Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Académica Profesional de Estadística

Análisis de regresión logística aplicada a la educación

Monografía

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Juan Manuel MANCO POMACAJA

ASESOR

Rosa Ysabel ADRIAZOLA CRUZ

Lima, Perú

2007

Referencia bibliográfica

Manco, J. (2007). *Análisis de regresión logística aplicada a la educación*. Monografía para optar el título de Licenciado en Estadística. Escuela Académico Profesional de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

ANÁLISIS DE REGRESIÓN LOGÍSTICA APLICADA A LA EDUCACIÓN

Juan Manuel Manco Pomacaja

Monografía presentada a consideración del Cuerpo de Docentes de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el Título Profesional de Licenciado en Estadística.

Aprobada por:

Jurado Evaluador

Asesora

Lima – Perú
Enero - 2007

FICHA CATALOGRÁFICA

JUAN MANUEL MANCO POMACAJA

Análisis de Regresión Logística aplicada a la Educación. (Lima 2007).

III, 65p., (UNMSM, Licenciado, Estadística, 2007)

Monografía, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas. Estadística.

“Dedicado a mi padre y a mi madre que

siempre me apoyaron en mi

formación profesional”

RESUMEN

ANÁLISIS DE REGRESIÓN LOGÍSTICA APLICADA A LA EDUCACIÓN

JUAN MANUEL MANCO POMACAJA

ENERO – 2007

Orientador: Mg. Rosa Ysabel Adriazola Cruz
Titulo: Licenciado en Estadística.

La presente monografía trata de proporcionar una explicación general y a la vez detallada del Análisis de Regresión Logística, y así como también la construcción de un modelo a través de etapas en la cual se van seleccionando variables y al mismo tiempo eliminando otras.

El siguiente trabajo esta dirigido al área de Educación, en el cual se presenta una aplicación con respecto a los ingresantes de la Facultad de Ciencias Administrativas correspondiente al proceso de admisión 2006-I de la Universidad Nacional Mayor de San Marcos.

El objetivo del trabajo monográfico es conocer que características presentan los ingresantes según el colegio de procedencia (Estatal ó Particular).

La información de los ingresantes se ha obtenido a través de la Oficina Central de Admisión quien nos brindo la base de datos con variedad de variables relevantes, asimismo hacemos presente la absoluta reserva de información del ingresante.

SUMMARY

LOGISTIC REGRESSION ANALYSIS APPLIED TO EDUCATION

JUAN MANUEL MANCO POMACAJA

ENERO – 2007

Advisory: Mg. Rosa Ysabel Adriazola Cruz
Professional degree obtained: Licentiate in Statistics

This monograph seeks to provide an overview and also detaileAPOYARON d logistic regression analysis, and as well as the construction of a model through stages in which variables are chosen, while eliminating others.

The following is directed to the area of Education, in which an application is submitted with respect to entering the Faculty of Administrative Sciences for the admission process 2006-I of the National University of San Marcos.

The aim of this work is to know which features case presents as entering the school of origin (State or private).

The information from the entrants was obtained through the Admissions Office who provides the database with a variety of variables; we also present the total pool of incoming information.

ÍNDICE

Pág.

CAPÍTULO I

CAPÍTULO II

CAPÍTULO III

CAPÍTULO I

MODELO DE REGRESIÓN LOGÍSTICA

1.1 INTRODUCCIÓN

La Regresión Logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en diversas áreas del conocimiento, es un método multivariante que nos permite estimar la relación existente entre una variable dependiente cualitativa dicotómica que puede tomar únicamente dos valores: 1, presencia (con probabilidad p); y 0 la ausencia (con probabilidad 1-p) y un conjunto de variables independientes (explicativas) que pueden ser cuantitativas o cualitativas, asimismo se debe obtener una función lineal de las variables independientes que permita clasificar a las unidades de análisis en una de las subpoblaciones o grupos establecidos por los dos valores de la variable dependiente.

La ecuación del modelo de Regresión Logística es exponencial la cual es linealizada a través de una transformación logarítmica.

Una característica del modelo es su utilidad en situaciones prácticas de investigación en que la respuesta y las variables que contribuyan más en el aumento o disminución de la probabilidad de un éxito.

En el modelo de regresión logística si una variable explicativa de tipo categórica tiene “c” niveles habrá que generar “c-1” variables ficticias (dummy) a fin de que todos los posibles valores de la variable queden bien representadas en el modelo.

Por otro lado si todas las variables explicativas son categóricas, se utilizará el modelo log lineal.

1.2. REGRESIÓN LOGÍSTICA

1.2.1. OBJETIVOS

a. El objetivo principal de este método es el de modelar la influencia de las variables explicativas (regresoras) en la probabilidad de ocurrencia de un suceso particular.
b. Determinar el modelo (más parsimonioso) de probabilidad que mayor describa la relación entre la variable respuesta y un conjunto de variables explicativas.

En el caso del modelo de regresión logística simple establece la siguiente relación entre la probabilidad $\pi(x)$ y un valor de $X = x$ está dada por:

$$\pi(x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad (1.1)$$

A continuación se presentan los siguientes gráficos: 2-1 y 2-2 donde se visualizan mejor esta relación cuando $x \to \infty, \pi(x) \to 0$ si $\beta_1 < 0$ y $\pi(x) \to 1$ si $\beta_1 > 0$. Por otro lado si $|\beta_1|$ tiende a crecer, la tasa de incremento o de disminución de $\pi(x)$ aumenta en ambas situaciones y a medida que β_1 se

acerca a 0, la curva se va inclinando hasta convertirse en una línea horizontal (cuando β_1=0).

Gráfico 2-1

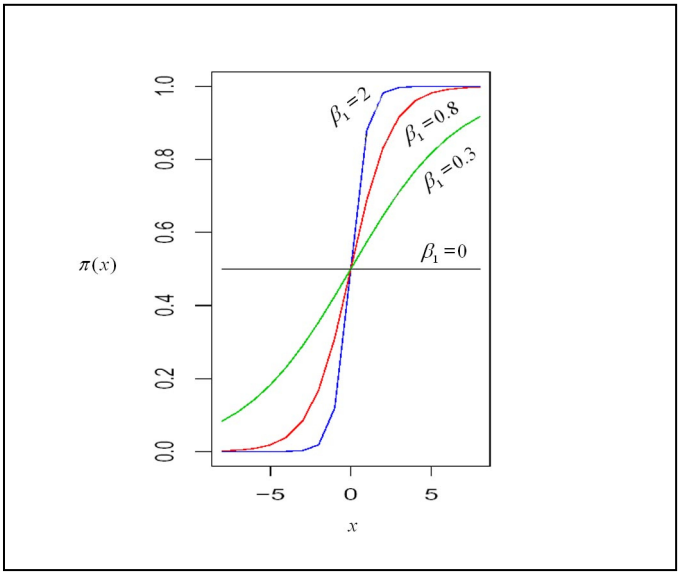
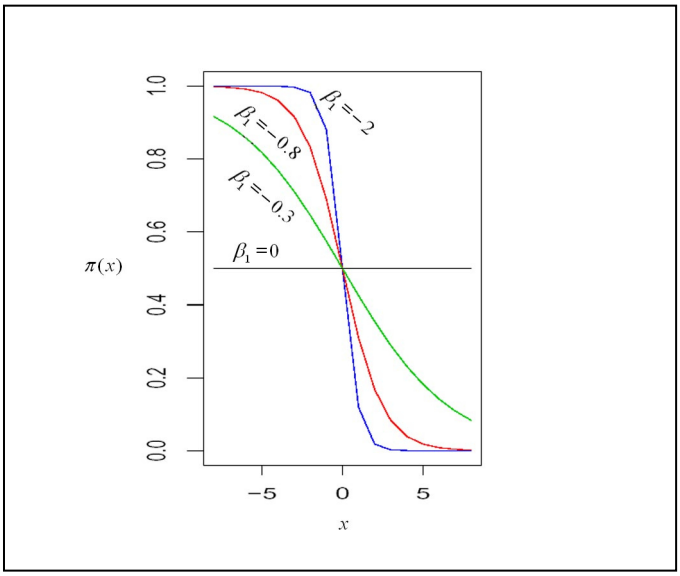


Gráfico 2-2



De la ecuación (1.1) puede obtenerse lo siguiente:

$$\frac{\pi(x)}{1-\pi(x)} = e^{(z)} = e^{(\beta_0+\beta_1 x)} = e^{(\beta_0)} e^{(\beta_1 x)}$$

Teniendo en cuenta que: $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = z$

La razón $\frac{\pi(x)}{1-\pi(x)}$ es llamada también razón de probabilidades o razón de ventaja y representa la oportunidad de éxito.

Esto es imprescindible para la interpretación de β_1, puesto que la razón de probabilidades se incrementa en e^{β_1} por cada unidad de incremento en x.

En el caso de la regresión logística múltiple,

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{(\beta_0+\beta_1 x_1+.....+\beta_p x_p)}$$

La razón de probabilidades se incrementa en e^{β_i} por cada unidad de incremento en x_i cuando se mantienen fijos los niveles de las otras x_j.

1.2.2. REGRESIÓN LOGÍSTICA VERSUS REGRESIÓN LINEAL

Se utiliza la regresión logística por los siguientes inconvenientes de la regresión lineal:

- ✓ Debido a que la variable Y toma sólo dos valores, los supuestos de normalidad y Homocedasticidad de los errores no son satisfechas. Si se utiliza la regresión lineal y los valores pronosticados pueden ser mas grande que uno y menores que cero sobrepasando los ejes X, dichos valores son teóricamente inadmisibles debido a que las probabilidades están en el intervalo [0,1].

- ✓ Una de las suposiciones de regresión lineal es que la varianza de Y es constante a través de X (Homocedasticidad). Esto no se cumple con una variable binaria, porque los errores son heterocedásticos: var(error) = $\pi(x)(1-\pi(x))$, donde π es la probabilidad del evento. Como π depende de la matriz de datos X, implica el no cumplimiento de este supuesto.

Como una solución a estos inconvenientes se utilizan los modelos lineales generalizados (MLG) los cuales extienden los modelos de regresión convencionales en el uso de variables respuesta que no tienen distribución normal.

