CAPÍTULO 2

MODELO DE REGRESIÓN LOGÍSTICA

2.1 INTRODUCCIÓN

La Regresión Logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas.

El Análisis de Regresión Logística tiene la misma estrategia que el Análisis de Regresión Lineal Múltiple, el cual se diferencia esencialmente del Análisis de Regresión Logística por que la variable dependiente es métrica; en la práctica el uso de ambas técnicas tienen mucha semejanza, aunque sus enfoques matemáticos son diferentes.

La variable dependiente o respuesta no es continua, sino discreta (generalmente toma valores 1,0). Las variables explicativas pueden ser cuantitativas o cualitativas; y la ecuación del modelo no es una función lineal de partida, sino exponencial; si bien, por sencilla transformación logarítmica, puede finalmente presentarse como una función lineal.

Así pues el modelo será útil en frecuentes situaciones prácticas de investigación en que la respuesta puede tomar únicamente dos valores: 1, presencia (con probabilidad p); y 0, ausencia (con probabilidad $1-p$).

El modelo será de utilidad puesto que, muchas veces, el perfil de variables puede estar formado por caracteres cuantitativos y cualitativos; y se pretende hacer participar a todos ellos en una única ecuación conjunta.

El modelo puede acercarse mas a la realidad ya que muchos fenómenos, como los del campo epidemiológico, se asemejan más a una curva que a una recta. Además la curva exponencial elegida como mejor ajuste, puede ser transformada logarítmicamente en una ecuación lineal de todas las

variables, siendo así que el aparato matemático estudiado para la regresión lineal múltiple será aplicable; aunque el investigador tenga, al final, que deshacer la transformación para interpretar sus conclusiones.

Si para el Modelo de Regresión Logística una variable regresora de tipo categórica tiene c niveles habrá que generar c-1 variables ficticias (dummy) a fin que todas las posibilidades de la variable queden bien representadas en el modelo logístico.

Cuando todas las variables regresoras son categóricas entonces se usa el modelo Log lineal, ver Mc Cullagh (1983).

2.2 OBJETIVOS DE LA REGRESIÓN LOGÍSTICA

El objetivo primordial de esta técnica es el de modelar como influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular.

Sistemáticamente tiene dos objetivos:

1. Investigar como influye en la probabilidad de ocurrencia de un suceso, la presencia o no de diversos factores y el valor o nivel de los mismos.
2. Determinar el modelo más parsimonioso y mejor ajustado que siendo razonable describa la relación entre la variable respuesta y un conjunto de variables regresoras.

2.3 REGRESIÓN LOGÍSTICA Y OTROS MÉTODOS RELACIONADOS

El objetivo general de la Regresión Logística es predecir la probabilidad de un evento de interés en una investigación, así como identificar las variables predictoras útiles para tal predicción.

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

El Análisis de Regresión Lineal Múltiple y el Análisis Discriminante son dos métodos eficaces pero plantean problemas cuando la variable respuesta es binaria.

En el Análisis de Regresión Lineal Múltiple cuando la variable respuesta toma solo dos valores, se violan los supuestos de necesarios para efectuar inferencias, los problemas que se plantean son:

1. La distribución de los errores aleatorios no es normal.
2. Los valores predictados no pueden ser interpretados como probabilidades como en la Regresión Logística, porque no toman valores dentro del intervalo [0,1].

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, pueden ser diferentes también; para que la regla de predicción sea óptima, Johnson (1982).

La Regresión Logística requiere mucho menos supuestos que el AD, por ello cuando satisfacen los supuestos requeridos para el AD, la Regresión Logística trabaja bien.

A continuación se describirá un paralelo entre la Regresión Lineal Múltiple y la Regresión Logística, debido a que ambos tienen el mismo objetivo, predecir la variable respuesta a partir de las variables regresoras.

2.4 REVISIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

La diferencia básica entre los modelos del Análisis de Regresión Lineal Múltiple y de la Regresión Logística es naturaleza de la relación entre la variable respuesta y las variables regresoras.

Para el Análisis de Regresión Lineal Múltiple, consideremos y una variable respuesta cuantitativa y $x_1, x_2, ..., x_k$ variables regresoras o llamadas también explicativas; y se desea describir la relación que hay entre la variable respuesta y las variables explicativas, si entre la variable respuesta y las regresoras hay una relación lineal se espera que:

$$E(y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} \quad , (2.1)$$

para i=1, 2, ..., n

donde:

y_i es el valor de la variable respuesta cuantitativa para el i-ésimo objeto.

b_j ; j = 0, 1, 2,....,k son los parámetros.

Siendo n el número de objetos u observaciones.

Aunque (2.1) no de valores exactos, se espera que varíe linealmente con las variables regresoras, esto es:

$$E(y_i \mid \underset{\rho}{x}_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik},$$

para i=1, 2, ..., n

(2.2)

siendo $\underset{\rho}{x}_i^T = (x_{i0}, x_{i1}, x_{i2}, ..., x_{in})$ la i-ésima observación, con $x_{i0} = 1$, (2.2) toma valores reales y en forma vectorial es:

$$E(y_i \mid \underset{\rho}{x}_i) = \underset{\rho}{x}_i^T \underset{\rho}{b} \quad (2.3)$$

donde $\underset{\rho}{b}^T = (b_0, b_1, b_2, ..., b_k)$ es el vector de parámetros.

Pero en (2.3) hay otras variables regresoras que pueden influir linealmente sobre y_i , por tanto cada valor de y_i está variando alrededor de $E(y_i)$ a esa variación lo denotamos con e_i , esto es:

$$\begin{aligned} e_i &= y_i - E(y_i \mid \underset{\rho}{x}_i) \\ &= y_i - \underset{\rho}{x}_i^T \underset{\rho}{b} \end{aligned} \qquad (2.4)$$

de (2.4):

$$y_i = \underset{\rho}{x}_i^T \underset{\rho}{b} + e_i \qquad (2.5)$$

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + e_i \quad ,$$

para i=1, 2, ..., n

(2.6)

llamado Modelo de Regresión Lineal Múltiple poblacional, a e_i se le llama error aleatorio y tiene las siguientes propiedades:

$$\begin{aligned} &E(e_i) = 0 \\ &V(e_i) = s^2 \\ &Cov(e_i, e_j) = 0 \quad \forall\, i \neq j \\ &Cov(e_i, X_j) = 0 \end{aligned} \qquad (2.7)$$

las variables regresoras no son variables aleatorias y el comportamiento de y es la respuestas a aquellas, así mismo e_i es una variable aleatoria no observable.

