x}_{o1}\right)_{i-1},\ 2 \times nv0 + (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right],\ \left(\bar{x}_{o0}\right)_{i+1} + \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p + \left(\bar{x}_{o1}\right)_p \right)^T =$$

$$= \frac{1}{2} \times \left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i, \ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p \right) \times S_{oc}^{-1} \times \left(\left(\bar{x}_{o0}\right)_1 + \left(\bar{x}_{o1}\right)_1, \ldots, 2 \times \frac{nv0}{nv1 - nv0} + \left(\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right), \ldots, \left(\bar{x}_{o0}\right)_p + \left(\bar{x}_{o1}\right)_p \right)^T =$$

$$= \frac{1}{2} \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right) \times S_{oc}^{-1} \times \left(\left(\bar{x}_{o0}\right)_1 + \left(\bar{x}_{o1}\right)_1, \ldots, 2 \times \frac{nv0}{nv1 - nv0} + \left(\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right), \ldots, \left(\bar{x}_{o0}\right)_p + \left(\bar{x}_{o1}\right)_p \right)^T =$$

$$= \frac{1}{2} \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right) \times S_{oc}^{-1} \times \left(\left(\bar{x}_{o0}\right)_1 + \left(\bar{x}_{o1}\right)_1, \ldots, \left(\left(\bar{x}_{o0}\right)_i + \left(\bar{x}_{o1}\right)_i\right), \ldots, \left(\bar{x}_{o0}\right)_p + \left(\bar{x}_{o1}\right)_p \right)^T + \left[\frac{1}{2} \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right) \times \left(S_{oc}^{-1}\right) \times \left(0, \ldots, 2 \times \frac{nv0}{(nv1 - nv0)}, \ldots, 0 \right)^T \right] =$$

$$= \hat{m}_o + \left[\left(\bar{x}_{o0} - \bar{x}_{o1}\right) \times \left(S_{oc}^{-1}\right) \times \left(0, \ldots, \frac{nv0}{(nv1 - nv0)}, \ldots, 0 \right)^T \right] = \hat{m}_o + \left[\left(\bar{x}_{o0} - \bar{x}_{o1}\right) \times \left(S_{oc}^{-1}\right)_i \times \frac{nv0}{(nv1 - nv0)} \right]$$

Estimaciones de los casos

Estimación del caso j-ésimo en el juego de datos original:

$$\left(\hat{y}_o\right)_j = \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \left(x_{oj}\right)$$

Estimación del caso j-ésimo en el juego de datos original cuando $\left(x_{oj}\right)_i = 0$:

$$\left(\hat{y}_o\right)_j = \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \ldots\ldots \\ \left(x_{oj}\right)_i = 0 \\ \ldots\ldots \\ \left(x_{oj}\right)_p \end{bmatrix}$$

Estimación del caso j-ésimo en el juego de datos original cuando $\left(x_{oj}\right)_i = 1$:

$$\left(\hat{y}_o\right)_j = \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \ldots\ldots \\ \left(x_{oj}\right)_i = 1 \\ \ldots\ldots \\ \left(x_{oj}\right)_p \end{bmatrix}$$

Estimación del caso j-ésimo en el juego de datos modificado cuando $(x_{mj})_i = nv0$ y corresponde con $(x_{oj})_i = 0$ en el juego de datos original:

$$(\hat{y}_m)_j = (\bar{x}_{m0} - \bar{x}_{m1})^T \times S_{mc}^{-1} \times (x_{mj}) =$$

$$= (\bar{x}_{o0} - \bar{x}_{o1})^T \times S_{oc}^{-1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \frac{1}{nv1 - nv0} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} (x_{oj})_1 \\ \dots \\ nv0 \\ \dots \\ (x_{oj})_p \end{bmatrix} =$$

$$= (\bar{x}_{o0} - \bar{x}_{o1})^T \times S_{oc}^{-1} \times \begin{bmatrix} (x_{oj})_1 \\ \dots \\ \frac{nv0}{nv1 - nv0} \\ \dots \\ (x_{oj})_p \end{bmatrix} =$$

$$= (\bar{x}_{o0} - \bar{x}_{o1})^T \times S_{oc}^{-1} \times \begin{bmatrix} (x_{oj})_1 \\ \dots \\ 0 \\ \dots \\ (x_{oj})_p \end{bmatrix} + \left\{ (\bar{x}_{o0} - \bar{x}_{o1})^T \times S_{oc}^{-1} \times \begin{bmatrix} 0 \\ \dots \\ \frac{nv0}{nv1 - nv0} \\ \dots \\ 0 \end{bmatrix} \right\} =$$

$$= (\hat{y}_o)_j + \left\{ (\bar{x}_{o0} - \bar{x}_{o1})^T \times S_{oc}^{-1} \times \begin{bmatrix} 0 \\ \dots \\ \frac{nv0}{nv1 - nv0} \\ \dots \\ 0 \end{bmatrix} \right\} = (\hat{y}_o)_j + \left[(\bar{x}_{o0} - \bar{x}_{o1})^T \times (S_{oc}^{-1})_i \times \frac{nv0}{nv1 - nv0} \right]$$