Los modelos lineales generalizados están conformados por tres componentes:

1. Un componente aleatorio que identifica la variable respuesta Y y su distribución de probabilidad, perteneciente a la familia exponencial.

$$f(y_i,\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

   Donde θ_i y y_i representan: el parámetro del modelo y la variable respuesta para la i-esima observación respectivamente y a, b y Q son funciones reales.

2. Un componente sistemático que representa a las variables regresoras a través de una función lineal. Esto es, un vector $(z_1,\ldots\ldots,z_n)$ de variables explicativas a través de un modelo lineal. Dado x_{ij} el valor de la regresora j para el sujeto i entonces

$$z_i = \sum_{j=0}^{p} \beta_j \chi_{ij} \qquad (2.1)$$

   Con $x_{io} = 1 \ \forall i$

3. Una función de enlace que conecta los componentes aleatorio y sistemático. Dada $\mu_i = E(Y_i), i = 1,....., n.$ El modelo enlaza a μ_i con z_i por $z_i = g(\mu_i)$ donde la función de enlace g es monótona y diferenciable. Así, g enlaza $E(Y_i)$ con las variables explicativas a través de la fórmula

$$g(\mu_i) = z_i$$

En el modelo de regresión clásico el objetivo es estimar la media condicional de Y dado X_i así, la función de enlace es la función identidad $g(\mu_i) = \mu_i$ y el componente sistemático es una función lineal de las variables x. Esto es:

$$\mu_i = \sum_j^p \beta_j \chi_{ij}$$

En el modelo de regresión logística se tienen n variables aleatorias binomiales y_i, $i=1,\ldots\ldots n.$ Para cada y_i se conoce la cantidad de ensayos m_i y un vector de p variables predictoras asociado a $x_i = (x_{i1}, x_{i2,\ldots\ldots} x_{ip})'$. La probabilidad de éxito $\pi(x_i)$ depende de x_i, entonces

$$y_i / x_i \sim Bin\ (m_i, \pi(x_i)) \qquad i = 1, \ldots\ldots, n.$$

Para esta variable la media y varianza son dada por

$$\mu_i = E(y_i / x_i) = m_i \pi(x_i) \tag{2.2}$$

$$V_i = V(y_i / x_i) = m_i \pi(x_i)(1 - \pi(x_i)) \tag{2.3}$$

respectivamente.

En la mayoría de los casos m_i =1, es decir, y_i es una variable de Bernoulli que toma el valor 1 si la unidad en estudio presenta característica de interés está presente y 0 si no esta presente. Para estimar la probabilidad de $\pi(x_i)$ se utiliza una función Kernel M aplicada a z_i. La función Kernel para la regresión logística

debe estar entre 0 y 1. La más frecuentemente utilizada es la función logística dada por:

$$\pi(x_i) = M(z_i) = \frac{e^{z_i}}{1+e^{z_i}} = \frac{1}{1+e^{-z_i}} \tag{2.4}$$

Con z_i dado por (2.1). Al resolver la ecuación (2.4) para z_i se obtiene:

$$\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = z_i \tag{2.5}$$

Entonces, para el modelo de regresión logística el componente sistemático está dado por: $z_i = \sum_j \beta_j x_{ij}$ y la función logit o de enlace que hace al modelo lineal es:

$$g(\pi(x_i)) = \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \tag{2.6}$$

La razón de probabilidades $\frac{\pi(x)}{1-\pi(x)}$ representa la oportunidad de éxito.

También conocido como: ODDS RATIO = $\frac{\pi(x)}{1-\pi(x)}$

El ODDS asociado a un suceso es el cociente entre la probabilidad de que ocurra un "éxito" frente a la probabilidad de que no ocurra, siendo $\pi(x)$ la probabilidad de éxito.

Por ejemplo, sean dos juegos, en uno (x=0) se apuesta sobre la salida de una cierta cara en una tirada de un dado, y en otro (x=1) sobre la salida de una cara en la tirada de una moneda.

Evidentemente, la probabilidad de ganar es para el dado p/(x=0)=1/6 y para la moneda p/(x=1)=1/2.

El odds ratio para este ejemplo es:

$$OR = \frac{p/q/(x=1)}{p/q/(x=0)} = \frac{\frac{1/2}{1/2}}{\frac{1/6}{5/6}} = 5$$

El odds para la moneda es 5 veces el odds del dado, es decir, a la larga la razón de partidas ganadas/ perdidas es 5 veces mayor para la moneda que para el dado.

1.2.3. REGRESIÓN LOGÍSTICA Y MÉTODOS AFINES

El objetivo general de la regresión logística es predecir la probabilidad de un evento de interés de una investigación, así como identificar las variables predictoras útiles para la estimación.

Se pueden dar varios métodos multivariantes para predecir una variable respuesta de naturaleza dicotómica a partir de un grupo de variables regresoras.

En el caso del análisis de regresión lineal múltiple y el análisis discriminante que son métodos eficaces pero adolecen en algunos puntos, por ejemplo en el análisis de regresión lineal múltiple cuando la variable respuesta tiene solo dos valores, se violan los supuestos necesarios, por tal motivo se plantean lo siguiente:

- ✓ La distribución de los errores aleatorios no es normal.
- ✓ Los valores predictados no pueden ser interpretados como probabilidades como en el regresión logística, porque no toman valores dentro del intervalo [0,1].

Del mismo modo el análisis discriminante permite la predicción de pertenencia de la unidad de análisis a uno de los dos grupos pre-establecidos pero se

requiere que se cumplan los supuestos de multinormalidad de las variables regresoras y la igualdad de matrices de covarianzas de los dos grupos.

Por tanto, la regresión logística requiere menos supuestos que el análisis discriminante.

1.3. REGRESIÓN LOGÍSTICA MULTINOMIAL

El modelo de regresión logística es el más empleado cuando la variable respuesta es binomial, asimismo puede ampliarse este modelo a variables respuesta multinomiales.

Por ejemplo una variable respuesta Y tiene J categorías: 0,1,2,….J-1. Del mismo modo que en el caso binomial, en el cual la variable respuesta es parametrizada en términos de la función logit de $Y = 1$ vs $Y = 0$, en el caso multinomial se tienen J-1 funciones logit comparando Y=1 vs Y=0, Y =2 vs Y=0,……, Y=J-1 vs Y=0; las demás funciones logit pueden ser obtenidas de manera semejante a las combinaciones anteriores.

Dado $x_i = (x_{i1},x_{i2},.....,x_{ip})'$ el vector de p predictoras asociado a y_i y dado $\pi_{ji} = \pi_j(x_i) = p(Y = s / x_i)$ para s=0,1,……,J-1, entonces las J-1 funciones logit son:

$$g_1(x_i) = \ln\left(\frac{\pi_{1i}}{\pi_{0i}}\right) = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + ... + \beta_{1p}x_{ip}$$

$$g_2(x_i) = \ln\left(\frac{\pi_{2i}}{\pi_{0i}}\right) = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + ... + \beta_{2p}x_{ip}$$

$$\vdots$$

$$g_{J-1}(x_i) = \ln\left(\frac{\pi_{(J-1)i}}{\pi_{0i}}\right) = \beta_{(J-1)0} + \beta_{(J-1)1}x_{i1} + \beta_{(J-1)2}x_{i2} + ... + \beta_{(J-1)p}x_{ip}$$

Se conoce que,

$$\pi_{0i} + \pi_{1i} + ... + \pi_{(J-1)i} = 1,$$

Esto implica

$$1 + \frac{\pi_{1i}}{\pi_{0i}} + ... + \frac{\pi_{(J-1)i}}{\pi_{0i}} = \frac{1}{\pi_{0i}}$$

Equivalentemente,

$$1 + e^{g_1(x_i)} + ... + e^{g_{J-1}(x_i)} = \frac{1}{\pi_{0i}}$$

y así se obtiene

$$\pi_{0i} = \frac{1}{1 + e^{g_1(x_i)} + ... + e^{g_{J-1}(x_i)}}$$

Con el objetivo de obtener una expresión para π_{1i}, multiplicamos a ambos lados π_{1i} de la ecuación

$$(\pi_{1i})\pi_{0i} = \frac{1}{1 + e^{g_1(x_i)} + ... + e^{g_{J-1}(x_i)}}(\pi_{1i})$$

Entonces

$$\pi_{1i} = \frac{e^{g_1(x_i)}}{1 + e^{g_1(x_i)} + ... + e^{g_{J-1}(x_i)}}$$

Similarmente multiplicamos π_{ji} en ambos lados de la siguiente ecuación, resultando

$$\pi_{Ji} = \frac{e^{g_J(x_i)}}{1 + e^{g_1(x_i)} + ... + e^{g_{J-1}(x_i)}}, \quad J = 1,...., J-1$$

Si se consideran variables regresoras categóricas, por ejemplo si una variable esta conformada por d niveles será necesario incorporar d-1 variables indicadoras, o comúnmente llamado variables dummy.

CAPÍTULO 2

ESTIMACIONES

2.1. ESTIMACIÓN DE PARAMETROS EN REGRESIÓN LOGÍSTICA

En regresión lineal el método más utilizado para la estimación de parámetros es el de mínimos cuadrados, este método radica en hallar los valores de β que minimizan la suma de cuadrados de las desviaciones de los valores observados de Y con respecto a los valores predichos por el modelo.

El método de mínimos cuadrados proporciona estimadores con propiedades estadísticas que se cumplen solamente bajo los supuestos de la regresión lineal. Todo lo contrario sucede con el modelo de regresión logística que no cumple con los supuestos de la regresión lineal, por lo tanto el método de mínimos cuadrados no produce estimaciones eficaces en la regresión logística.

Por tal motivo el método más adecuado es el de máxima verosimilitud, que suministra los valores de los parámetros desconocidos que maximizan la probabilidad de contar un buen grupo de datos observados.

De este modo se presenta el siguiente procedimiento:

1. Se construye la función de verosimilitud, este a su vez indica la probabilidad de los datos observados en función del vector de parámetros desconocidos:

$$\beta = (\beta_0, \beta_1, ..., \beta_p)'$$

En el caso de la regresión logística, si Y es codificada como cero o uno, implica que la expresión para $\pi(x)$ dada en (2.4) resulta la probabilidad condicional de que Y sea igual a 1 dado x y la cantidad 1-$\pi(x)$ resulta la probabilidad condicional de que Y sea igual a 0 dado x.