Generalizando el Modelo de Regresión Lineal Múltiple, (2.6), mediante el álgebra matricial está dada por:

$$\underset{\rho}{y} = X\underset{\rho}{b} + \underset{\rho}{e} \qquad (2.8)$$

donde:

$\underset{\rho}{y}^T = (y_1, y_2, ..., y_n)$, vector de variables respuestas observadas

$X = (1, \underset{\rho}{x}_1, \underset{\rho}{x}_2, ..., \underset{\rho}{x}_k)$ matriz de rango completo y con

$$\underset{\sim}{x}_i^T = (1, x_{i1}, x_{i2}, ..., x_{ik})$$

$$\underset{\sim}{b}^T = (b_0, b_1, ..., b_k)$$

$$\underset{\sim}{e}^T = (e_1, e_2, ..., e_n)$$

así mismo (2.1) en forma matricial es:

$$E(\underset{\sim}{y}) = X\underset{\sim}{b} \quad (2.9)$$

el objetivo es estimar los parámetros del modelo (2.6), los mismos que son estimados mediante el método de mínimos cuadrados.

Sea $\hat{y}_i$ la estimación de y_i, entonces:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + ... + \hat{b}_k x_{ik},$$

para i=1, 2, ..., n (2.10)

o equivalentemente:

$$\hat{y}_i = \underset{\sim}{x}_i^T \underset{\sim}{\hat{b}} \quad (2.11)$$

siendo:

$\underset{\sim}{\hat{b}}^T = (\hat{b}_0, \hat{b}_1, ..., \hat{b}_k)$, el vector de parámetros estimados.

(2.10) en su forma matricial es:

$$\underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{b}} \quad (2.12)$$

los residuos ordinarios r_i es la contraparte muestral de e_i y está dado por:

$r_i = y_i - \hat{y}_i$ para i=1, 2, ..., n (2.13)

en forma vectorial es:

$$\underset{\sim}{r} = \underset{\sim}{y} - \underset{\sim}{\hat{y}} \quad (2.14)$$

El estimador de $\underset{\sim}{b}$ se obtiene usando el método de mínimos cuadrados, ver que consiste en minimizar la suma de cuadrados del error y está dada por:

$$SCE = \sum_{i=1}^{n} e_i^2 = \underset{\sim}{e}^T \underset{\sim}{e} = \left(\underset{\sim}{y} - \underset{\sim}{\hat{y}}\right)^T \left(\underset{\sim}{y} - \underset{\sim}{\hat{y}}\right)$$

con respecto a $\underset{\sim}{b}$, esta suma de cuadrados se expresa en forma cuadrática como::

$$\left(\underset{\sim}{y} - X\underset{\sim}{b}\right)^T \left(\underset{\sim}{y} - X\underset{\sim}{b}\right) \quad (2.15)$$

al minimizarlo se obtiene que:

$$\left(X^T X\right)\underset{\sim}{\hat{b}} = X^T \underset{\sim}{y} \quad (2.16)$$

llamada ecuaciones normales.

Como $\left(X^T X\right)$ es invertible, por que es simétrica de tamaño (k+1)x(k+1) y de rango completo, entonces la solución del sistema lineal es:

$$\underset{\sim}{\hat{b}} = \left(X^T X\right)^{-1} X^T \underset{\sim}{y} \quad (2.17)$$

que es sensible a observaciones pobremente ajustados y a los puntos extremos de X, Montgomery y Peck (1992).

El vector (2.12) de valores estimados para el vector de variables respuesta es:

$$\underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{b}}$$

$$\underset{\sim}{\hat{y}} = X\left(X^T X\right)^{-1} X^T \underset{\sim}{y} \quad (2.18)$$

a la matriz $X\left(X^T X\right)^{-1} X^T \underset{\sim}{y}$, se le llama matriz de cambio o de proyección denotada por H, entonces (2.18) es:

$$\underset{\sim}{\hat{y}} = H\underset{\sim}{y} \quad (2.19)$$

El vector de residuos es:

$$\underset{\sim}{r} = \underset{\sim}{y} - \underset{\sim}{\hat{y}}$$

$$\underset{\sim}{r} = \underset{\sim}{y} - H\underset{\sim}{y} \quad (2.20)$$

$$\underset{\sim}{r} = \left(I - H\right)\underset{\sim}{y} \quad (2.21)$$

$$\vec{r} = M\vec{y} \qquad (2.22)$$

El vector $\vec{r}$ describe las desviaciones de los valores observados de los ajustados y la matriz M es el subespacio en el cual cae $\vec{r}$.

El vector residual es importante para detectar puntos ´extraños¨. A la matriz H se le llama matriz sombrero o de proyección, ver Cook y Weisberg (1982). Ahora veamos como queda expresado la suma de cuadrados de los residuos, denotada por SCE:

$$SCE = (\vec{y} - \hat{\vec{y}})^T (\vec{y} - \hat{\vec{y}}) \qquad (2.23)$$

Reemplazando $\hat{\vec{y}}$ por $X\hat{\vec{b}}$: (2.23) es:

$$SCE = \left(\vec{y} - X\hat{\vec{b}}\right)^T \left(\vec{y} - \hat{\vec{b}}\right) \qquad (2.24)$$

Y reemplazando $\hat{\vec{b}}$ por $\left(X^T X\right)^{-1} X^T \vec{y}$:

$$SCE = \vec{y}^T \vec{y} - \vec{y}^T X \left(X^T X\right)^{-1} X^T \vec{y} \qquad (2.25)$$

La matriz $H = X\left(X^T X\right)^{-1} X^T$, entonces:

$$SCE = \vec{y}^T \vec{y} - \vec{y}^T H\vec{y} \qquad (2.26)$$

$$SCE = \vec{y}^T (I - H)\vec{y} \qquad (2.27)$$

Sea $M = I - H$, entonces (2.27) es:

$$SCE = \vec{y}^T M\vec{y} \qquad (2.28)$$

Bajo el supuesto que $\vec{e} \sim N\left(\vec{0}, s^2 I_n\right)$, las observaciones $y_1, y_2, ..., y_n$ son independientes y distribuidas cono una Normal n-variada con $E(\vec{y}) = X\vec{b}$ con matriz de varianzas y covarianzas $s^2 I_n$.

En la Regresión Logística, se estima la probabilidad de que un evento ocurra; es decir, el valor esperado de y dado las variables regresoras, debe tomar valores entre 0 y 1. La relación entre las variables regresoras y la dependiente no es lineal. Las estimaciones de probabilidad estarán siempre entre 0 y 1, así, el valor de la variable respuesta se puede definir como una probabilidad de que ocurra o no un evento sujeto a control.