Estimación del caso j-ésimo en el juego de datos modificado cuando $\left(x_{mj}\right)_i = nv1$ y corresponde con $\left(x_{oj}\right)_i = 1$ en el juego de datos original:

$$\left(\hat{y}_m\right)_j = \left(\bar{x}_{m0} - \bar{x}_{m1}\right)^T \times S_{mc}^{-1} \times \left(x_{mj}\right) =$$

$$= \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \dots \\ nv1 \\ \dots \\ \left(x_{oj}\right)_p \end{bmatrix} =$$

$$= \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \dots \\ \dfrac{nv1}{nv1 - nv0} \\ \dots \\ \left(x_{oj}\right)_p \end{bmatrix} = \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \dots \\ \left(\dfrac{nv1 - nv0}{nv1 - nv0}\right) + \left(\dfrac{nv0}{nv1 - nv0}\right) \\ \dots \\ \left(x_{oj}\right)_p \end{bmatrix} =$$

$$= \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} \left(x_{oj}\right)_1 \\ \dots \\ 1 \\ \dots \\ \left(x_{oj}\right)_p \end{bmatrix} + \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} 0 \\ \dots \\ \dfrac{nv0}{nv1 - nv0} \\ \dots \\ 0 \end{bmatrix} =$$

$$= \left(\hat{y}_o\right)_j + \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \begin{bmatrix} 0 \\ \dots \\ \dfrac{nv0}{nv1 - nv0} \\ \dots \\ 0 \end{bmatrix} = \left(\hat{y}_o\right)_j + \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times \left(S_{oc}^{-1}\right)_i \times \left(\frac{nv0}{nv1 - nv0}\right)$$

Relación entre la estimación del punto de corte (m) y la estimación del caso j-ésimo en el juego de datos original y en el juego de datos modificado:

$$\hat{m}_o > (\hat{y}_o)_j \Leftrightarrow \hat{m}_o + \left[(\bar{x}_{o0} - \bar{x}_{o1})^T \times (S_{oc}^{-1})_i \times \frac{nv0}{nv1 - nv0} \right] > (\hat{y}_o)_j + \left[(\bar{x}_{o0} - \bar{x}_{o1})^T \times (S_{oc}^{-1})_i \times \frac{nv0}{nv1 - nv0} \right] \Leftrightarrow$$
$$\Leftrightarrow \hat{m}_m > (\hat{y}_m)_j$$

$$\hat{m}_o \leq (\hat{y}_o)_j \Leftrightarrow \hat{m}_o + \left[(\bar{x}_{o0} - \bar{x}_{o1})^T \times (S_{oc}^{-1})_i \times \frac{nv0}{nv1 - nv0} \right] \leq (\hat{y}_o)_j + \left[(\bar{x}_{o0} - \bar{x}_{o1})^T \times (S_{oc}^{-1})_i \times \frac{nv0}{nv1 - nv0} \right] \Leftrightarrow$$
$$\Leftrightarrow \hat{m}_m \leq (\hat{y}_m)_j$$

Por lo tanto la clasificaciones de los casos no cambian cuando se cambia la codificación.

Así como la clasificación de los casos al usar 0 y 1 en la variable binaria x_i o al usar nv0 y nv1 no cambia entonces tampoco cambian las clasificaciones de casos entre usar 0 y 1 o usar v0 y v1 donde v0≠nv0, v1≠nv1. Si llamamos a las estimaciones de el punto de corte (m) y de los casos en el juego de datos con valores v0 y v1 con $\hat{y}_v, \hat{m}_v$, entonces se cumple lo siguiente:

$$(\hat{y}_o)_j \geq \hat{m}_o \Leftrightarrow (\hat{y}_m)_j \geq \hat{m}_m \text{ y } (\hat{y}_o)_j \geq \hat{m}_o \Leftrightarrow (\hat{y}_v)_j \geq \hat{m}_v$$

$$(\hat{y}_m)_j \geq \hat{m}_m \Leftrightarrow (\hat{y}_o)_j \geq \hat{m}_o \Leftrightarrow (\hat{y}_v) \geq \hat{m}_v$$

$$(\hat{y}_o)_j < \hat{m}_o \Leftrightarrow (\hat{y}_m)_j < \hat{m}_m \text{ y } (\hat{y}_o)_j < \hat{m}_o \Leftrightarrow (\hat{y}_v)_j < \hat{m}_v$$

$$(\hat{y}_m)_j < \hat{m}_m \Leftrightarrow (\hat{y}_o)_j < \hat{m}_o \Leftrightarrow (\hat{y}_v) < \hat{m}_v$$

Entonces cuando clasifica un caso dentro de la clase 0 en el juego de datos con valores nv0 y nv1 también clasifica al mismo caso dentro de la clase 0 en el juego de datos con los valores v0 y v1, viceversa. También cuando clasifica un caso dentro de la clase 1 en el juego de datos con los valores nv0 y nv1 también lo clasifica dentro de la clase 1 en el juego de datos con los valores v0 y v1, y viceversa. Entonces para cualquier par de valores con los que se codifique una variable binaria siempre da la misma codificación.

Si se quiere ver que pasa al cambiar el código de más de una variable dummy, entonces la demostración se hace para el cambio en una variable y luego el juego de datos modificado se usa como juego de datos original, es decir la matriz de varianzas-covarianzas combinada, los vectores de medias y los casos del juego de datos modificado pasan a ser los del juego de datos original para el próximo paso, que es la modificación de la otra variable, se cumple la misma demostración, tampoco cambia las clasificaciones de los casos. Lo mismo ocurre para cualquier cantidad de cambios de código.