Así, para los pares (x_i, y_i) en los cuales y_i=1 la contribución de la función de verosimilitud es $\pi(x_i)$ y para los pares en los que y_i=0 la contribución a la función de verosimilitud es 1-$\pi(x_i)$. En conclusión del par (x_i, y_i) a la función de verosimilitud es:

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

La función de verosimilitud es obtenida como el producto de los n términos, ya que las observaciones se asumen independientes.

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

Aplicando *ln* ambos lados:

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

2. Luego se halla el vector de derivadas parciales de $L(\beta)$ con respecto a β

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta} \qquad \text{con} \qquad u_k = \frac{\partial L(\beta)}{\partial \beta_k},\ k = 0, ..., p$$

donde p es la cantidad de parámetros, Cada una de las p+1 ecuaciones se iguala a cero y se resuelve para β_k, lo cual forma un conjunto de p+1 ecuaciones en p+1 incógnitas.

Aplicando logaritmo natural en (2.4):

$$\ln[\pi(x_i)] = z_i - \ln(1+e^{z_i})$$

$$\ln[1-\pi(x_i)] = -\ln(1+e^{z_i})$$

$$\frac{\partial \ln[\pi(x_i)]}{\partial \beta_0} = 1 - \frac{e^{z_i}}{1+e^{z_i}} = 1 - \pi(x_i)$$

$$\frac{\partial \ln[\pi(x_i)]}{\partial \beta_j} = x_{ij} - \frac{x_{ij}e^{z_i}}{1+e^{z_i}} = x_{ij}[1-\pi(x_i)]$$

$$\frac{\partial \ln[1-\pi(x_i)]}{\partial \beta_0} = -\frac{e^{z_i}}{1+e^{z_i}} = -\pi(x_i)$$

$$\frac{\partial \ln[1-\pi(x_i)]}{\partial \beta_j} = -\frac{x_{ij}e^{z_i}}{1+e^{z_i}} = -x_{ij}\pi(x_i)$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}\{y_i(1-\pi(x_i)) + (1-y_i)(-\pi(x_i))\} = \sum_{i=1}^{n}\{y_i - \pi(x_i)\}$$

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\{y_i x_{ij}(1-\pi(x_i)) + (1-y_i)(-x_{ij}\pi(x_i))\} = \sum_{i=1}^{n}\{x_{ij}[y_i - \pi(x_i)]\}$$

Luego: $U(\beta) = \left(\sum_{i=1}^{n}\{y_i - \pi(x_i)\}, \sum_{i=1}^{n}\{x_{ij}[y_i - \pi(x_i)]\}, ..., \sum_{i=1}^{n}\{x_{ip}[y_i - \pi(x_i)]\}\right)'$

$$= X'(Y - \prod)'$$

Con $Y = (y_1, ...y_n)'$, $\prod = (\pi(x_1), ...\pi(x_n))'$ y la matriz de datos de X dada por:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & x'_1 \\ 1 & x'_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x'_n \end{pmatrix}$$

Por tanto, las ecuaciones de verosimilitud son los siguientes:

$$\sum_{i=1}^{n} \{y_i - \pi(x_i)\} = 0$$

$$\sum_{i=1}^{n} \{x_{ij}[y_i - \pi(x_i)]\} = 0$$

A través de las cuales se obtendrán los estimadores.

2.2. PRUEBAS DE SIGNIFICANCIA PARA LOS PARAMETROS

Del mismo modo que el Modelo de Regresión Múltiple, en el Modelo de Regresión Logística se utilizan pruebas con distintos fines, siendo lo siguiente:

a. Verificar si una variable regresora tiene coeficiente igual a cero.
b. Verificar si un grupo de variables regresoras tienen coeficientes igual a cero.
c. Verificar la calidad global del modelo.

A continuación se presentara una descripción de cada prueba:

2.2.1. PRUEBA DE WALD

Al igual que el estadístico t en la regresión lineal múltiple, el estadístico de Wald en la regresión logística juega el mismo rol sobre las variables incluidas

en la ecuación. Por ejemplo si una variable independiente X_j seleccionada, y β_j es el parámetro asociado a X_j en la ecuación de regresión logística, el estadístico de Wald permite contrastar la hipótesis nula:

$$H_0 : \beta_j = 0 \quad j = 1,2,.....,k$$

$$H_1 : algún\ \beta_j \neq 0$$

Esta es una prueba asintótica para máximos verosímiles. Por otro lado los parámetros estimados en los modelos logísticos tienen una distribución normal para muestras grandes.

Por otro lado no es muy recomendable utilizar la estadística de Wald cuando se tiene un coeficiente demasiado grande, mejor es crear un modelo con y sin esa variable y basarse en la prueba de hipótesis de la diferencia entre los dos modelos.

Estadístico de prueba:

$$W = \frac{(\bar{\beta}_j)^2}{\sigma^2(\bar{\beta}_j)} \sim X^2_{\alpha,k+1}$$

Nivel de significancia: α

Regla de decisión:

Si la estadística de Wald, $W > X^2_{\alpha,k}$ se rechazará la hipótesis H_0 con un nivel de significancia fijado α, concluiremos que la variable explicativa influye en la probabilidad del suceso caso contrario si el p-valor asociado al estadístico de Wald es menor que α se rechazará la hipótesis nula.

2.2.2. PUNTUACIÓN EFICIENTE DE RAO

Como ya hemos visto anteriormente el estadístico de Wald hace la misma función que el estadístico *t* en la regresión lineal múltiple sobre las variables incluidas en la ecuación, la puntuación eficiente de Rao realiza el mismo papel para las variables no incluidas en la ecuación.

Por ejemplo el parámetro β_j está asociado a la variable X_j , supuesto que entrara en la ecuación de regresión en el siguiente paso. El estadístico puntuación eficiente de Rao nos permite contrastar la hipótesis:

$$H_0 : \beta_j = 0$$

Con un nivel de significancia α

La hipótesis indica que si la variable X_j ha sido seleccionada en el siguiente paso, la información que contribuiría no sería significativa. Se rechazará la hipótesis nula si el p-valor asociado al estadístico puntuación eficiente de Rao es menor que α .

Así en el proceso de selección de variables, se seleccionara la variable que presente el mínimo p-valor asociado al estadístico eficiente de Rao.

2.3. BONDAD DE AJUSTE DEL MODELO

La bondad de ajuste es analizar cuan probables son los resultados muestrales a partir del modelo ajustado, para evaluar la bondad de ajuste se utilizará el logaritmo del cociente de verosimilitud y la prueba de Hosmer-Lemeshow.

2.3.1. ESTADÍSTICO -2logL

Esta prueba estadística evalúa los cambios que se producen cuando se adiciona o se extrae una variable, donde L es la función de verosimilitud del modelo, y puede variar entre 0 y 1, si el modelo se ajusta a los datos implicará una verosimilitud igual a 1, de allí que -2logL=0.

Diremos que el modelo se ajusta a la data si tiene un valor pequeño de -2logL que es el logaritmo de la verosimilitud y se distribuye como una X^2 (Ji-cuadrado), cuando el modelo solo incluye la constante los grados de libertad es igual al número de casos menos uno (n-1), y cuando se incluye una variable independiente sigue una distribución X^2 con n-k-1 grados de libertad, en el modelo de regresión logística simple es n-2, la diferencia entre estos dos valores de -2logL se denomina “Desvianza”, que prueba si la variable x_i es significativa, se define como:

D = -2logL (verosimilitud del modelo sin la variable / verosimilitud del modelo con la variable)

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\pi(x_i)}{y_i}\right) + (1-y_i)\log\left(\frac{1-\pi(x_i)}{1-y_i}\right)\right]$$

Hipótesis:

H_0 : El modelo se ajusta a los datos observados

H_1 : El modelo no se ajusta a los datos observados

Nivel de significancia: α

Estadístico de prueba:

$$D \sim X^2 con\ n-k-1\ grados de libertad$$

Regla de decisión:

Si $D < X^2_{\alpha,(n-k-1)}$ no rechazamos H_0, el modelo ajustado es significativo.

2.3.2. PRUEBA DE HOSMER - LEMESHOW

Esta prueba evalúa la bondad de ajuste del modelo, mejor dicho el grado en que la probabilidad predicha coincide con la observada, construyendo una tabla de contingencia, dividiendo la muestra en aproximadamente 10 grupos iguales a partir de las probabilidades estimadas, comparando las frecuencias observadas con las esperadas en cada uno de estos grupos a través de la prueba X^2 con j-2 grados de libertad, en donde j es el número de grupos formados.

Se calcula los deciles de las probabilidades estimadas $\pi(x_i)$; $i = 1,.....,n$ y $D_1, D_2, ..., D_9$ y divide los datos observados en 10 categorías dadas por:

$$A_j = \{\bar{\pi}(x_i) \in \lfloor D_{j-1}, D_j \rfloor / \, i \in \{1,2,.....,n\}\} \quad j = 1,2,.....,10$$

Donde $D_0 = 0, D_{10} = 1$

Sean:

N_j = número de casos en A_j ; $j = 1,2,.....,10$

O_j = número de $y_i = 1$ en A_j ; $j = 1,2,.....,10$

$$\bar{\pi}(x_j) = \frac{1}{n_j} \sum_{i \in A_j} \bar{\pi}(x_i) \quad ; \quad j = 1,2,.....,10$$

Hipótesis:

H_0 : El modelo es adecuado

H_1 : El modelo no es adecuado

Nivel de significancia: α

Estadístico de prueba:

$$X^2 = \sum_{j=1}^{10} \frac{\left(O_j - n_j \bar{\pi}(x_j)\right)^2}{n_j \bar{\pi}(x_j)(1 - \bar{\pi}(x_j))}$$

Regla de decisión:

Si $X^2 \geq X^2_{\alpha, j-2}$ rechazamos H_0 y concluimos que el modelo no es adecuado a un nivel de significancia α.

2.4. ANÁLISIS DE RESIDUOS PARA LA REGRESIÓN LOGÍSTICA

A continuación se presentará varios tipos de residuos que permitirán si una observación es discordante o no.