En la Regresión Logística, se seleccionan los coeficientes, del modelo, que hacen que los resultados sean los más "probables". Como el modelo de Regresión Logística no es lineal, se requiere de un algoritmo iterativo para estimar los parámetros.

En las secciones siguientes se detallarán los aspectos teóricos y la aplicación de la Regresión Logística.

2.5 REGRESIÓN LOGÍSTICA SIMPLE

Este modelo tiene la forma

$$y_i = b_0 + b_1 x_i + e_i \quad \text{para } i = 1,2, \quad ,n \qquad (2.29)$$

De esto se deduce que:

Si $y = 1, e_i = 1 - b_0 - b_1 x_i$ (2.30)

Si $y = 0, e_i = -b_0 - b_1 x_i$ (2.31)

Por tanto e_i, no puede tener distribución normal debido a que toma valores discretos, el Modelo de Regresión Lineal Simple, no es aplicable para el caso de variable respuesta de tipo dicotómico.

En el Análisis de Regresión Lineal simple, el punto inicial del proceso de estimación del modelo es un gráfico de dispersión de la variable respuesta versus la regresora, pero este gráfico resulta limitado cuando sólo hay dos valores posibles para la variable respuesta, por tanto se debe usar otros gráficos, éstos resultan de la suavización de los valores de la variable respuesta, representando después los valores de la variable respuesta versus la regresora.

La notación que se usará en el presente trabajo para la Regresión Logística es misma que emplea Hosmer y Lemeshow (2000).

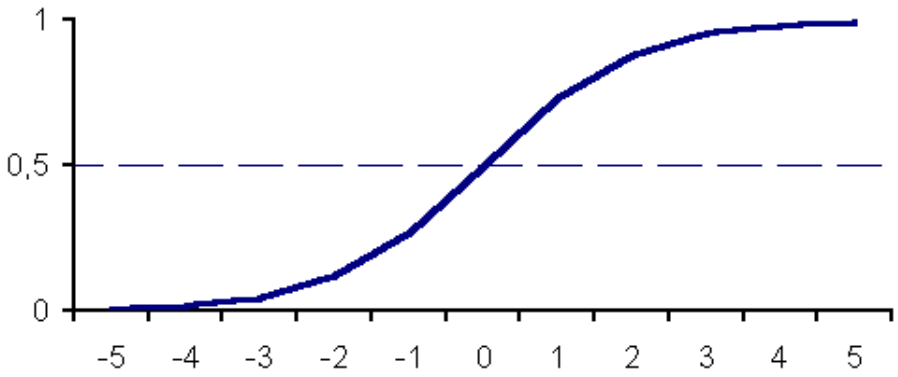
Sea $p(x) = E(y|x)$ (2.32)

Que representa la media condicional de $y = 1$ dado x, donde $p(x)$ representa la probabilidad de que ocurra $y = 1$, ciertamente no se espera que (2.32) tenga relación lineal dentro del rango de la variable regresora.

¿Qué hay de la relación entre $p(x)$ y x para valores intermedios de x? Se espera una relación curvilínea. Para cualquier valor grande de x, $p(x)$ tomará valores cercanos a 1 y para valores pequeños de x, $p(x)$ tomará

valores cercanos a cero. El gráfico que muestra el comportamiento de $p(x)$ versus x es:

FIGURA Nº 2.1



curva en forma de S o sigmoide que tiene las propiedades requeridas para $p(x)$ y que tiene las propiedades de una función de distribución de probabilidad acumulada, para esta probabilidad se usa la función de distribución acumulada de la distribución logística dada por:

$$p(x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \qquad (2.33)$$

(2.33) tiene un gráfico similar a la Figura Nº 2.1, cuando $b_0 < 0$ y $b_1 > 0$, además este modelo toma valores en el intervalo $[0,1]$.

Cuando , $P[y = 1] = 0.5$ el valor de x es: $\frac{- b_0}{b_1}$, que brinda información muy útil.

Una transformación de $p(x)$ que es lo central del estudio de la Regresión Logística es la transformación logit, transformación que se define en términos $p(x)$ como:

$$g(x) = Ln\left[\frac{p(x)}{1-p(x)}\right] = b_0 + b_1 x \qquad (2.34)$$

Lo importante de esta transformación es que tiene muchas propiedades semejantes al Modelo de Regresión Lineal simple, por ejemplo es lineal en sus parámetros, puede ser continua y puede tomar cualquier valor real dependiendo de x.

Para el Modelo de Regresión Lineal simple, la variable respuesta, de (2.4) se expresa como:

$$y = E(y|x) + e \qquad (2.35)$$

para la variable respuesta dicotómica lo expresamos como:

$$y = p(x) + e \qquad (2.36)$$

veamos que ocurre con este modelo:

Si $y = 1$, $e_i = 1 - p(x)$ y tiene probabilidad $p(x)$

Si $y = 0$, $e_i = -p(x)$ y tiene probabilidad $1 - p(x)$

Entonces e_i tiene distribución binomial con media cero y varianza $p(x)[1 - p(x)]$. Por tanto la distribución condicional de la variable respuesta tiene distribución de probabilidad binomial con media $p(x)$.

El lado izquierdo de (2.34) se llama también logaritmo de ODDS RATIO o razón de probabilidades de $y = 1$ contra $y = 0$, especificamente:

$$ODDS\ RATIO = \frac{p(x)}{1-p(x)} \qquad (2.37)$$

o también llamado razón de ventaja a favor de éxito.

2.6 REGRESIÓN LOGÍSTICA MÚLTIPLE

En esta sección se generaliza el Modelo de Regresión Logística Simple tratado en la sección anterior, es decir consideraremos más de una variable regresora, en donde por lo menos una es de tipo cuantitativo.