7.2.4 Suma de Cuadrados Entre

Se define:

SCE_0=Suma de Cuadrados Entre del juego de datos original.
SCE_m=Suma de Cuadrados Entre del juego de datos modificado.
$\overline{\overline{x_o}} = \overline{x_o}$ total = media total del juego de datos original.
$\overline{\overline{x_m}} = \overline{x_m}$ total = media total del juego de datos modificado.

$$SCE_o = n_0 \times \left(\overline{x_{o0}} - \overline{\overline{x_o}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o0}} - \overline{\overline{x_o}}\right) + n_1 \times \left(\overline{x_{o1}} - \overline{\overline{x_o}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o1}} - \overline{\overline{x_o}}\right)$$

$$\left(\overline{x_{o0}} - \overline{\overline{x_o}}\right) = \left(\frac{n_1}{n_0 + n_1}\right)\left(\overline{x_{o0}} - \overline{x_{o1}}\right)$$

$$\left(\overline{x_{o1}} - \overline{\overline{x_o}}\right) = \left(\frac{n_0}{n_0 + n_1}\right)\left(\overline{x_{o1}} - \overline{x_{o0}}\right)$$

$$n_0 \times \left(\overline{x_{o0}} - \overline{\overline{x_o}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o0}} - \overline{\overline{x_o}}\right) = n_0 \times \left(\frac{n_1}{n_0 + n_1}\right)^2 \times \left(\overline{x_{o0}} - \overline{x_{o1}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o0}} - \overline{x_{o1}}\right)$$

$$n_1 \times \left(\overline{x_{o1}} - \overline{\overline{x_o}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o1}} - \overline{\overline{x_o}}\right) = n_1 \times \left(\frac{n_0}{n_0 + n_1}\right)^2 \times \left(\overline{x_{o1}} - \overline{x_{o0}}\right)^T \times S_{oc}^{-1} \times \left(\overline{x_{o1}} - \overline{x_{o0}}\right)$$

$$SCE_m = n_0 \times \left(\overline{x_{m0}} - \overline{\overline{x_m}}\right)^T \times S_{mc}^{-1} \times \left(\overline{x_{m0}} - \overline{\overline{x_m}}\right) + n_1 \times \left(\overline{x_{m1}} - \overline{\overline{x_m}}\right)^T \times S_{mc}^{-1} \times \left(\overline{x_{m1}} - \overline{\overline{x_m}}\right)$$

$$\left(\overline{x_{m0}} - \overline{\overline{x_m}}\right) = \left(\frac{n_1}{n_0 + n_1}\right)\left(\overline{x_{m0}} - \overline{x_{m1}}\right) =$$

$$= \left(\frac{n_1}{n_0 + n_1}\right)\left(\left(\overline{x_{o0}}\right)_1 - \left(\overline{x_{o1}}\right)_1, \ldots, \left(\overline{x_{o0}}\right)_{i-1} - \left(\overline{x_{o1}}\right)_{i-1}, \ (nv1 - nv0) \times \left[\left(\overline{x_{o0}}\right)_i - \left(\overline{x_{o1}}\right)_i\right], \ \left(\overline{x_{o0}}\right)_{i+1} - \left(\overline{x_{o1}}\right)_{i+1}, \ldots, \left(\overline{x_{o0}}\right)_p - \left(\overline{x_{o1}}\right)_p\right)$$

$$\left(\overline{x_{m1}} - \overline{\overline{x_m}}\right) = \left(\frac{n_0}{n_0 + n_1}\right)\left(\overline{x_{m1}} - \overline{x_{m0}}\right) =$$

$$= \left(\frac{n_0}{n_0 + n_1}\right)\left(\left(\overline{x_{o1}}\right)_1 - \left(\overline{x_{o0}}\right)_1, \ldots, \left(\overline{x_{o1}}\right)_{i-1} - \left(\overline{x_{o0}}\right)_{i-1}, \ (nv1 - nv0) \times \left[\left(\overline{x_{o1}}\right)_i - \left(\overline{x_{o0}}\right)_i\right], \ \left(\overline{x_{o1}}\right)_{i+1} - \left(\overline{x_{o0}}\right)_{i+1}, \ldots, \left(\overline{x_{o1}}\right)_p - \left(\overline{x_{o0}}\right)_p\right)$$

$$n_0 \times \left(\bar{\bar{x}}_{m0} - \bar{\bar{x}}_m\right)^T \times S_{mc}^{-1} \times \left(\bar{\bar{x}}_{m0} - \bar{\bar{x}}_m\right) =$$

$$= n_0 \times \left(\frac{n_1}{n_0 + n_1}\right)\left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} - \left(\bar{x}_{o1}\right)_{i-1},\ (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i\right],\ \left(\bar{x}_{o0}\right)_{i+1} - \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p\right) \times$$

$$\times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times S_{oc}^{-1} \times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times$$

$$\times \left(\frac{n_1}{n_0 + n_1}\right)\left(\left(\bar{x}_{o0}\right)_1 - \left(\bar{x}_{o1}\right)_1, \ldots, \left(\bar{x}_{o0}\right)_{i-1} - \left(\bar{x}_{o1}\right)_{i-1},\ (nv1 - nv0) \times \left[\left(\bar{x}_{o0}\right)_i - \left(\bar{x}_{o1}\right)_i\right],\ \left(\bar{x}_{o0}\right)_{i+1} - \left(\bar{x}_{o1}\right)_{i+1}, \ldots, \left(\bar{x}_{o0}\right)_p - \left(\bar{x}_{o1}\right)_p\right)^T =$$

$$= n_0 \left(\frac{n_1}{n_0 + n_1}\right)^2 \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right)^T \times S_{oc}^{-1} \times \left(\bar{x}_{o0} - \bar{x}_{o1}\right) = n_0 \times \left(\bar{x}_{o0} - \bar{\bar{x}}_o\right)^T \times S_{oc}^{-1} \times \left(\bar{x}_{o0} - \bar{\bar{x}}_o\right)$$

$$n_1 \times \left(\bar{\bar{x}}_{m1} - \bar{\bar{x}}_m\right)^T \times S_{mc}^{-1} \times \left(\bar{\bar{x}}_{m1} - \bar{\bar{x}}_m\right) =$$