2.4.1. RESIDUOS DE PEARSON

Está definido de la siguiente manera:

$$r_i = \frac{y_i - m_i \bar{\pi}(x_i)}{\sqrt{\left(m_i \bar{\pi}(x_i)(1 - \bar{\pi}(x_i))\right)}}$$

Donde:

Si los valores de la variable de respuesta están agrupados, y_i representa el número de veces que $y = 1$ entre las m_i repeticiones de X_i.

El residual de Pearson es similar al residual estudentizado usado en la regresión lineal. Así un residual de Pearson en valor absoluto mayor que 2 indica un dato anormal.

La estadística X^2 de Pearson es la suma de cuadrados de los residuales de Pearson.

$$X^2_p = \sum_{i=1}^{I} r_i^2$$

2.4.2. RESIDUOS DE DESVIANZA

Los residuos de desvianza tienen el mismo signo que $y_i - m_i\hat{\pi}(x_i)$ y están definidos por:

$$D_i = \pm\left\{2\left[y_i \ln\left(\frac{y_i}{m_i\hat{\pi}(x_i)}\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i(1-\hat{\pi}(x_i))}\right)\right]\right\}^{1/2}$$

Si el residual de desviana es mayor que 4 en valor absoluto entonces la observación correspondiente es anormal.

La desviana es igual a la suma de cuadrados de los residuales desviana:

$$X_D^2 = \sum_{i=1}^{I} d_i^2$$

$$X_D^2 = \sum_{i=1}^{I} d_i^2 \sim X_{i-k}^2$$

Donde k es el número de parámetros estimados en el modelo logit.

2.5. EVALUACIÓN DE LA CAPACIDAD PREDICTIVA DEL MODELO

Para conocer la calidad predictiva del modelo se puede utilizar la tabla de clasificación.

2.5.1. TABLA DE CLASIFICACIÓN

La ecuación del modelo ya diseñado nos proporciona una probabilidad P(y=1|X), lo que nos permite predecir a partir de ella para cada sujeto un valor de y (Y_{pred}), tal que si P(y =1|X) ≤ 0.5 entonces Y_{pred} =0, y si P(y =1|X) > 0.5 entonces Y_{pred} =1. Estos valores predichos de Y pueden enfrentarse a los valores reales de Y (Y_{obs}) de la muestra, obteniendo una tabla de 2x2 de la que

es posible determinar la tasa global de clasificaciones correctas, la sensibilidad, la especificidad, el valor predictivo positivo, el valor predictivo negativo y el llamado índice de Youden (sensibilidad + especificidad - 1); mayores valores del índice de Youden denotarán una mejor capacidad predictiva.

Sin embargo, las tablas de clasificación y sus correspondientes índices son malos parámetros para comparar distintos modelos, pues sensibilidad y especificidad dependen, no del ajuste del modelo, sino de la distribución de probabilidades de la muestra sobre la que se calculan.

En la tabla se enfrentan los valores estimados y observados.

Grupo Observado	Grupo Estimado		Total Marginal
	0	1	
0	n_{11}	n_{12}	$n_{11} + n_{12}$
1	n_{21}	n_{22}	$n_{21} + n_{22}$
Total Marginal	$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

2.6. OTRAS PRUEBAS DE SIGNIFICANCIA CON RESPECTO A LOS PARÁMETROS DEL MODELO DE REGRESIÓN LOGÍSTICA

2.6.1. PRUEBA DE LIKELIHOOD-RATIO

Esta prueba también se le conoce como prueba de chi-cuadrado y se utiliza para comprobar si los coeficientes de las variables regresoras son igual a cero.

Se utilizará la prueba de razón de verosimilitud, para elegir un modelo.

Sea la hipótesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_q = 0$$

$$H_1 : \beta_j \neq 0, \; al\ menos\ un\ j=1,2,...,q$$

Estadístico de Prueba:

$$X_q^2 = -2\ln\left(\frac{L_{p-q}}{L_p}\right) = -2[\ln L_{p-q} - \ln L_p]$$

Nivel de significancia: α

Esta transformación logarítmica origina una prueba estadística chi-cuadrado. Esta prueba es recomendada para emplear cuando se construye un modelo a través de “backward stepwise”

Los coeficientes de las variables retiradas son iguales a cero, esto conlleva a que la estadística X_q^2 tenga una distribución asintótica. Cuando el p-valor asociado a la estadística es pequeña implicara que una ó más de las q variables retiradas tienen coeficiente distinto de cero.

2.6.2. PRUEBA DE CHI-CUADRADO DE PEARSON

Esta prueba es de mucha importancia ya que evalúa el modelo ajustado en forma global. El objetivo de esta prueba es el de comparar los valores observados y_i en relación a las probabilidades estimadas π_i.

Sea la hipótesis:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = ... = \beta_q = 0$$

$$H_1 : \beta_j \neq 0, \; al\ menos\ un\ j=0,1,2,...,k$$

Sea la estadística:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

Bajo la hipótesis H_0 que el modelo se ajusta bien a los valores observados. La X^2 tiene una distribución asintótica chi-cuadrado $X^2_{(n-(k+1))}$

Si el p-valor asociado a la estadística es un valor pequeño indicara que el modelo no se ajusta bien con el modelo teórico.

CAPÍTULO 3

APLICACIÓN

3.1 INTRODUCCIÓN

El presente trabajo monográfico pretende construir un modelo y determinar en cuanto influyen las variables regresoras, en este caso se está realizando un análisis con respecto a los ingresantes de la Facultad de Ciencias Administrativas de la Universidad Nacional Mayor de San Marcos y determinar que variables son significativas para conocer si el alumno proviene de un colegio particular.

Partiendo de esta premisa, se ha estudiado una serie de variables que puedan predecir, por ejemplo el INGRESO FAMILIAR, POSTULACIONES, VIVIENDA, SALUD, PAGO DE COLEGIO, PREPARACIÓN, entre otros.

A continuación se tomará una muestra de los ingresantes a la Facultad de Ciencias Administrativas correspondiente al Proceso de Admisión 2006-I; posteriormente se ira analizando con el Análisis de Regresión Logística.

3.2 DESCRIPCIÓN DEL PROBLEMA

Como primer paso se realizará métodos estadísticos exploratorios.

3.2.1 DISEÑO MUESTRAL

3.2.1.1 POBLACIÓN OBJETIVO:

Los ingresantes a la Facultad de Ciencias Administrativas correspondiente al proceso de admisión 2006-I.

3.2.1.2 UNIDAD ELEMENTAL:

Un ingresante a la Facultad de Ciencias Administrativas correspondiente al Proceso de Admisión 2006-I.

3.2.1.3 MARCO MUESTRAL:

Se utilizará como marco muestral el listado de todos los ingresantes a la Facultad de Ciencias Administrativas del Proceso de Admisión 2006-I, obtenidos de la Oficina Central de Admisión de la UNMSM.

3.2.1.4 TIPO DE MUESTREO:

En la siguiente investigación se utilizará un Muestreo Aleatorio Simple proporcional al tamaño.

3.2.1.5 PROCEDIMIENTO DE SELECCIÓN DE LA MUESTRA:

Se tomo una muestra utilizando el diseño de Muestreo Aleatorio Simple proporcional al tamaño, en este caso la Facultad de Ciencias Administrativas está conformada por tres Escuelas Académico Profesionales.

A continuación se presenta la formula del tamaño de muestra total:

$$n = \frac{NP(1-P)}{(N-1)D + P(1-P)}$$

Donde $D = \left(\frac{E}{Z}\right)^2$; asimismo se tomará como margen de error de un 5% (E=0.05) y un nivel de confianza del 95% (Z=0.05).

Donde:

N =152 ingresantes a la Facultad de Ciencias Administrativas (2006-I)

n = Total de ingresantes de la muestra

E = Margen de error (E = 5%)

Reemplazando los valores se obtiene lo siguiente:

Tabla N °1

Escuela Académico Profesional	Total de ingresantes	Porcentaje	Tamaño de la muestra
Administración	72	47.4	52
Adm. de Turismo	27	17.8	19
Adm. de Negocios Internacionales	53	34.9	38
Total	152	100	109

En la tabla N °1 se muestra la distribución de la muestra según Escuela Académico Profesional considerando los porcentajes de ingresantes en cada una de las E.A.P.

Obtenidos los tamaños de muestra en cada una de las E.A.P. se procedió a seleccionar aleatoriamente los datos.

Para el análisis respectivo las variables en estudio consideradas son las siguientes:

A) VARIABLE DEPENDIENTE

- ✓ Tipo de Colegio: Indica el colegio de procedencia del alumno.

$$X_i = \left\{ \begin{array}{l} 0, Estatal \\ 1, Particular \end{array} \right\}$$

B) VARIABLES INDEPENDIENTES

- ✓ Egresó del Colegio: Indica el año en que egresó el alumno.

- ✓ Preparación: Indica la modalidad de preparación del alumno.

$$X_i = \left\{ \begin{array}{l} 1, Individual \\ 2, Grupo\ de\ estudio \\ 3, Academia \\ 4, Cepusm \end{array} \right\}$$

- ✓ Edad: Indica la edad del ingresante.
- ✓ Postulaciones: Indica el número de veces que ha postulado anteriormente a la Universidad San Marcos.

$$X_i = \left\{ \begin{array}{l} 0, Ninguna\ vez \\ 1, Una\ sola\ vez \\ 2, Dos\ veces \\ 3, Tres\ ó\ más\ veces \end{array} \right\}$$

- ✓ Ingreso familiar: Indica el ingreso mensual familiar del alumno.

$$X_i = \left\{ \begin{array}{l} 1, Hasta\ 400\ nuevos\ soles \\ 2, De\ 401\ a\ 800\ nuevos\ soles \\ 3, De\ 801\ a\ 1200\ nuevos\ soles \\ 4, De\ 1201\ a\ 1600\ nuevos\ soles \\ 5, De\ 1601\ a\ más \end{array} \right\}$$

- ✓ Salud: Indica la atención de salud de la familia del alumno.

$$X_i = \left\{ \begin{array}{l} 1, Área\ de\ salud, posta \\ 2, Hospital\ nacional \\ 3, EsSalud \\ 4, Médico\ particular \end{array} \right\}$$

- ✓ Pago de colegio: Indica el promedio mensual de pago en el colegio.

$$X_i = \left\{ \begin{array}{l} 1, Hasta\ 100\ nuevos\ soles \\ 2, De\ 101\ a\ 300\ nuevos\ soles \\ 3, De\ 301\ a\ 400\ nuevos\ soles \\ 4, De\ 401\ a\ más \end{array} \right\}$$

- ✓ Vivienda: Indica la ubicación de la vivienda del alumno.

$$X_i = \left\{ \begin{array}{l} 1, Urbanización\ residencial \\ 2, Urbanización\ popular, conjunto\ habitacional \\ 3, Qu\,\mathrm{int}\,a\ ó\ callejón \\ 4, Asentamiento\ humano \end{array} \right\}$$

- ✓ Nota final: Puntaje obtenido en el Proceso de Admisión.