2.6.1 MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

Sea el vector de variables regresoras $\underset{\rho}{x}^T = (x_1, x_2, ..., x_k)$ por el momento asumiremos que están medidas por lo menos bajo escala intervalar. Sea la probabilidad condicional para que la variable respuesta sea igual a 1, denotado por:

$$P(y = 1 | \underset{\rho}{x}) = p(\underset{\rho}{x}) \qquad (2.39)$$

el logaritmo del Modelo de Regresión Logística Múltiple está dado por:

$$g(\underset{\mu}{x}_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}$$, para i=1, 2, ..., n

(2.40)

entonces el Modelo de Regresión Logística Múltiple es:

$$p(\underset{\rho}{x}) = \frac{e^{g(\underset{\mu}{x})}}{1 + e^{g(\underset{\rho}{x})}} \qquad (2.41)$$

Al igual que en el caso de Regresión Lineal Múltiple si es necesario usar variables regresoras categóricas, por ejemplo si una tiene c niveles será necesario incorporar c-1 variables ficticias o llamadas dummy., así entonces el logit para un modelo con k variables regresoras y una variable categórica, por ejemplo la j-ésima es:

$$g(\underset{\rho}{x}) = b_0 + b_1 x_{i1} + ... + \sum_{l=1}^{c-1} b_{jl} D_{jl} + b_k x_{ik} \quad ,$$

para i=1, 2, ..., n

(2.42)

2.6.2 AJUSTE DEL MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE

El ajuste se efectúa a través del uso de los métodos de máxima verosimilitud, los mismos que se encuentran en los softwares estadísticos que permiten analizar datos mediante este método.

Asumiremos que disponemos de una muestra n observaciones independientes

$(\underset{\rho}{x_i}, y_i)$, i=1,2, ..., n ; donde y_i toma valores 0 ó 1, para estimar

$\underset{\rho}{b}^T = (b_0, b_1, \ldots, b_k)$ que es el vector de parámetros desconocidos.

Para el Modelo de Regresión Lineal Múltiple se usa el método de Mínimos Cuadrados para estimar $\underset{\rho}{b}$, el cual minimiza la suma de cuadrados del error, pero cuando la variable respuesta es binaria aplicar este método no provee las mismas propiedades cuando es usado en variables respuestas continuas.

Por ello se usará el método de Máxima Verosimilitud, ya que obtendremos parámetros estimados que maximizan la probabilidad de obtener un conjunto de datos observados.

La función de verosimilitud expresa la probabilidad de los datos observados como una función de parámetros desconocidos. Los Estimadores de Máxima Verosimilitud de esos parámetros son aquellos que están en concordancia con los datos observados.

Consideremos el Modelo de Regresión Lineal Múltiple con mayor detalle, supongamos que se dispone de n objetos u observaciones donde para cada uno de ellos existe una respuesta que puede ser:

$$y_i = 0 \quad o \quad y_i = 1$$

Sea $\underset{\varpi}{y}^T = (y_1, y_2, \ldots, y_n)$ donde $y_i \sim B(1, p_i)$ y sea

$\underset{\rho}{x}_i^T = (1, x_{i1}, \ldots\ldots, x_{ik})$ la i-ésima observación para las k variables explicativas.

Así el Modelo de Regresión Logística está dada por la expresión (2.40):

$$P[y_i = 1 | \underset{\rho}{x}_i] = p(\underset{\rho}{x}_i) = \frac{e^{g(\underset{\rho}{x}_i)}}{1 + e^{g(\underset{\rho}{x}_i)}} \qquad (2.43)$$

equivalentemente

$$P[y_i = 1 | \underset{\rho}{x}_i] = \frac{Exp\left(b_0 + \sum_{j=1}^{k} b_j x_{ij}\right)}{1 + Exp\left(b_0 + \sum_{j=1}^{k} b_j x_{ij}\right)} \qquad (2.44)$$

y la probabilidad de que y_i sea igual a cero es:

$P[y_i = 0 | \underset{\rho}{x}_i] = 1 - P[y_i = 1 | \underset{\rho}{x}_i]$, *entonces :*

$$P[y_i = 0 | \underset{\rho}{x}_i] = \frac{1}{1 + Exp\left(b_0 + \sum_{j=1}^{k} b_j x_{ij}\right)} \qquad (2.45)$$

para facilitar la notación usaremos la variable indicadora $x_{i0} = 1, \quad i = 1,2,\ldots,n$.

Entonces (2.44) y (2.45) son respectivamente:

$$P[y_i = 1 | \underset{\rho}{x}_i] = p(\underset{\rho}{x}_i) = \frac{e^{\underset{\rho}{b}\underset{\rho}{x}_i^T}}{1 + e^{\underset{\rho}{b}\underset{\rho}{x}_i^T}} \qquad (2.46)$$

$$P[y_i = 0 \mid \vec{x}_i] = 1 - p(\vec{x}_i) = \frac{1}{1 + e^{\vec{b}\vec{x}_i^T}} \qquad (2.47)$$

donde: $\vec{x}_i^T = (x_{i0}, x_{i1}, \ldots, x_{ik})$, es el vector que contiene los valores de las variables explicativas

$\vec{b}^T = (b_0, b_1, \ldots, b_k)$ es el vector de parámetros a ser estimado.

El i-ésimo logito es:

$$l_i = Ln\left(\frac{p_i}{1 - p_i}\right) = \sum_{j=0}^{k} b_j x_{ij} \qquad (2.48)$$

como vemos, (2.48) es una función lineal simple del vector de observaciones $\vec{x}_i$ llamada transformación logística de la probabilidad p_i o simplemente Logit o Logito de la ecuación, a la expresión (2.48) también se le llama Modelo Logístico Lineal.

A fin de obtener la estimación máximo verosímil para el vector $\vec{b}$, escribimos la función de densidad de probabilidad del vector $\vec{y}$ el cual es proporcional a n funciones $B(1, p_i)$, esto es:

$$\begin{aligned} f(y_i; p_i) &= \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \\ &= \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i}\right)^{y_i} (1 - p_i) \\ &= \left\{\prod_{i=1}^{n} (1 - p_i)\right\} \left\{\prod_{i=1}^{n} Exp\left[Ln\left(\frac{p_i}{1 - p_i}\right)^{y_i}\right]\right\} \\ &= \left\{\prod_{i=1}^{n} (1 - p_i)\right\} Exp\left[\sum_{i=1}^{n} y_i Ln\left(\frac{p_i}{1 - p_i}\right)\right] \end{aligned} \qquad (2.49)$$

Reemplazando (2.48) en (2.49), se obtiene:

$$f(y_i;p_i)=\left[\prod_{i=1}^{n}(1-p_i)\right]Exp\left[\sum_{i=1}^{n}y_i\sum_{j=0}^{k}b_j x_{ij}\right]$$
$$=\left[\prod_{i=1}^{n}(1-p_i)\right]Exp\left[\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)b_j\right] \quad (2.50)$$