$$= \left(\frac{n_0}{n_0 + n_1}\right)\left(\left(\bar{x}_{o1}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots, \left(\bar{x}_{o1}\right)_{i-1} - \left(\bar{x}_{o0}\right)_{i-1},\ (nv1 - nv0) \times \left[\left(\bar{x}_{o1}\right)_i - \left(\bar{x}_{o0}\right)_i\right],\ \left(\bar{x}_{o1}\right)_{i+1} - \left(\bar{x}_{o0}\right)_{i+1}, \ldots, \left(\bar{x}_{o1}\right)_p - \left(\bar{x}_{o0}\right)_p\right) \times$$

$$\times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times S_{oc}^{-1} \times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} \times$$

$$\times \left(\frac{n_0}{n_0 + n_1}\right)\left(\left(\bar{x}_{o1}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots, \left(\bar{x}_{o1}\right)_{i-1} - \left(\bar{x}_{o0}\right)_{i-1},\ (nv1 - nv0) \times \left[\left(\bar{x}_{o1}\right)_i - \left(\bar{x}_{o0}\right)_i\right],\ \left(\bar{x}_{o1}\right)_{i+1} - \left(\bar{x}_{o0}\right)_{i+1}, \ldots, \left(\bar{x}_{o1}\right)_p - \left(\bar{x}_{o0}\right)_p\right)^T =$$

$$= n_1 \times \left(\frac{n_0}{n_0 + n_1}\right)^2 \times \left(\bar{x}_{o1} - \bar{x}_{o0}\right)^T \times S_{oc}^{-1} \times \left(\bar{x}_{o1} - \bar{x}_{o0}\right) = n_1 \times \left(\bar{x}_{o1} - \bar{\bar{x}}_o\right)^T \times S_{oc}^{-1} \times \left(\bar{x}_{o1} - \bar{\bar{x}}_o\right)$$

Entonces $SCE_m = SCE_o$

7.2.5 Suma de Cuadrados Dentro

La distancia de un caso j-ésimo, perteneciente a la clase 0, hacia la media de la clase 0, del juego de datos original:

$$\left(x_{o0j} - \bar{x}_{o0}\right)^T \times S_{oc}^{-1} \times \left(x_{o0j} - \bar{x}_{o0}\right)$$

En los casos en que la variable i-ésima vale 0:

$$\left(x_{o0j} - \bar{x}_{o0}\right) = \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\left(\bar{x}_{o0}\right)_i, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

En los casos en que la variable i-ésima vale 1:

$$\left(x_{o0j} - \bar{x}_{o0}\right) = \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, 1 - \left(\bar{x}_{o0}\right)_i \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

La distancia de el caso j-ésimo, perteneciente a la clase 0, del juego de datos modificado:

$$\left(x_{m0j} - \bar{x}_{m0}\right)^T \times S_{mc}^{-1} \times \left(x_{m0j} - \bar{x}_{m0}\right)$$

En los casos en que la variable i-ésima, en el juego de datos original, valía 0:

$$\left(x_{m0j} - \bar{x}_{m0}\right) = \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, nv0 - \left[nv0 + \bar{x}_{o0} \times \left(nv1 - nv0\right)\right], \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] =$$
$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\bar{x}_{o0} \times \left(nv1 - nv0\right)_i, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

$$\left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\bar{x}_{o0} \times \left(nv1 - nv0\right)_i, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] \times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} =$$

$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

$$\left(x_{m0j} - \bar{x}_{m0}\right)^T \times S_{mc}^{-1} \times \left(x_{m0j} - \bar{x}_{m0}\right) =$$
$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] \times S_{oc}^{-1} \times \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, -\bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]^T =$$

$$= \left(x_{o0j} - \bar{x}_{o0}\right)^T \times S_{oc}^{-1} \times \left(x_{o0j} - \bar{x}_{o0}\right) \text{ cuando en el juego de datos original, la variable i-ésima valía 0.}$$

En los casos en que la variable i-ésima en el juego de datos original, valía 1:

$$\left(x_{m0j} - \bar{x}_{m0}\right) = \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, nv1 - \left[nv0 + \bar{x}_{o0} \times (nv1 - nv0)\right], \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] =$$
$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, \ldots nv1 - nv0 - \bar{x}_{o0} \times (nv1 - nv0), \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] =$$
$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, \left(1 - \bar{x}_{o0}\right) \times (nv1 - nv0), \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

$$\left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, \left(1 - \bar{x}_{o0}\right) \times (nv1 - nv0)_i, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] \times \begin{bmatrix} 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \dfrac{1}{nv1 - nv0} & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & 0 & 0 & \ldots & 1 \end{bmatrix} =$$

$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, 1 - \bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]$$

$$\left(x_{m0j} - \bar{x}_{m0}\right)^T \times S_{mc}^{-1} \times \left(x_{m0j} - \bar{x}_{m0}\right) =$$
$$= \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, 1 - \bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right] \times S_{oc}^{-1} \times \left[\left(x_{o0j}\right)_1 - \left(\bar{x}_{o0}\right)_1, \ldots\ldots, 1 - \bar{x}_{o0i}, \ldots\ldots, \left(x_{o0j}\right)_p - \left(\bar{x}_{o0}\right)_p\right]^T =$$

$= \left(x_{o0j} - \bar{x}_{o0}\right)^T \times S_{oc}^{-1} \times \left(x_{o0j} - \bar{x}_{o0}\right)$ cuando en el juego de datos original, la variable i-ésima valía 1.

Por lo tanto, si en cada caso no cambia entonces $SCD_m = SCD_o$.

7.3 Cálculo de Suma de Cuadrados Dentro

n_1 = cantidad de casos pertenecientes a la clase 1.

n_2 = cantidad de casos pertenecientes a la clase 2.

$n = n_1 + n_2$ = cantidad total de casos.

x_{1i} = i – ésimo caso perteneciente a la clase 1.

$(x_{1i})_j$ = valor correspondiente a la variable en la posición j, del caso en posición i, perteneciente a la clase 1.

x_{2i} = i – ésimo caso perteneciente a la clase 2.