A continuación se aplicará el Análisis de Regresión Logística que no deja ser un caso particular del modelo discriminante, en la cual la variable dependiente tiene exclusivamente dos categorías, a su vez que parte unos supuestos menos restrictivos y permite introducir variables independientes categóricas en el modelo.

A partir de (x_{i1},.....,x_{ip}), i=1,.....n, es una muestra de n observaciones de las variables independientes X_{i1},.....,X_{ip}, en los dos grupos de individuos establecidos por los dos valores de la variable dependiente *Y*, lo cual se trata de obtener una combinación lineal de las variables independientes que permita estimar las probabilidades de que un ingresante pertenezca a cada una de las dos subpoblaciones.

Como primer paso es obtener una combinación lineal de las variables independientes: EGRESÓ DEL COLEGIO, PREPARACIÓN, EDAD, POSTULACIONES, INGRESO FAMILIAR, SALUD, PAGO DE COLEGIO, VIVIENDA Y NOTA FINAL que permita estimar las probabilidades de pertenecer a cada uno de los grupos determinados por la variable dependiente TIPO DE COLEGIO.

La probabilidad de que un ingresante pertenezca al segundo grupo, $\pi(x)$ vendrá dada por:

$$\pi(x) = \frac{1}{1+e^{-z}}$$

Siendo z la combinación lineal:

$$z = \beta_1 Egreso\,del\,Colegio + \beta_2\,\mathrm{Pr}\,eparación + \beta_3 Edad + \beta_4 Postulaciones + \beta_5 Ingreso\,familiar + \beta_6 Salud + \beta_7 Pago\,de\,colegio + \beta_8 Vivienda + \beta_9 Nota\ final + \beta_0$$

Donde $\beta_0, \beta_1, ...\beta_p$ son los parámetros a estimar. La probabilidad de que para el i-ésimo ingresante provenga de un colegio particular.

$$\pi(x_i) = \frac{1}{1 + e^{-(\beta_1 Egreso\ del\ Colegio + ... + \beta_9 Nota\ final + \beta_0)}}$$

Si dicha probabilidad es igual o superior a 0.5 el ingresante será clasificado en la categoría “PARTICULAR”, caso contrario será clasificado en la categoría “ESTATAL”.

El objetivo de este trabajo es estimar a través del Análisis de Regresión Logística, para cualquier ingresante y determinar la probabilidad de que colegio proviene el alumno, para ello se procederá estimar una función *z* tal que no solo permita estimar la probabilidad de que colegio provienen, si no además que garantice de alguna manera que, cuando se trate de ingresantes para los que se desconoce a cual de los grupos pertenecen, en términos de las probabilidades estimadas sean correctas.

Si el porcentaje de ingresantes clasificados correctamente es elevado, es de esperar que la función *z* proporcione resultados más precisos en relación al TIPO DE COLEGIO para cualquier ingresante.

Antes de proceder el ingreso de los datos al software estadístico SPSS se verificará si algunas variables presentan dos ó más categorías, lo cual conllevará a la generación de nuevas variables ficticias (dummy) tantas variables como el total de categorías menos uno.

Una vez obtenidos los datos a través del software estadístico, se aplicara el análisis de regresión logística sobre las variables independientes EGRESÓ DEL COLEGIO, PREPARACIÓN, EDAD, POSTULACIONES, INGRESO

FAMILIAR, SALUD, PAGO DE COLEGIO, VIVIENDA, NOTA FINAL y la variable dependiente TIPO DE COLEGIO.

Se tomará como punto de corte: 0.5 y un nivel de significancia α = 0.05

Los resultados se disponen en los siguientes cuadros:

Cuadro N° 1

Resumen del procesamiento de los casos

Casos no ponderados		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	109	100.0
	Casos perdidos	0	.0
	Total	109	100.0
Casos no seleccionados		0	.0
Total		109	100.0

El cuadro indica el número de datos que representa la muestra de los ingresantes de la Facultad de Ciencias Administrativas, asimismo se puede observar que no hay casos perdidos.

Cuadro N° 2

Codificación de la variable dependiente: Tipo de colegio

Valor original	Valor interno
Estatal	0
Particular	1

La variable dependiente estará representada por la variable categórica TIPO DE COLEGIO codificada por dos valores: 0 Estatal y 1 Particular.

Cuadro N° 3

Creación de variables dummy

		Frecuencia	Codificación			
			(1)	(2)	(3)	(4)
Ingreso familiar	Hasta 400 soles	8	1.000	.000	.000	.000
	De 401 a 800 soles	44	.000	1.000	.000	.000
	De 801 a 1200 soles	35	.000	.000	1.000	.000
	De 1201 a 1600 soles	12	.000	.000	.000	1.000
	De 1601 a más	10	.000	.000	.000	.000
Postulaciones	Ninguna vez	33	1.000	.000	.000	
	Una sola vez	40	.000	1.000	.000	
	Dos veces	14	.000	.000	1.000	
	Tres o más veces	22	.000	.000	.000	
Vivienda	Urb. Residencial	46	1.000	.000	.000	
	Urb. Popular / Conj. Habitacional	47	.000	1.000	.000	
	Quinta o callejón	3	.000	.000	1.000	
	Asentamiento humano	13	.000	.000	.000	
Salud	Área de salud / Posta	36	1.000	.000	.000	
	Hospital nacional	29	.000	1.000	.000	
	EsSalud	33	.000	.000	1.000	
	Médico particular	11	.000	.000	.000	
Pago de colegio	Hasta 100 soles	68	1.000	.000	.000	
	De 101 a 300 soles	37	.000	1.000	.000	
	De 301 a 400 soles	3	.000	.000	1.000	
	De 401 a más soles	1	.000	.000	.000	
Preparación	Individual	18	1.000	.000	.000	
	Grupo de estudio	3	.000	1.000	.000	
	Academia	65	.000	.000	1.000	
	Cepusm	23	.000	.000	.000	

En el cuadro N °3 se observa que la variable INGRESO FAMILIAR será necesario generar cuatro variables ficticias: Ingreso familiar (1), Ingreso familiar (2), Ingreso familiar (3) e Ingreso familiar (4), una por cada uno de los cuatro primeros códigos numéricos, tales que sus valores son iguales a 1 cuando el valor en la variable INGRESO FAMILIAR sea igual al código correspondiente y a 0 en el resto de los casos. Además, las cuatro nuevas

variables tomarán valor 0 cuando el valor de INGRESO FAMILIAR sea igual a “De 1601 a más”.

Bloque 0: Bloque inicial

Cuadro N° 4

Tabla de clasificación[a,b]

			Pronosticado		
			Tipo de colegio		Porcentaje correcto
	Observado		Estatal	Particular	
Paso 0	Tipo de colegio	Estatal	57	0	100.0
		Particular	52	0	.0
	Porcentaje global				52.3

a. En el modelo se incluye una constante.

b. El valor de corte es 0.5

En este cuadro se observa la clasificación de los datos sin considerar información cualquiera de las variables independientes o lo que es equivalentemente, solo el número de casos en cada grupo, el porcentaje de casos correctamente clasificados es 52.3%, posteriormente se irán añadiendo las variables más significativas.

Cuadro N° 5

Variables en la ecuación

		B	S.E.	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-.092	.192	.229	1	.632	.912

Aquí se propone un modelo nulo logit $(\pi) = \beta_0$, lo cual indica que el Exp (B) no tiene mucha relevancia.

SELECCIÓN DE VARIABLES

Se utilizará el método Forward Stepwise para la selección de variables, que nos permitirá que variables independientes aportan más en la probabilidad de pertenecer al grupo si proviene de un colegio particular o no.

Cuadro N° 6

Variables que no están en la ecuación

			Puntaje	gl	Sig.
Paso 0	Variables	Egreso del col.	10.059	1	.002
		Preparación	18.836	3	.000
		Preparación(1)	7.815	1	.005
		Preparación(2)	.445	1	.505
		Preparación(3)	18.516	1	.000
		Edad	9.053	1	.003
		Postulaciones	3.533	3	.317
		Postulaciones (1)	3.157	1	.076
		Postulaciones (2)	.186	1	.667
		Postulaciones (3)	.926	1	.336
		Ingreso familiar	23.947	4	.000
		Ingreso familiar(1)	1.784	1	.182
		Ingreso familiar(2)	9.755	1	.002
		Ingreso familiar(3)	.015	1	.901
		Ingreso familiar(4)	10.446	1	.001
		Salud	3.439	3	.329
		Salud(1)	2.897	1	.089
		Salud(2)	.256	1	.613
		Salud(3)	.275	1	.600
		Pago de colegio	65.487	3	.000
		Pago de col(1)	65.477	1	.000
		Pago de col(2)	55.217	1	.000
		Pago de col(3)	3.382	1	.066
		Vivienda	13.519	3	.004
		Vivienda(1)	3.853	1	.050
		Vivienda(2)	.373	1	.541
		Vivienda(3)	2.814	1	.093
		Nota_Final	7.015	1	.008
	Estadísticos globales		77.831	22	.000

En el cuadro se puede observar que la variable candidata a ser seleccionada en el primer paso, es la que presenta el mínimo p-valor asociado al estadístico Puntuación Eficiente de Rao ó también se puede considerar el que presente mayor valor estadístico, en esta caso es “Pago de colegio” con una puntuación de 65.487 y su correspondiente p-valor es “Sig.=0.000”, que es menor que α=0.05 lo cual conllevará a que sea seleccionada.