El logaritmo natural de la función (2.50), llamado función soporte es:

$$l(p_i;y_i)=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)b_j+\sum_{i=1}^{n}\ln(1-p_i) \quad (2.51)$$

pero (2.47) $:1-p_i=\left[1+Exp\left(\vec{b}^T\vec{x}_i\right)\right]^{-1}$, entonces

$$\text{Ln}\,(1-p_i)=-Ln\left[1+Exp\left(\vec{b}^T\vec{x}_i\right)\right]$$

$$\text{Ln}\,(1-p_i)=-Ln\left[1+Exp\left(\sum_{j=0}^{k}b_j x_{ij}\right)\right] \quad (2,52)$$

reemplazando (2.52) en (2.51), se obtiene:

$$l(p_i;y_i)=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)b_j-\sum_{i=1}^{n}Ln\left[1+Exp\sum_{j=0}^{k}b_j x_{ij}\right] \quad (2.53)$$

como vemos (2.53) es una función que ya no depende de p_i sino de b_j solamente, entonces lo denotamos como:

$$L(\vec{b})=\sum_{j=0}^{k}\left(\sum_{i=1}^{n}y_i x_{ij}\right)b_j-\sum_{i=1}^{n}Ln\left[1+Exp\left(\sum_{j=0}^{k}b_j x_{ij}\right)\right] \quad (2.54)$$

es una función que depende exclusivamente del vector $\overset{\rho}{b}$.

Definamos como:

$$t_j = \sum_{i=1}^{n} y_i x_{ij} \qquad (2.55)$$

entonces reemplazando (2.55) en (2.54) se tiene:

$$L\left(\overset{\rho}{b}\right) = \sum_{j=0}^{k} b_j t_j - \sum_{i=1}^{n} Ln\left[1 + Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)\right] \qquad (2.56)$$

Como (2.56) es una función exclusiva del vector de parámetros $\overset{\rho}{b}$, por el Teorema de Factorización de Fisher-Neyman, Bickel y Doksum (1976), se tiene que $t_j \text{ para } j = 0,1,....,k$ son estadísticas suficientes para los parámetros b_j, para j = 0,1,...,k.,

La variable aleatoria t_j dada en la expresión (2.56) es la suma de algunos de los términos de la matriz de diseño X, es decir se incluyen en la suma solamente los elementos que corresponden a una respuesta del tipo $y = 1$.

Las ecuaciones de verosimilitud, se obtienen derivando (2.54) con respecto a los elementos de $\overset{\rho}{b}$ e igualando a cero:

$$\frac{\partial L}{\partial b_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \left[\frac{Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)}{1 + Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)}\right] \qquad (2.57)$$

las ecuaciones de verosimilitud de (2.57) son:

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \hat{p}_i = 0 \qquad j = 0,1,2....., k \qquad (2.58)$$

siendo $x_{i0} = 1$, equivalentemente (2.58) es:

$$\sum_{i=1}^{n} x_{ij}\left(y_i - \hat{p}_i\right) = 0 \qquad j = 0,1,2.....,k \qquad (2.59)$$

donde:

$$\hat{p}_i = \frac{Exp\left(\sum_{j=0}^{k} \hat{b}_j x_{ij}\right)}{1 + Exp\left(\sum_{j=0}^{k} \hat{b}_j x_{ij}\right)} \text{ ; para i=1,2,...,n}$$

es el estimador máximo verosímil de p_i y se obtiene mediante $\hat{b}_j$ y el vector $\underset{\rho}{x}_i$

La expresión (2.58) en su forma matricial es:

$$X^T\left(\underset{\rho}{y} - \underset{\rho}{\hat{p}}\right) = X\underset{\rho}{S} = \underset{\rho}{0} \qquad (2.60)$$

Estas ecuaciones son parecidas a las ecuaciones normales obtenidas para estimar el Modelo de Regresión Lineal Múltiple, pero son no lineales en $\underset{\rho}{b}$, lo cual hace que se use un método iterativo para determinar los valores del vector $\underset{\rho}{b}$.

La obtención de $\hat{b}_j$ mediante métodos iterativos; para $j = 0,1,\ldots,k$ se tratará en la siguiente sección, ahora obtendremos la varianza y covarianza de $\underset{\rho}{b}$.

Sea $X_{(nxp)}$ la matriz de diseño, con p=k+1, con elementos:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & . & x_{2k} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & . & x_{nk} \end{bmatrix}$$

Las ecuaciones de verosimilitud en su forma matricial, de la expresión (2.60):

$$X^T \hat{y} = X^T \hat{p} \quad (2.61)$$

donde $\hat{p}^T = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$

$$\sum_{i=1}^{n} x_{ij}(y_i - p_i) = 0 \quad (2.62)$$

El método de estimación de las varianzas y covarianzas lo obtendremos de la matriz de segunda derivada parcial de (2.62): y tiene la forma:

$$\frac{\partial^2 L}{\partial b_j^2} = -\sum_{i=1}^{n} x_{ij}^2 p_i (1 - p_i) \quad (2.63)$$

para j=0, 1, 2, ..., k

reemplazando: la ecuación para p_i en (2.63)

$$\frac{\partial^2 L}{\partial b^2{}_j} = -\sum_{i=1}^{n} \frac{x_{ij}^2 Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)}{\left[1 + Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)\right]^2} \quad (2.64)$$

para $j = 0,1,\ldots,k$

$$\frac{\partial^2 L}{\partial b_j \partial b_l} = -\sum_{i=1}^{n} x_{ij} x_{il} p_i (1 - p_i) \qquad (2.65)$$

$$\text{para j, l} = 0,1,2,\ldots,\text{k}$$

reemplazando:

$$\frac{\partial^2 L}{\partial b_j \partial b_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \frac{Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)}{\left[1 + Exp\left(\sum_{j=0}^{k} b_j x_{ij}\right)\right]^2} \qquad (2.66)$$

Tanto (2.64) como (2.65) no son funciones de y_i , entonces la matriz de observación y la matriz de segunda derivada esperada son idénticas.

Ahora bien la matriz que contiene el negativo de las ecuaciones (2.64) y (2.66) se denota con $\mathrm{I}\left(\underset{\sim}{b}\right)$, llamada Matriz de Información; las varianzas y covarianzas de $\hat{b}_j$ se obtienen tomando la inversa de esta matriz, esto es:

$$Cov\left(\underset{\sim}{b}\right) = \mathrm{I}^{-1}\left(\underset{\sim}{b}\right) \qquad (2.67)$$

ver Cordeiro (1992).

Los estimadores de la varianza y covarianza, denotada por $\hat{Cov}\left(\underset{\sim}{\hat{b}}\right)$, se obtiene evaluando

$Cov\left(\underset{\sim}{b}\right)$ en $\underset{\sim}{\hat{b}}$.