$(x_{1i})_j$ = valor correspondiente a la variable en la posición j, del caso en posición i, perteneciente a la clase 2.

$\overline{x_1} = \bar{x}$ de los casos pertenecientes a la clase 1.

$(\bar{x}_1)_j$ = valor correspondiente a la variable en la posición j, de $\bar{x}$ de los casos pertenecientes a la clase 1.

$\overline{x_2} = \bar{x}$ de los casos pertenecientes a la clase 2.

$(\bar{x}_2)_j$ = valor correspondiente a la variable en la posición j, de $\bar{x}$ de los casos pertenecientes a la clase 2.

p = cantidad de variables.

$$S_1 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \overline{x_1}) \times (x_{1i} - \overline{x_1})'}{n_1 - 1}, \quad S_2 = \frac{\sum_{i=1}^{n_2}(x_{2i} - \overline{x_2}) \times (x_{2i} - \overline{x_2})'}{n_2 - 1}$$

$$S_c = \frac{(n_1 - 1) \times S_1 + (n_2 - 1) \times S_2}{(n_1 - 1) + (n_2 - 1)}$$

$$Llamamos\ S = [(n_1 - 1) + (n_2 - 1)] \times S_c = \sum_{i=1}^{n_1}(x_{1i} - \overline{x_1}) \times (x_{2i} - \overline{x_2})' + \sum_{i=1}^{n_2}(x_{2i} - \overline{x_2}) \times (x_{2i} - \overline{x_2})'$$

$$S_c = \left[\frac{1}{(n_1 - 1) + (n_2 - 1)}\right] \times S \Rightarrow S_c^{-1} = [(n_1 - 1) + (n_2 - 1)] \times S^{-1}$$

$$(n_1 - 1)S_1 = \sum_{i=1}^{n_1}(x_{1i} - \overline{x_1}) \times (x_{1i} - \overline{x_1})' =$$

$$= \begin{bmatrix} \sum_{i=1}^{n_1}[(x_{1i})_1 - (\overline{x_1})_1]^2 & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_1 - (\overline{x_1})_1] \times [(x_{1i})_j - (\overline{x_1})_j] & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_1 - (\overline{x_1})_1] \times [(x_{1i})_p - (\overline{x_1})_p] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n_1}[(x_{1i})_j - (\overline{x_1})_j] \times [(x_{1i})_1 - (\overline{x_1})_1] & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_j - (\overline{x_1})_j]^2 & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_j - (\overline{x_1})_j] \times [(x_{1i})_p - (\overline{x_1})_p] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n_1}[(x_{1i})_p - (\overline{x_1})_p] \times [(x_{1i})_1 - (\overline{x_1})_1] & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_p - (\overline{x_1})_p] \times [(x_{1i})_j - (\overline{x_1})_j] & \cdots & \sum_{i=1}^{n_1}[(x_{1i})_p - (\overline{x_1})_p]^2 \end{bmatrix}$$

$$(n_2-1)S_2=\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)\times\left(x_{2i}-\overline{x_2}\right)^t=$$

$$=\begin{bmatrix}
\sum_{i=1}^{n_2}\left[(x_{2i})_1-(\overline{x_2})_1\right]^2 & \cdots & \sum_{i=1}^{n_1}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right] \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{i=1}^{n_1}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right] & \cdots & \sum_{i=1}^{n_2}\left[(x_{2i})_j-(\overline{x_2})_j\right]^2 & \cdots & \sum_{i=1}^{n_1}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right] \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{i=1}^{n_1}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right] & \cdots & \sum_{i=1}^{n_2}\left[(x_{2i})_p-(\overline{x_2})_p\right]^2
\end{bmatrix}$$

$$S=(n_1-1)\times S_1+(n_2-1)\times S_2=$$

$$=\begin{bmatrix}
\sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]^2+\sum_{i=1}^{n_2}\left[(x_{2i})_1-(\overline{x_2})_1\right]^2 & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right] \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]^2+\sum_{i=1}^{n_2}\left[(x_{2i})_j-(\overline{x_2})_j\right]^2 & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right] \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]+\sum_{i=1}^{n_1}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right] & \cdots & \sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]^2+\sum_{i=1}^{n_2}\left[(x_{2i})_p-(\overline{x_2})_p\right]^2
\end{bmatrix}$$

$$SCD1=\sum_{i=0}^{n_1}\left(x_{1i}-\overline{x_1}\right)^t S_c^{-1}\left(x_{1i}-\overline{x_1}\right)$$

$$SCD2=\sum_{i=0}^{n_2}\left(x_{2i}-\overline{x_2}\right)^t S_c^{-1}\left(x_{2i}-\overline{x_2}\right)$$

$$SCD=SCD_1+SCD_2=\sum_{i=1}^{n_1}\left(x_{1i}-\overline{x_1}\right)^t S_c^{-1}\left(x_{1i}-\overline{x_1}\right)+\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)^t S_c^{-1}\left(x_{2i}-\overline{x_2}\right)=$$

$$=\left[(n_1-1)+(n_2-1)\right]\times\left[\sum_{i=1}^{n_1}\left(x_{1i}-\overline{x_1}\right)^t S^{-1}\left(x_{1i}-\overline{x_1}\right)+\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)^t S^{-1}\left(x_{2i}-\overline{x_2}\right)\right]$$

Se ve el aporte de cada caso de la clase 1 x_{1i} a SCD_1: la multiplicación por $[(n_1-1)+(n_2-1)]$ se agrega al final.