MÉTODO FORWARD PARA LA SELECCIÓN DE VARIABLES

Bloque 1: Método por pasos hacia delante (Wald)

Cuadro N° 7

Variables en la ecuación

		B	S.E.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Pagcol			23.256	3	.000	
	Pagcol(1)	-22.743	40193.049	.000	1	1.000	.000
	Pagcol(2)	-17.619	40193.049	.000	1	1.000	.000
	Pagcol(3)	.000	46410.912	.000	1	1.000	1.000
	Constante	21.203	40193.049	.000	1	1.000	2E+009
Paso 2[b]	Tipoprep			12.735	3	.005	
	Tipoprep(1)	-.120	.935	.016	1	.898	.887
	Tipoprep(2)	.150	1.515	.010	1	.921	1.161
	Tipoprep(3)	-2.988	.905	10.892	1	.001	.050
	Pagcol			22.787	3	.000	
	Pagcol(1)	-24.335	40192.991	.000	1	1.000	.000
	Pagcol(2)	-18.479	40192.991	.000	1	1.000	.000
	Pagcol(3)	.000	46410.862	.000	1	1.000	1.000
	Constante	24.191	40192.991	.000	1	1.000	3E+010
Paso 3[c]	Tipoprep			11.195	3	.011	
	Tipoprep(1)	.436	1.063	.168	1	.682	1.546
	Tipoprep(2)	.429	1.624	.070	1	.792	1.536
	Tipoprep(3)	-2.697	.940	8.226	1	.004	.067
	Pagcol			23.136	3	.000	
	Pagcol(1)	-24.337	40193.146	.000	1	1.000	.000
	Pagcol(2)	-17.972	40193.146	.000	1	1.000	.000
	Pagcol(3)	-.060	45695.337	.000	1	1.000	.941
	Nota_Final	.019	.009	4.494	1	.034	1.019
	Constante	18.990	40193.146	.000	1	1.000	2E+008

a. Variable introducida en el paso 1: Pago de colegio.

b. Variable introducida en el paso 2: Preparación.

c. Variable introducida en el paso 3: Nota_Final.

Luego de ser seleccionado la variable “PAGO DE COLEGIO” se procederá a incluir en el primer paso del cuadro N° 7 (“Variables en la ecuación”). Recordemos que la variable “Pago de colegio” era categórica y que a partir de sus valores, se crearon las variables Pago de colegio(1), Pago de colegio(2) y Pago de colegio(3), pero al ser seleccionada la variable “Pago

de colegio", las tres variables generadas entran en bloque, es decir se consideran como un solo grupo.

Continuando con el proceso de selección de variables, la candidata a ser seleccionada en el Cuadro N° 8 (Variables que no están en la ecuación) es PREPARACIÓN ya que presenta el máximo valor en el estadístico Puntuación Eficiente de Rao con 18.855 y el p-valor asociado es "Sig.=0.000" que es menor que α=0.05, por lo tanto esta variable se incluirá en el segundo paso del Cuadro N° 7 (Variables en la ecuación).

El siguiente paso sería tratar de eliminar variables en el Cuadro N° 7 (Variables en la ecuación) por ejemplo entre las variables obtenidas hasta el segundo paso, la variable a ser eliminada sería la que presente el máximo p-valor asociado al estadístico de Wald, en este caso sería "Pago de colegio(1), Pago de colegio(2) y Pago de colegio(3) ya que presentan p-valores mayores que α=0.05, pero recordemos que estas variables son tratadas en bloque, por lo que no se podría eliminar individualmente, por tanto los únicos p-valores a eliminar serían los correspondientes a PAGO DE COLEGIO y PREPARACIÓN y ninguno de estos son mayores que α=0.05, en consecuencia ninguna de las dos variables puede ser eliminada.

La siguiente candidata a seleccionar es "NOTA FINAL" ya que presenta el menor p-valor asociado al estadístico Puntuación Eficiente de Rao con "Sig.=0.026" con una puntuación de 4.925, luego será seleccionada en el tercer paso del Cuadro N °7 (Variables en la ecuación).

Volviendo al Cuadro N° 7, la candidata a ser eliminada será la que presente el mayor p-valor asociado al estadístico de Wald en este caso sería la

variable "NOTA FINAL" ya que su p-valor es 0.034, pero presenta el inconveniente que es menor que α =0.05, implica que no será eliminada.

Regresando al Cuadro N° 8 (Variables que no están en la ecuación) la candidata a ser seleccionada es "INGRESO FAMILIAR" ya que cuenta con la mayor puntuación con 2.201 pero el p-valor asociado al estadístico Puntuación Eficiente de Rao es 0.699 mayor que α =0.05 lo cual implica que no sea seleccionada la variable.

Luego de que ninguna otra variable pueda ser seleccionada o eliminada, la estimación de los parámetros de la función z se realizará a partir de los valores de las variables PREPARACIÓN, PAGO DE COLEGIO Y NOTA FINAL.

ESTIMACIÓN DE LOS PARAMETROS

La estimación de la función z a partir de los valores de las variables seleccionadas será:

$$\bar{z} = 0.436\,\mathrm{Pr}\,eparación(1) + 0.429\,\mathrm{Pr}\,eparación(2) - 2.697\,\mathrm{Pr}\,eparación(3) - 24.337\,Pagodecolegio(1) - 17.972\,Pagodecolegio(2) - 0.06\,Pagodecolegio(3) + 0.019\,Notafinal + 18.99$$

Sabemos que:

$$\lg\left(\frac{\pi(x)}{1-\pi(x)}\right) = z \qquad \longrightarrow \quad \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0}(e^{\beta_1})^{X_1}.....(e^{\beta_P})^{X_P}$$

Una expresión alternativa al modelo de regresión logística es:

$$\frac{\bar{\pi}(x)}{1-\bar{\pi}(x)} = e^{0.436\,\mathrm{Pr}\,eparación(1)}\,e^{0.429\,\mathrm{Pr}\,eparación(2)}\,e^{-2.697\,\mathrm{Pr}\,eparación(3)}\,e^{-24.337\,Pagodecolegio(1)}\,e^{-17.972\,Pagodecolegio(2)}\,e^{-0.06\,Pagodecolegio(3)}\,e^{0.019\,Notafinal}\,e^{18.99}$$

Cuadro N °8

Variables que no están en la ecuación

			Puntuación	gl	Sig.
Paso 1	Variables	Egreso del col.	4.461	1	.035
		Preparación	18.855	3	.000
		Prep(1)	3.810	1	.051
		Prep(2)	1.477	1	.224
		Prep(3)	18.793	1	.000
		Edad	1.177	1	.278
		Postulaciones	.588	3	.899
		Postulac1)	.200	1	.655
		Postulac(2)	.100	1	.752
		Postulac(3)	.323	1	.570
		Ingreso familiar	5.823	4	.213
		Ingfam(1)	1.195	1	.274
		Ingfam(2)	.045	1	.833
		Ingfam(3)	.054	1	.816
		Ingfam(4)	4.867	1	.027
		Salud	2.705	3	.439
		Salud(1)	1.788	1	.181
		Salud(2)	.591	1	.442
		Salud(3)	.001	1	.970
		Vivienda	2.028	3	.567
		Vivienda(1)	.084	1	.772
		Vivienda(2)	.704	1	.401
		Vivienda(3)	.673	1	.412
		Nota_Final	7.647	1	.006
	Estadísticos globales		26.676	19	.112
Paso 2	Variables	Egreso del col.	3.313	1	.069
		Edad	.593	1	.441
		Postulaciones	.729	3	.866
		Postulac(1)	.294	1	.588
		Postulac(2)	.111	1	.740
		Postulac(3)	.034	1	.854
		Ingreso familiar	1.484	4	.829
		Ingfam(1)	.129	1	.720
		Ingfam(2)	.136	1	.712
		Ingfam(3)	.384	1	.535
		Ingfam(4)	.890	1	.346
		Salud	.786	3	.853
		Salud(1)	.182	1	.669
		Salud(2)	.747	1	.388
		Salud(3)	.162	1	.687
		Vivienda	.874	3	.832
		Vivienda(1)	.638	1	.425
		Vivienda(2)	.728	1	.393
		Vivienda(3)	.140	1	.709
		Nota_Final	4.925	1	.026
	Estadísticos globales		8.260	16	.941
Paso 3	Variables	Egresó del col.	.150	1	.698
		Edad	.012	1	.913
		Postulaciones	.006	3	1.000
		Postulac(1)	.003	1	.957
		Postulac(2)	.004	1	.948
		Postulac(3)	.002	1	.965
		Ingreso familiar	2.201	4	.699
		Ingfam(1)	.377	1	.539
		Ingfam(2)	.228	1	.633
		Ingfam(3)	.756	1	.385
		Ingfam(4)	1.318	1	.251
		Salud	.781	3	.854
		Salud(1)	.014	1	.905
		Salud(2)	.750	1	.387
		Salud(3)	.175	1	.676
		Vivienda	.900	3	.825
		Vivienda(1)	.711	1	.399
		Vivienda(2)	.698	1	.404
		Vivienda(3)	.120	1	.729
	Estadísticos globales		5.096	15	.991

Bondad de ajuste del modelo

La bondad de ajuste es analizar cuán probables son los resultados a partir del modelo ajustado.

Denominamos verosimilitud a la probabilidad de los resultados observados y se basa en comparar el número de ingresantes observados en la segunda población con el número esperado si el modelo fuera valido.

Para comprobar que el modelo es adecuado, una alternativa sería contrastar, en una única hipótesis nula, que todos los parámetros correspondientes al conjunto de variables incluidas en el modelo son iguales a cero.

Para contrastar la hipótesis nula de que, en cada etapa, para todas las variables incluidas en el modelo los parámetros asociados son nulos utilizaremos el estadístico Chi-cuadrado.

Esto nos permitirá evaluar en cada paso la mejora obtenida en el estadístico Chi-cuadrado con relación al anterior.

Cuadro N° 9

Prueba omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	78.306	3	.000
	Bloque	78.306	3	.000
	Modelo	78.306	3	.000
Paso 2	Paso	18.326	3	.000
	Bloque	96.632	6	.000
	Modelo	96.632	6	.000
Paso 3	Paso	5.174	1	.023
	Bloque	101.806	7	.000
	Modelo	101.806	7	.000

En el cuadro se puede observar que el proceso de selección el estadístico Chi-cuadrado para el modelo con la variable PAGO DE COLEGIO como única independiente es igual a 78.306, como se trata del primer paso coincide con el valor del cambio en el paso. El p-valor asociado al estadístico para el modelo es menor que 0.05, implicará que se rechace la hipótesis nula de que los parámetros asociados a las tres variables generadas a partir de los valores de PAGO DE COLEGIO son nulos.