Entonces la matriz de información estimada, matricialmente tiene la forma:

$$\hat{\mathrm{I}}\left(\underset{\sim}{\hat{b}}\right) = X'VX \qquad (2.68)$$

V es una matriz diagonal, esto es:

$$V = Diag\left[\hat{p}_i (1 - \hat{p}_i)\right]$$

de tamaño nxn, además (2.68) es:

$$\hat{C}ov\left(\hat{\beta}\right) = \left(X'VX\right)^{-1} \qquad (2.69)$$

y es de tamaño (k+1)(k+1)

escribiremos los elementos de la matriz (2.69)

$$\hat{C}ov\left(\hat{\beta}\right) = \begin{bmatrix} \hat{s}^2\left(\hat{b}_0\right) & \hat{s}\left(\hat{b}_0, \hat{b}_1\right) \ldots\ldots \hat{s}\left(\hat{b}_0, \hat{b}_k\right) \\ . & \hat{s}\left(\hat{b}_1\right) \ldots\ldots \hat{s}\left(\hat{b}_1, \hat{b}_k\right) \\ . & . \ldots\ldots . \\ . & . \ldots\ldots . \\ . & . \ldots\ldots \hat{s}^2\left(\hat{b}_k\right) \end{bmatrix}$$

donde:

$\hat{s}^2\left(b_j\right)$ es la varianza estimada de $\hat{b}_j$

$\hat{s}\left(\hat{b}_j, \hat{b}_l\right)$ es la covarianza estimada de $\hat{b}_j$ y $\hat{b}_l$

$\hat{s}\left(\hat{b}_j\right)$ es el error estandar de $\hat{b}_j$

La matriz (2.69) será muy útil cuando se discuta el ajuste y la evaluación del Modelo de Regresión Logística.

2.6.3 MÉTODO DE NEWTON – RAPHSON PARA ESTIMAR LOS PARÁMETROS DEL MODELO DE REGRESIÓN LOGÍSTICA.

Este es un método para resolver ecuaciones no lineales, como las obtenidas en (2.57) o equivalentemente en (2.58), y requieren una solución mediante métodos iterativos para hallar la estimación de los parámetros que es el máximo de la función (2.54).

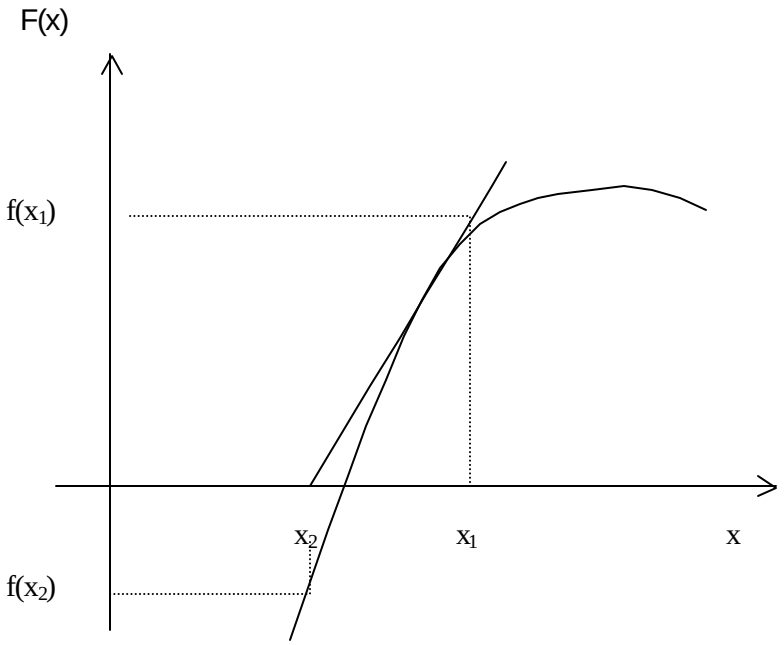
Uno de los métodos más usados para resolver ecuaciones de este tipo, es el de Newton-Raphson, porque converge rápidamente. En la figura Nº 2.2 se ilustra el método.

Tomando como estimación inicial x_1, se prolonga la tangente a la curva en este punto hasta interceptar con el eje de las abscisas al cual llamaremos punto x_2, entonces se toma a x_2 como la siguiente aproximación.

Este proceso continua hasta que un valor de x_2 haga que la función sea nula o suficientemente cercana a cero.

Para la estimación del vector $\vec{b}$ se quiere hallar el máximo de una función; entonces usa la primera derivada, porque se anula en el punto máximo y la segunda derivada para calcular las tangentes. En nuestro caso es necesario hallar la segunda derivada para obtener la matriz de varianzas y covarianzas de los parámetros estimados.

Figura Nº 2.2 Interpretación Geométrica del Método Iterativo de Newton Raphson.



Entonces se usa el siguiente esquema iterativo:

$$\hat{\beta}(t+1) = \hat{\beta}(t) + \left[I\left(\hat{\beta}(t)\right)\right]^{-1} S\left(\hat{\beta}(t)\right) \qquad (2.70)$$

donde:

$S\left(\hat{\beta}\right)$ y $I\left(\hat{\beta}\right)$ son las funciones de Score y de Información respectivamente.

La función Score es un vector de tamaño k+1, donde el ¡ésimo elemento de acuerdo a (2.57) es:

$$\frac{\partial L}{\partial b_j} = \sum_{i=1}^{n} \left(y_i - p_i^{(t)} \right) x_{ij} \quad (2.71)$$

La cual es similar a la expresión (2.59):

$$\sum_i x_{ij} \left(y_i - \hat{p}_i \right) = 0 \quad j = 1, 2, \ldots, k$$

La Función de información es una matriz de tamaño (k+1)(k+1) donde el ij ésimo elemento (l,j) es:

$$\frac{\partial^2 l}{\partial b_j \partial b_l} = -\frac{\partial}{\partial b_l} \left[\sum_{i=1}^{n} x_{ij} \left(y_i - p_i \right) \right]$$

$$= -\frac{\partial}{\partial b_l} \left[\sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \frac{e^{\underline{b}^T \underline{x}_i}}{1 + e^{\underline{b}^T \underline{x}_i}} \right]$$

$$= \sum_{i=1}^{n} x_{ij} \left[\frac{e^{\underline{b}^T \underline{x}_i} x_{il} \left(1 + e^{\underline{b}^T \underline{x}_i} \right) - e^{\underline{b}^T \underline{x}_i} x_{il} e^{\underline{b}^T \underline{x}_i}}{\left(1 + e^{\underline{b}^T \underline{x}_i} \right)^2} \right]$$

$$= \sum_{i=1}^{n} \frac{x_{ij} x_{il} e^{\underline{b}^T \underline{x}_i}}{\left(1 + e^{\underline{b}^T \underline{x}_i} \right)^2}$$

$$= \sum_{i=1}^{n} x_{ij} x_{il} p_i \left(1 - p_i \right) \quad \text{j=0,1,....,k} \quad ; \quad \text{l} \quad \text{=0,1,....,k}$$