$$\left(x_{1i}-\overline{x_1}\right)^t S^{-1}\left(x_{1i}-\overline{x_1}\right)=$$

$$=\left\{\begin{bmatrix}\left[(x_{1i})_1-(\overline{x_1})_1\right]\times(S^{-1})_{11}+ & \cdots & \left[(x_{1i})_1-(\overline{x_1})_1\right]\times(S^{-1})_{1j}+ & \cdots & \left[(x_{1i})_1-(\overline{x_1})_1\right]\times(S^{-1})_{1p}+\\ +\ldots\ldots+ & \cdots & +\ldots\ldots+ & \cdots & +\ldots\ldots+\\ +\left[(x_{1i})_j-(\overline{x_1})_j\right]\times(S^{-1})_{j1}+ & \cdots & +\left[(x_{1i})_j-(\overline{x_1})_j\right]\times(S^{-1})_{jj}+ & \cdots & +\left[(x_{1i})_j-(\overline{x_1})_j\right]\times(S^{-1})_{jp}+\\ +\ldots\ldots+ & \cdots & +\ldots\ldots+ & \cdots & +\ldots\ldots+\\ +\left[(x_{1i})_p-(\overline{x_1})_p\right]\times(S^{-1})_{p1} & \cdots & +\left[(x_{1i})_p-(\overline{x_1})_p\right]\times(S^{-1})_{pj} & \cdots & +\left[(x_{1i})_p-(\overline{x_1})_p\right]\times(S^{-1})_{pp}\end{bmatrix}\right\}\times\left(x_{1i}-\overline{x_1}\right)$$

$$=\left\{\begin{bmatrix}(S^{-1})_{11}\times\left[(x_{1i})_1-(\overline{x_1})_1\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{j1}\times\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{p1}\times\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1j}\times\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jj}\times\left[(x_{1i})_j-(\overline{x_1})_j\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{pj}\times\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1p}\times\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jp}\times\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{pp}\times\left[(x_{1i})_p-(\overline{x_1})_p\right]^2\end{bmatrix}\right\}$$

Aporte a Suma de Cuadrados Dentro de todos los casos pertenecientes a la clase 1:

$$\sum_{i=1}^{n_1}\left(x_{1i}-\overline{x_1}\right)^t\times S^{-1}\times\left(x_{1i}-\overline{x_1}\right)=$$

$$=\left\{\begin{bmatrix}(S^{-1})_{11}\times\sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{j1}\times\sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{p1}\times\sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_1-(\overline{x_1})_1\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1j}\times\sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jj}\times\sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{pj}\times\sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]\times\left[(x_{1i})_j-(\overline{x_1})_j\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1p}\times\sum_{i=1}^{n_1}\left[(x_{1i})_1-(\overline{x_1})_1\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jp}\times\sum_{i=1}^{n_1}\left[(x_{1i})_j-(\overline{x_1})_j\right]\times\left[(x_{1i})_p-(\overline{x_1})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{pp}\times\sum_{i=1}^{n_1}\left[(x_{1i})_p-(\overline{x_1})_p\right]^2\end{bmatrix}\right\}$$

Aporte a Suma de Cuadrados Dentro de todos los casos pertenecientes a la clase 2:

$$\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)^t\times S^{-1}\times\left(x_{2i}-\overline{x_2}\right)=$$

$$\begin{bmatrix}(S^{-1})_{11}\times\sum_{i=1}^{n_2}\left[(x_{2i})_1-(\overline{x_2})_1\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{j1}\times\sum_{i=1}^{n_2}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{p1}\times\sum_{i=1}^{n_2}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_1-(\overline{x_2})_1\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1j}\times\sum_{i=1}^{n_2}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jj}\times\sum_{i=1}^{n_2}\left[(x_{2i})_j-(\overline{x_2})_j\right]^2+\\ +\ldots\ldots+\\ +(S^{-1})_{pj}\times\sum_{i=1}^{n_2}\left[(x_{2i})_p-(\overline{x_2})_p\right]\times\left[(x_{2i})_j-(\overline{x_2})_j\right]\end{bmatrix}+\ldots+\begin{bmatrix}(S^{-1})_{1p}\times\sum_{i=1}^{n_2}\left[(x_{2i})_1-(\overline{x_2})_1\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{jp}\times\sum_{i=1}^{n_2}\left[(x_{2i})_j-(\overline{x_2})_j\right]\times\left[(x_{2i})_p-(\overline{x_2})_p\right]+\\ +\ldots\ldots+\\ +(S^{-1})_{pp}\times\sum_{i=1}^{n_2}\left[(x_{2i})_p-(\overline{x_2})_p\right]^2\end{bmatrix}$$

Aporte a Suma de Cuadrados Dentro de todos los casos del juego de datos:

$$\sum_{i=1}^{n_1}\left(x_{1i}-\overline{x_1}\right)^t\times S^{-1}\times\left(x_{1i}-\overline{x_1}\right)+\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)^t\times S^{-1}\times\left(x_{2i}-\overline{x_2}\right)=$$

$$=\begin{bmatrix}\left(S^{-1}\right)_{11}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_1-\left(\overline{x_1}\right)_1\right]^2+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_1-\left(\overline{x_2}\right)_1\right]^2\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{j1}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_j-\left(\overline{x_1}\right)_j\right]\times\left[\left(x_{1i}\right)_1-\left(\overline{x_1}\right)_1\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_j-\left(\overline{x_2}\right)_j\right]\times\left[\left(x_{2i}\right)_1-\left(\overline{x_2}\right)_1\right]\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{p1}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_p-\left(\overline{x_1}\right)_p\right]\times\left[\left(x_{1i}\right)_1-\left(\overline{x_1}\right)_1\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_p-\left(\overline{x_2}\right)_p\right]\times\left[\left(x_{2i}\right)_1-\left(\overline{x_2}\right)_1\right]\end{matrix}\right\}\end{bmatrix}+\ldots+\begin{bmatrix}\left(S^{-1}\right)_{1j}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_1-\left(\overline{x_1}\right)_1\right]\times\left[\left(x_{1i}\right)_j-\left(\overline{x_1}\right)_j\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_1-\left(\overline{x_2}\right)_1\right]\times\left[\left(x_{2i}\right)_j-\left(\overline{x_2}\right)_j\right]\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{jj}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_j-\left(\overline{x_1}\right)_j\right]^2+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_j-\left(\overline{x_2}\right)_j\right]^2\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{pj}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_p-\left(\overline{x_1}\right)_p\right]\times\left[\left(x_{1i}\right)_j-\left(\overline{x_1}\right)_j\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_p-\left(\overline{x_2}\right)_p\right]\times\left[\left(x_{2i}\right)_j-\left(\overline{x_2}\right)_j\right]\end{matrix}\right\}\end{bmatrix}+\ldots+$$