Continuando con el segundo paso, al ingresar la variable PREPARACIÓN en el modelo, el valor del estadístico chi-cuadrado para el modelo con las variables PAGO DE COLEGIO y PREPARACIÓN aumenta hasta 96.632 produciendo un incremento de 18.326 así mismo el p-valor asociado al cambio en el paso es menor que 0.05, por lo que se rechazará la hipótesis nula de que la mejora no es significativa. Por último, en el tercer paso, al ingresar la variable NOTA FINAL en el modelo, el valor del estadístico chi-cuadrado para el modelo con las variables PAGO DE COLEGIO, PREPARACIÓN y NOTA FINAL se incrementa a 101.806 con un aumento de 5.174 también el p-valor asociado al cambio en el paso es menor que 0.05 lo cual implica que la mejora es significativa.

Cuadro N° 10

Resumen de los modelos

Paso	-2 Log de la verosimilitud	R cuadrado de Cox & Snell	R cuadrado Nagelkerke
1	72.570	.512	.684
2	54.244	.588	.784
3	49.071	.607	.810

El R cuadrado de Cox & Snell se basa en la comparación de la verosimilitud del modelo final con el modelo inicial (modelo que incluye solamente la constante). El R cuadrado de Nagelkerke consiste en una corrección del anterior.

En el Cuadro N° 10 se puede observar en el paso 3 el R cuadrado de Nagelkerke es de 0.810 por lo que podría interpretarse que el modelo de regresión logística explica el comportamiento de la variable dependiente TIPO DE COLEGIO al 81 %.

VALIDACIÓN DE RESULTADOS

La clasificación de los individuos en uno u otro grupo se realizará a partir de la probabilidad estimada de pertenecer al segundo grupo (Ingresantes provenientes de colegios particulares).

En el Cuadro N° 11 presenta el resumen de los datos clasificados, el cual se observa una mejora en cada etapa, al inicio del proceso, antes de considerar la información cualquiera de las variables independientes, el porcentaje de casos correctamente clasificados era del 52.31%.

Cuadro N° 11

Tabla de clasificación[a]

Observado			Pronosticado		
			Tipo de colegio		Porcentaje correcto
			Estatal	Particular	
Paso 1	Tipo de colegio	Estatal	56	1	98.2
		Particular	12	40	76.9
	Porcentaje global				88.1
Paso 2	Tipo de colegio	Estatal	55	2	96.5
		Particular	11	41	78.8
	Porcentaje global				88.1
Paso 3	Tipo de colegio	Estatal	54	3	94.7
		Particular	5	47	90.4
	Porcentaje global				92.7

a. El valor de corte es 0.5

Al introducir la variable PAGO DE COLEGIO el porcentaje aumentó al 88.1%, luego al considerar la otra variable PREPARACIÓN en el segundo paso no vario y por último en el paso 3 se ingreso la otra variable NOTA FINAL más las otras variables anteriores el porcentaje de casos correctamente clasificados es igual a 92.7%, lo cual conlleva un incremento total del inicio al final del proceso, del 40.39%.

Para una mejor visualización se presentaran algunos gráficos en el anexo N° 1 y una tabla de la prueba de Hosmer and Lemeshow en el anexo N° 2.

Y por último en el anexo N° 3 se muestran los valores de las probabilidades de pertenecer al segundo grupo y el grupo en el que ha sido clasificado el caso, respectivamente.

La sensibilidad del modelo es 90.4%. y la especificidad es del 94.7% lo cual indica que el modelo ajustado es más específico que sensible.

Entonces podemos concluir que la información aportada por las variables PAGO DE COLEGIO, PREPARACIÓN Y NOTA FINAL es significativa.

CONCLUSIONES

- De acuerdo a los resultados obtenidos con el Análisis de Regresión Logística se concluye que las variables más significativas relacionadas a los alumnos que proceden de colegios particulares son: el tipo de preparación, el pago al colegio y por ultimo el rendimiento que obtuvo en la nota final del examen de admisión.

  Asimismo se observa que las variables ingreso familiar, ubicación de la vivienda, edad, salud, año que egresó del colegio y el número de veces que ha postulado a la universidad San Marcos no son características de los alumnos que provengan de colegios particulares.

  Haciendo un análisis de cada variable vemos que el tipo de preparación influye mucho si el alumno procede de un colegio particular, tal es el caso de los alumnos que se prepararon en el Centro Preuniversitario de San Marcos (Cepusm).

  Otra variable influyente es el pago de colegio, en la cual las categorías con mayor pago en el colegio tienden a diferenciar los colegios particulares de los estatales.

  Por ultimo, un dato importante es la presencia de la variable nota final que hace una diferencia al tipo de colegio de procedencia, este es un caso particular, ya que en otras facultades la variable no es muy relevante.

Por otro lado la procedencia de los ingresantes según ubicación de la vivienda como urbanizaciones residenciales, urbanizaciones

populares, conjunto habitacionales y asentamientos humanos no fue relevante para la construcción del modelo.

La variable ingreso familiar que se clasifica en varias categorías de acuerdo a la situación económica, al final demostró no ser influyente en el análisis por lo que no se consideró en el modelo.

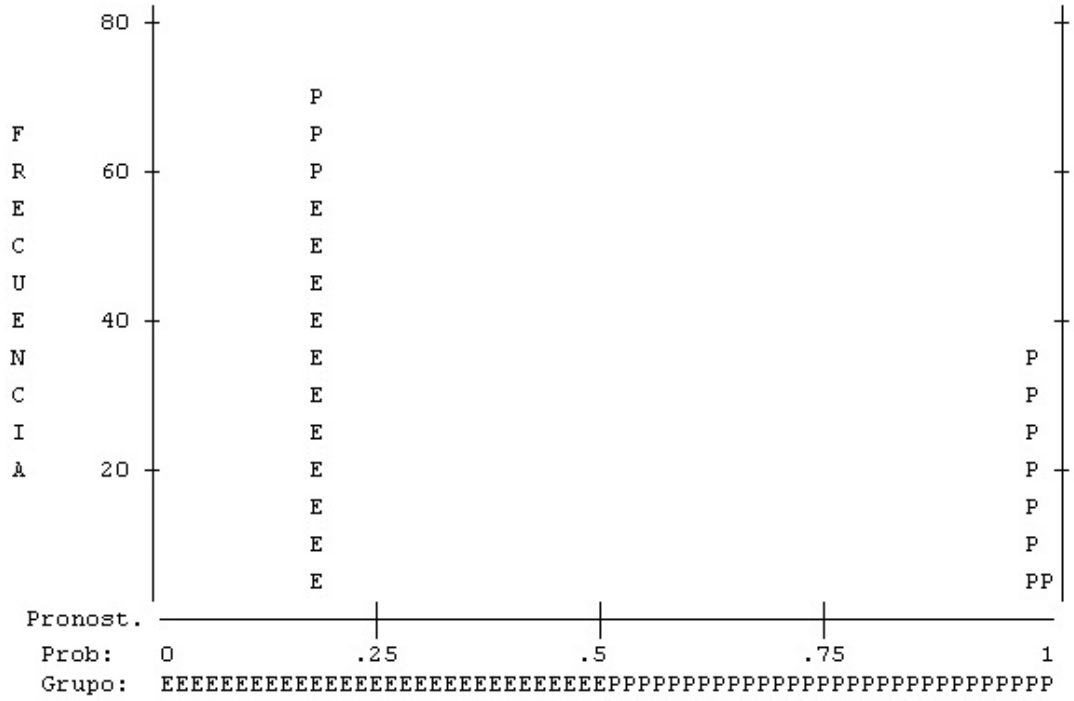
Sugerencias:

Debido a que la base de datos de los ingresantes de la Facultad de Ciencias Administrativas es muy pequeña se recomienda para posteriores investigaciones anexar una encuesta con respecto a su situación laboral, a la vez también realizar algunas coordinaciones con el Sistema Único de Matrícula para que nos brinde información de los alumnos en cuanto a su promedio ponderado y el número de créditos aprobados y comprobar si existe alguna relación entre el rendimiento académico y los colegios particulares.

ANEXO N °1

Paso: 1

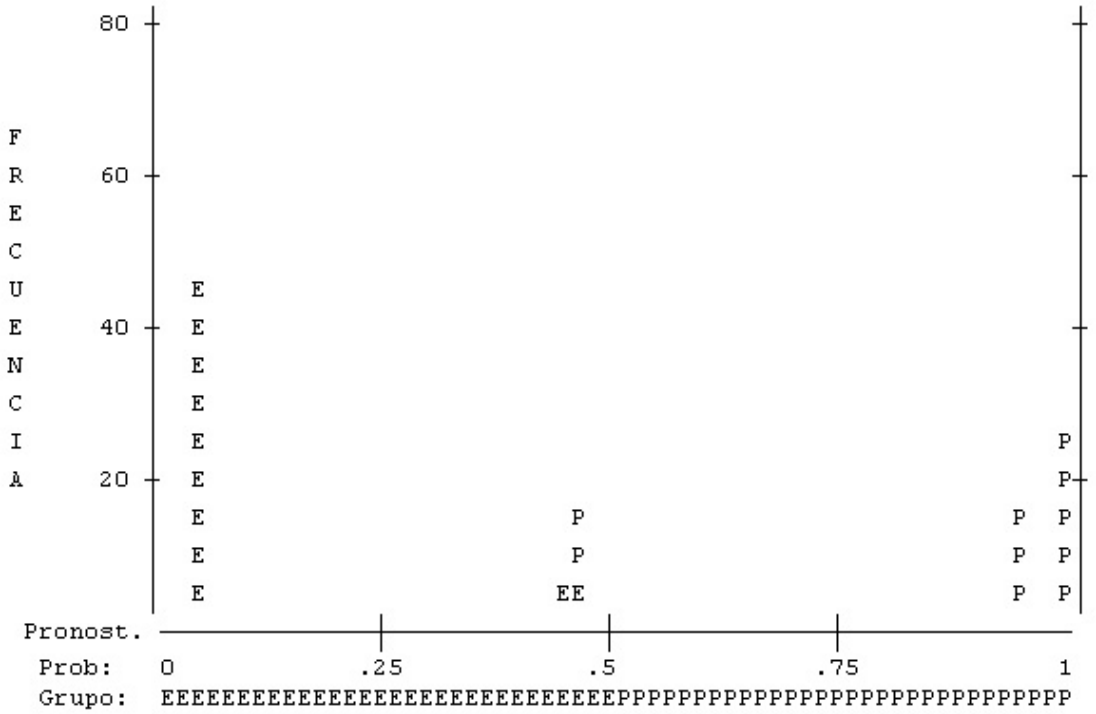
Probabilidades estimadas y los grupos observados



Valor de corte es 0.5
Simbolos: E - Estatal
P - Particular
Cada simbolo representa 5 Casos.