(2.72)

donde $\hat{p}^{(t)}$, es la t-ésima aproximación para $\hat{p}$, obtenida de $\hat{b}^{(t)}$ mediante:

$$p_i^{(t)} = \frac{Exp\left(\sum_{j=0}^{k} b_j^{(t)} x_{ij}\right)}{\left[1 + Exp\left(\sum_{j=0}^{k} b_j^{(t)} x_{ij}\right)\right]} \qquad (2.73)$$

Entonces el próximo valor reemplazando en (2.70) es:

$$\hat{b}^{(t+1)} = \hat{b}^{(t)} - \left\{X^T V^{(t)} X\right\}^{-1} X^T \left(y - \hat{p}^{(t)}\right) \qquad (2.74)$$

donde $V^{(t)} = Diag\left[p_i^{(t)}\left(1 - p_i^{(t)}\right)\right]$

La expresión (2.70) se usa para obtener $\hat{p}^{(t+1)}$ y así sucesivamente. Después de dar un valor inicial $\hat{b}^{(0)}$, se usa (2.70) para obtener $\hat{p}^{(0)}$ y para t>0 las iteraciones siguientes se efectúan usando (2.70) y (2.71).

En el límite, $\hat{p}^{(t)}$ y $\hat{b}^{(t)}$ converge a los EMV de $\hat{p}$ y converge en general en 5 o 6 iteraciones.

Existen software estadísticos como el SAS y el SPSS con programas para estimar una regresión logística usando el método descrito. Una ventaja de este método es que en el paso final del proceso iterativo se obtiene la inversa de la función de información, que es asintóticamente la matriz de varianzas y covarianzas del vector $\hat{b}$ y permiten efectuar inferencias sobre los parámetros basado en la teoría normal. Para mayor información se recomienda a Affifi y Clark (1998).

2.6.4 INTERPRETACIÓN DE LOS COEFICIENTES DEL MODELO ESTIMADO

Recordamos del modelo de regresión múltiple que el valor de un coeficiente significaba el cambio en unidades de la variable dependiente por cada unidad de la variable independiente a que se refiere el coeficiente, permaneciendo invariantes los valores del resto de variables independientes del modelo.

A nivel de coeficientes estimados exponencialmente la interpretación es muy similar y la diferencia estriba en que en este caso no se trata del cambio (incremento o disminución) de la probabilidad de la variable dependiente por cada unidad de cambio en las independientes, sino del incremento o disminución que se produce en el cociente entre P(Y=1) / P(Y=0), expresado por:

$$\frac{P(Y=1)}{P(Y=0)} = e^{B_0 + B_1 X_1 + BX_2 + \ldots + B_K X_K} \qquad (2.75)$$

Más aún, están expresados en logaritmos, por lo que sería necesario transformarlos (tomando los valores del antilogaritmo) de tal forma que se evalúe más fácilmente su efecto sobre la probabilidad. Los programas de computador lo hacen automáticamente calculando tanto el coeficiente real como el transformado. Utilizar este procedimiento no cambia en modo alguno la forma de interpretar el signo del coeficiente. Un coeficiente positivo aumenta la probabilidad, mientras que un valor negativo disminuye la probabilidad. Así pues si β es positivo, su transformación (antilog) será mayor a 1, y el odds ratio aumentará. Este aumento se produce cuando la probabilidad prevista de ocurrencia de un suceso aumenta y la probabilidad prevista de su no ocurrencia disminuye. Por lo tanto, el modelo tiene una elevada probabilidad de ocurrencia. De la misma forma, si β es negativo, el antilogaritmo es menor que 1 y el odds ratio disminuye. Un valor de cero equivale a un valor de 1, lo que no produce cambio en el odds. Hair (1999)

2.6.4 PRUEBA DE HIPÓTESIS PARA LOS COEFICIENTES DEL MODELO DE REGRESIÓN LOGÍSTICA.

Usualmente en la estimación del Modelo de Regresión Logística, como en el Modelo de Regresión Lineal Múltiple se efectúan pruebas con objetivos diferentes, siendo estos:

1. Determinar si una variable explicativa tiene coeficiente igual a cero.
2. Determinar si un conjunto de variables explicativas tienen coeficientes igual a cero.
3. Determinar la calidad del ajuste global del modelo.

Veamos para cada objetivo, como se efectúa el análisis.

2.6.5.1 PRUEBA DE WALD

Wald(1943) estudio una prueba asintótica para estimaciones máximos verosímiles, y aseveró que los parámetros estimados en los modelos logísticos tiene una Distribución Normal para muestras grandes.

Esta prueba se usa para evaluar la significancia estadística de cada variable explicativa o regresora.

Sea $\hat{p}^{(t)}$ que converge a los EMV de $\hat{p}$ y. $y_1, y_2, \ldots\ldots\ldots, y_n$ variables respuesta binaria independientes cuyas probabilidades satisfacen.

$$Logit(p_i) = \hat{x}_i^T \hat{b}$$

donde $p_i = P[y_i = 1/\hat{x}_i]$

Siendo $\hat{x}_i$ una observación que contiene los valores de las k variables explicativas con $\hat{x}_i^T = (1, x_{i1}, x_{i2}, \ldots\ldots x_{ik})$.

Sin perdida de generalidad, seleccionaremos b_j como el parámetro de interés.

Supóngase que las hipótesis son:

$$H_0 : b_j = b_{j0}$$
$$H_1 : b_j \neq b_{j0} \qquad (2.76)$$

sea $\hat{b}_j$ un EMV de b_j y sea:

$I^{-1} = \left(X^T V X\right)^{-1}$ la inversa de la matriz de información muestral, entonces la estadística de Wald para dócimar (2.75) es:

$$W = \frac{\left(\hat{b}_j - b_{j0}\right)^2}{\hat{s}^2\left(\hat{b}_j\right)} \qquad (2.77)$$

donde $\hat{s}\left(\hat{b}_j\right)$ es la estimación del error estándar de $\hat{b}_j$.