$$+\ldots+\begin{bmatrix}\left(S^{-1}\right)_{1p}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_1-\left(\overline{x_1}\right)_1\right]\times\left[\left(x_{1i}\right)_p-\left(\overline{x_1}\right)_p\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_1-\left(\overline{x_2}\right)_1\right]\times\left[\left(x_{2i}\right)_p-\left(\overline{x_2}\right)_p\right]\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{jp}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_j-\left(\overline{x_1}\right)_j\right]\times\left[\left(x_{1i}\right)_p-\left(\overline{x_1}\right)_p\right]+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_j-\left(\overline{x_2}\right)_j\right]\times\left[\left(x_{2i}\right)_p-\left(\overline{x_2}\right)_p\right]\end{matrix}\right\}+\\+\ldots\ldots+\\+\left(S^{-1}\right)_{pp}\times\left\{\begin{matrix}\sum_{i=1}^{n_1}\left[\left(x_{1i}\right)_p-\left(\overline{x_1}\right)_p\right]^2+\\+\sum_{i=1}^{n_2}\left[\left(x_{2i}\right)_p-\left(\overline{x_2}\right)_p\right]^2\end{matrix}\right\}\end{bmatrix}=$$

$$=\begin{bmatrix}\left(S^{-1}\right)_{11}\times S_{11}+&+\ldots+&\left(S^{-1}\right)_{1j}\times S_{1j}+&+\ldots+&+\left(S^{-1}\right)_{1p}\times S_{1p}+\\+\ldots+&\ldots&\ldots&\ldots&\ldots\\+\left(S^{-1}\right)_{j1}\times S_{j1}+&\ldots&+\left(S^{-1}\right)_{jj}\times S_{jj}+&\ldots&+\left(S^{-1}\right)_{jp}\times S_{jp}+\\+\ldots+&\ldots&\ldots&\ldots&\ldots\\+\left(S^{-1}\right)_{p1}\times S_{p1}+&\ldots&+\left(S^{-1}\right)_{pj}\times S_{pj}+&\ldots&+\left(S^{-1}\right)_{pp}\times S_{pp}\end{bmatrix}=$$

$$=\begin{bmatrix}\left(S^{-1}\right)_{11}\times S_{11}+&+\ldots+&\left(S^{-1}\right)_{j1}\times S_{1j}+&+\ldots+&+\left(S^{-1}\right)_{p1}\times S_{1p}+\\+\ldots+&\ldots&\ldots&\ldots&\ldots\\+\left(S^{-1}\right)_{1j}\times S_{j1}+&\ldots&+\left(S^{-1}\right)_{jj}\times S_{jj}+&\ldots&+\left(S^{-1}\right)_{pj}\times S_{jp}+\\+\ldots+&\ldots&\ldots&\ldots&\ldots\\+\left(S^{-1}\right)_{1p}\times S_{p1}+&\ldots&+\left(S^{-1}\right)_{jp}\times S_{pj}+&\ldots&+\left(S^{-1}\right)_{pp}\times S_{pp}\end{bmatrix}=$$

$$= \sum_{i=1}^{p} \left[\sum_{j=1}^{p} \left(S^{-1}\right)_{ij} \times S_{ji} \right]$$

$$S^{-1} \times S = I \Rightarrow (\textit{fila } i \textit{ de } S^{-1}) \times (\textit{columna } i \textit{ de } S) = 1 = \sum_{j=1}^{p} \left(S^{-1}\right)_{ij} \times S_{ji} \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{p} \left[\sum_{j=1}^{p} \left(S^{-1}\right)_{ij} \times S_{ji} \right] = \sum_{i=1}^{p} 1 = p$$

$$[(n_1 - 1) + (n_2 - 1)] \times \left[\sum_{i=1}^{n_1} \left(x_{1i} - \overline{x_1}\right)' S^{-1} \left(x_{1i} - \overline{x_1}\right) + \sum_{i=1}^{n_2} \left(x_{2i} - \overline{x_2}\right)' S^{-1} \left(x_{2i} - \overline{x_2}\right) \right] =$$

$$= [(n_1 - 1) + (n_2 - 1)] \times p = (n - 2) \times p$$

8 Bibliografía

[AIT/76] Aitchison J., Aitken C. G. G., "Multivariate binary discrimination by the kernel method", Biometrika, Biometrika Trust, Great Britain, volumen 63 número 3 -pag413-420. 1976.

[AND/72] Anderson J. A., "Separate sample logistic discrimination", Biometrika, Biometrika Trust, Great Britan, volumen 59, número 1 -pag19-35. 1972.

[AND/74] Anderson J. A. "Diagnosis by Logistic Discriminant Function: Further Practical Problems and Results", Journal of the Royal Statistical Society: Series C (Applied Statistics), Royal Statistical Society, London, volumen 23 número 3 -pag397-404. 1974.