Paso: 2

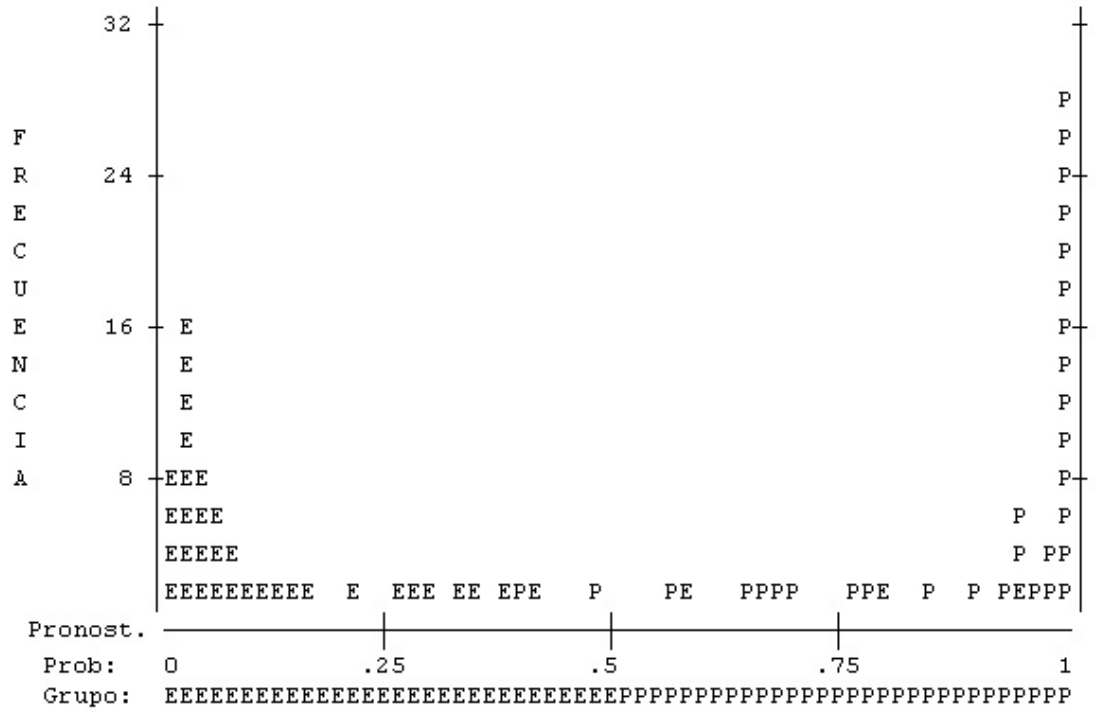
Probabilidades estimadas y los grupos observados



Valor de corte es 0.5
Simbolos: E - Estatal
P - Particular
Cada simbolo representa 5 Casos.

Paso: 3

Probabilidades estimadas y los grupos observados



Valor de corte es 0.5
Simbolos: E - Estatal
P - Particular
Cada simbolo representa 2 Casos.

ANEXO N °2

Prueba de Hosmer and Lemeshow

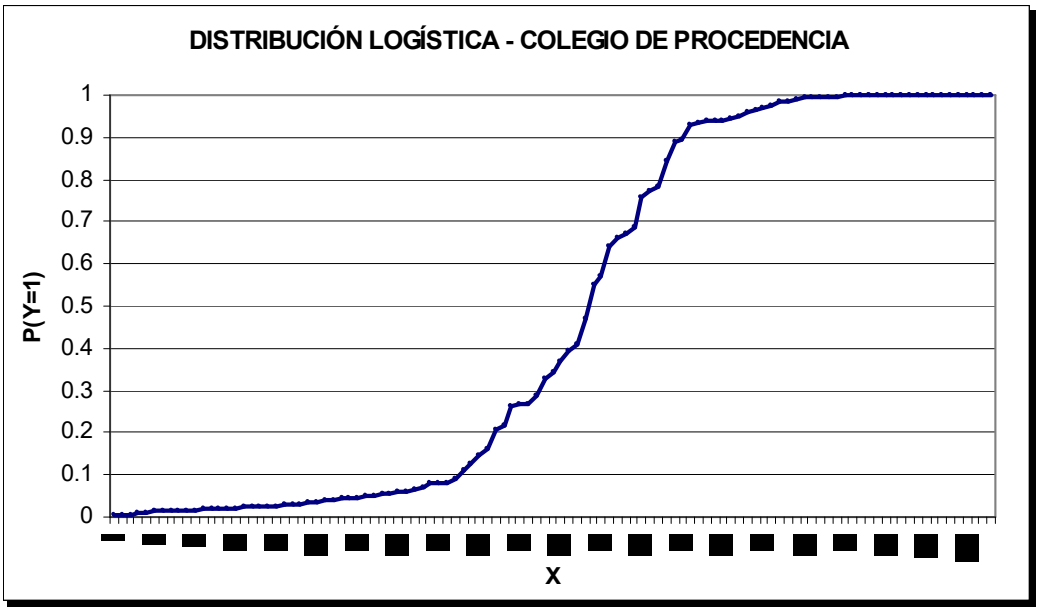
		colpro = Estatal		colpro = Particular		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	56	56.000	12	12.000	68
	2	1	1.000	40	40.000	41
Paso 2	1	44	44.077	2	1.923	46
	2	4	3.959	3	3.041	7
	3	7	6.967	6	6.033	13
	4	2	1.921	15	15.079	17
	5	0	.041	11	10.959	11
	6	0	.036	15	14.964	15
Paso 3	1	10	10.866	1	.134	11
	2	11	10.745	0	.255	11
	3	10	10.547	1	.453	11
	4	11	10.185	0	.815	11
	5	10	8.392	1	2.608	11
	6	3	4.822	8	6.178	11
	7	2	1.187	9	9.813	11
	8	0	.232	11	10.768	11
	9	0	.021	11	10.979	11
	10	0	.003	10	9.997	10

ANEXO N °3

Resumen de casos

n	Probabilidad de pertenecer al segundo grupo	Grupo al que ha sido clasificado
1	0.21466	0
2	0.26533	0
3	1	1
4	0.01505	0
5	0.03692	0
6	0.0179	0
7	0.99948	1
8	0.46942	0
9	0.34458	0
10	0.16013	0
11	0.08982	0
12	0.08255	0
13	0.00751	0
14	0.00758	0
15	0.99642	1
16	0.99883	1
17	0.3934	0
18	0.78419	1
19	0.04418	0
20	0.04187	0
21	0.00725	0
22	0.01701	0
23	0.03059	0
24	0.124	0
25	0.68558	1
26	0.01356	0
27	0.99987	1
28	0.40923	0
29	0.64391	1
30	0.93941	1
31	0.0202	0
32	0.02203	0
33	0.99694	1
34	0.08087	0
35	0.07845	0
36	0.77244	1
37	0.99199	1
38	0.99839	1
39	0.94085	1
40	0.02327	0
41	0.01726	0
42	0.99839	1
43	0.57044	1
44	0.93887	1
45	0.99964	1
46	0.67017	1
47	0.04866	0
48	0.3683	0
49	0.02379	0
50	0.55085	1
51	0.98583	1
52	0.03504	0
53	0.99756	1
54	0.94628	1

n	Probabilidad de pertenecer al segundo grupo	Grupo al que ha sido clasificado
55	0.99847	1
56	0.33054	0
57	0.97244	1
58	0.11257	0
59	1	1
60	0.02716	0
61	0.02896	0
62	0.20496	0
63	0.14498	0
64	0.02029	0
65	0.05066	0
66	0.95915	1
67	1	1
68	0.03237	0
69	0.7563	1
70	0.26187	0
71	0.05844	0
72	0.04759	0
73	0.88974	1
74	0.01749	0
75	0.06299	0
76	0.99457	1
77	0.05492	0
78	0.99929	1
79	0.92698	1
80	0.99786	1
81	0.06503	0
82	0.99822	1
83	0.28859	0
84	0.66001	1
85	0.84412	1
86	1	1
87	0.99746	1
88	0.97128	1
89	0.02711	0
90	0.89375	1
91	0.01524	0
92	0.99861	1
93	0.07273	0
94	0.99954	1
95	0.04048	0
96	0.26872	0
97	0.00658	0
98	0.02434	0
99	0.00973	0
100	0.98638	1
101	0.04486	0
102	0.99929	1
103	0.93594	1
104	0.94709	1
105	0.96675	1
106	0.05706	0
107	0.99829	1
108	0.01946	0
109	0.99606	1


DISTRIBUCIÓN LOGÍSTICA - COLEGIO DE PROCEDENCIA
P(Y=1)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
X

BIBLIOGRAFIA

[1] Visauta Vinacua (1998). Análisis estadístico con SPSS para Windows
McGraw – Hill / Interamericana de España, S.A.U. Madrid. España.

[2] Pia Veldt Larsen (2006). Logistic Regression
Department of Statistics. University of Southern Denmark

[3] David G. Kleinbaum (1994). Logistic Regression, a self-Learning Text
Springer - Verlag, New York

[4] Magdalena Ferrán Aranaz (2001). SPSS para Windows. Análisis Estadístico.
McGraw – Hill / Interamericana de España, S.A.U. Madrid. España.

[5] Carlos Veliz (2000). Regresión con datos categóricos.
Primera edición.

[6] Hosmer, D. Lemeshow, S. (1989). Applied Logistic Regression
John Wiley & Sons. New York, U.S.A.