[ASP/01] Asparoukhov O. K. Krzanowski W. J., "A comparition of discriminant procedures for binary variables", Computational Statistics & Data Analysis, Elsevier, volumen 38 – pag 139-160. 2001.

[BAR/95] Bar-Hen A., Daudin J. J., "Generalization of the Mahalanobis Distance in the Mixed Case", Journal of the Multivariate Analysis, volumen 53, - pag 332-342. 1995.

[BEI/85] Beitler P. J., Landis J. R., "A Mixed-Effects Model for Categorical Data", Biometrics, The Biometric Society, Washington D. C., volumen 41, número 4 - pag991-1000. 1985.

[CHA/00] Chaterjee, Sanprit, Hadi A. S., Price B., "Regression Analysis by example", Wiley, 2000.

[DAU/86] Daudin, J. J., "Selection of Variables in Mixed-Variable Discriminant Analysis", Biometrics, The Biometric Society Washington D. C., volumen 42, número 3,-pag473-481. 1986.

[DAU/99] Daudin J. J., Bar-Hen A., "Selection in Discriminant Analysis with continuous and discrete variables", Computational Statistics & Data Analysis, volumen 32, Elsevier, - pag 161-175. 1999.

[FIS/36] Fisher R. A., "The use of multiple measurement in taxonomic problems", Annals of Eugenics, volumen 7, pag 179-188. 1936.

[FIS/38] Fisher R. A., "The statistical Utilization of multiple measurements", Annals of Eugenics, volumen 8, pag. 376-386. 1938

[HAL/81] Hall P., "Optimal near neighbour estimator for use in discriminant analysis", Biometrika, Biometrika Trust, Grat Britain, volumen 68 número 2 -pag572-575. 1981.

[HOS/89] Hosmer D. W. Jr., Lemeshow S., "Applied Logistic Regression", Wiley, United Stated of America. 1989.

[JOH/02] Johnson R. A., Wichern D. W., "Applied Multivariate Statistical Analysis", Prentice Hall, New Jersey. 2000.

[KNO/82] Knoke J. D., "Discriminant Analysis with Discrete and Continuous Variables", Biometrics, The Biometric Society, Washington D. C., volumen 38, número 1, -pag191-200.1982.

[KRZ/75] Krzanowski W. J., "Discrimination and Classification Using Both Binary and Continuous Variables", Journal of the American Statistical Association, The American Statistical Association, Washington D. C., volumen 70, número 352 -pag782-790. 1975.

[KRZ/80] Krzanowski W. J., "Mixture of Continuous and Categorical Variables in Discriminant Analysis", Biometrics, The Biometric Society, Washington D. C., volumen 36, número 3, -pag493-499. 1980.

[KRZ/83] Krzanowski W. J., "Distance between populations using mixed continuous and categorical variables", Biometrika, Biometrika Trust, Great Britain , volumen 70 número 1 -pag235-243. 1983.

[KRZ/83] Krzanowski W. J., "Stepwise Location Model Choice in Mixed-variable Discrimination", Journal of the Royal Statistical Society: Series C (Applied Statistics), Royal Statistical Society, London, volumen 32, número 3, -pag260-266. 1983.

[KRZ/95] Krzanowski W. J., "Selection of variables, and assesment of their performance, in mixed-variable discriminant analysis", Computational Statistics & Data Analysis, volumen 19, Elsevier, -pag 419-431. 1995.

[LEO/05] de Leon A. R., Carrière, "A generalized Mahalanobis Distance for Mixed Data", Journal of Multivariate Analysis, Elsevier, volumen 92, -pag174-185. 2005.

[MAH/36] Mahalanobis P. C., "On the Generalized Distance in Statistics", National Institute of Sciences of India, vol 12, pág 49-55. 1936.

[MER/04] Merbouha A., Mkhadri A., "Regularization of the location model in discrimination with mixed discrete and continuous variables", Computational Statistics & Data Analysis, volumen 45, Elsevier, -563-576. 2004.

[MON/02] Montgomery D. C., Peck E. A., Vining G. G., "Introducción al Análisis de Regresión Lineal", Compañía Editorial Continental, México, 2002.

[MOR/98] Morales D., Pardo L., Zografos K. "Informational distances and related statistics in mixed continuous and categorical variables", Journal of Statistical Planning and Inference, Elsevier, North-Holland, volumen 75 -pág47-63. 1998.

[NUÑ/03] Núñez M., Villarroya A., Oller J. M., "Minimun Distance Probability Discriminant Analysis for Mixed Variables", Biometrics, The International Biometric Society, Washington D. C., volumen 59, número 2, -pag248-253. 2003.

[POH/04] Pohar M., Blas M., Turk S., "Comparition of Logistic Regression and Discriminant Analysis: A Simulation Study", Metodološki zvezki, volumen 1 número 1, -pag 143-161. 2004.

[PRE/78] Press J., Wilson S., "Choosing Between Logistic Regression and Discriminant Analysis", Journal of the American Statistical Association, The American Statistical Association, Washington D. C., volumen 73 número 364 -pag 699-705. 1978.

[VIL/95] Villarroya A., Ríos M., Oller J. M., "Discriminant Analysis Algorithm Based on a Distance Function and on Bayesian Decision", Biometrics, The International Biometric Society, Washington D. C., volumen 51, número 3, -pag908-919. 1995.

[VLA/82] Vlachonikolis I. G., Marriot F. H. C., "Discrimination with Mixed Binary and Continuous Data", Journal of the Royal Statistical Society: Series C (Applied Statistics), Royal Statistical Society, London, volumen 31 número 1 -pag23-31. 1982.

[VLA/90] Vlachonikolis I. G., "Predictive discrimination and classification with mixed binary and continuous variables", Biometrika, Biometrika Trust, Great Britain, volumen 77, número 3, -pag657-662. 1990.