Bajo H_0, $W \sim c^2_{(1)}$ y para n suficientemente grande se tiene que:

$$z = \frac{\hat{b}_j - b_{j0}}{\hat{s}\left(\hat{b}_j\right)} \sim N\left(\left(\frac{\hat{b}_j - b_{j0}}{\hat{s}\left(\hat{b}_j\right)}\right), 1\right) \qquad (2.78)$$

por tanto:

$z^2 \sim c^2_{(x,1)}$, es c^2 con parámetro de no centralización:

$$x = \frac{\left(b_j - b_{j0}\right)^2}{\hat{s}\left(\hat{b}_j\right)} \qquad (2.79)$$

ver Hanck y Donner (1977)

Pero la estadística W, tiene la propiedad que cuando el valor absoluto del coeficiente de Regresión es grande, el error estándar también lo es; esta situación hace que la estadística W sea pequeña y por tanto se puede rechazar b_j igual a cero, cuando en realidad no debería rechazarse.

Por tanto, cuando se encuentra que un coeficiente es grande, es preferible no usar la estadística de Wald para efectuar dócima individual. Sino se recomienda construir un modelo con y sin esa variable y basarse en la prueba de hipótesis de la diferencia entre los dos modelos, ver Hanck y Donner (1977).

Para las hipótesis estadísticas:

$$H_0 : b_j = 0$$
$$H_1 : b_j \neq 0 \qquad (2.80)$$

La estadística (2.77) es:

$$W = \frac{\left(\hat{b}_j\right)^2}{\hat{s}^2\left(\hat{b}_j\right)} \qquad (2.81)$$

Bajo H_0 , $W \sim c^2_{(1)}$ y para n suficientemente grande se tiene que:

$$z = \frac{\hat{b}_j}{\hat{s}\left(\hat{b}_j\right)} \sim N\left(\frac{\hat{b}_j}{\hat{s}\left(\hat{b}_j\right)}, 1\right) \qquad (2.82)$$

por tanto:

$$z^2 \sim c^2_{(x,1)}$$

z^2 se distribuye como una $c^2_{(x,1)}$ con parámetro de no centralización:

$$x = \frac{\left(b_j\right)^2}{\hat{s}^2\left(\hat{b}_j\right)} \qquad (2.83)$$

si la variable explicativa es categórica, los grados de libertad es igual al número de categorías o niveles de la variable menos uno.

2.6.5.2 PRUEBA CHI-CUADRADO

Esta prueba sirve para lograr el segundo objetivo propuesto al iniciarse la sección 2.6.4 y sirve para docimar los coeficientes del modelo logístico.

Para elegir un modelo, se usa la prueba de razón de verosimilitud, Bickel y Docksum (1977), para probar la hipótesis de que los coeficientes b_j correspondientes a las variables explicativas retiradas, digamos q variables explicativas, del modelo son iguales a cero, siendo la hipótesis estadísticas:

$$H_0 : b_1 = b_2 = ... = b_q = 0$$

$$H_1 : b_j \neq 0, para\ \ por\ \ lo\ \ menos\ \ un\ \ j = 1,2,..., q$$

.Esta prueba se basa en la siguiente estadística:

$$c_q^2 = -2\left[LnL_{p-q} - LnL_p\right] \quad (2.84)$$

Bajo la hipótesis de que los coeficientes de las variables retiradas son iguales a cero, la estadística (2.84) tiene una distribución asintótica $c_{(q)}^2$.

Valores altos para esta estadística indican que una o más de las *q* variables retiradas tienen coeficiente de regresión distinto de cero.

La estadística c_q^2 se usa también para probar si una variable explicativa determinada, por ejemplo x_k, muestra una asociación significativa (como factor de riesgo cuando se aplica a casos de enfermedades) para con la variable respuesta en la presencia de las demás variables $x_1, x_2, ..., x_{k-1}$.

2.6.5.3 ESTADISTICA CHI-CUADRADA DE PEARSON

Esta estadística sirve para lograr el objetivo número 3, es decir evaluar el modelo ajustado en forma global. La estadística se basa en la comparación de los valores observados, y_i ; y sus respectivas probabilidades estimadas, p_i.

Las hipótesis estadísticas para usar esta estadística son:

$$H_0 : b_o = b_1 = ... = b_k = 0$$

$$H_1 : b_j \neq 0, para\ por\ lo\ menos\ un\ j = 0,1,.,2,...,k$$

esta prueba se basa en la estadística Chi-cuadrado de Pearson, que está dada por:

$$c^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)} \quad (2.85)$$

o equivalentemente $c^2 = \sum_{i=1}^{n} \frac{r_i^2}{v_{ii}}$ (2.86)

donde:

$$r_i = (y_i - \hat{p}_i)$$

$$v_{ii} = Diag(\hat{V}) = \hat{p}_i(1-\hat{p}_i)$$

como observamos la estadística (2.86) es igual a (1.52).

Bajo la hipótesis nula, de que el modelo se ajusta bien a los valores observados, la

estadística (2.86) tiene distribución asintótica Chi-cuadrado $c^2_{(n-(k+1))}$.

Valores altos de la estadística Chi-cuadrado de Pearson indican discrepancias con el modelo teórico, Cordeiro (1992).

La estadística (2.86) es inestable cuando $\hat{p}_i$ toma valores cercanos a cero o uno, por ello tomar en cuenta esta observación, cuando se realiza el análisis.

2.6.5.4 DESVIANZA

Otra forma de probar el ajuste global del modelo, es mediante la estadística llamada Desvianza, propuesta por Nelder y Wederburn (1982), es análogo a la suma de cuadrados de los residuales del Modelo de Regresión Lineal Múltiple.

Las hipótesis estadísticas son:

$$H_0 : b_1 = ... = b_k = 0$$
$$H_1 : b_j \neq 0, para\ \ por\ \ lo\ \ menos\ \ un\ \ j = 1,.,2,...,k$$

Esta estadística se usa para evitar la inestabilidad de la estadística Chi-cuadrado de Pearson. La Desvianza esta dada por:

$$D_p = \sum_{i=1}^{n} d_i^2 \qquad (2.87)$$

donde :

$$d_i = \begin{cases} \sqrt{-2\log \hat{p}_i} & si\ y_i = 1 \\ & ; j = 1,2,...,n \\ \sqrt{-2\log(1-\hat{p}_i)} & si\ y_i = 0 \end{cases}$$

La Desvianza bajo la hipótesis nula, asintóticamente, es la misma que la distribución Chi-cuadrado de Pearson, es decir se distribuye $c^2_{(n-(k+1))}$ y mide la discrepancia o el desvio entre el modelo bajo investigación o actual y el modelo saturado.

La estadística (2.87) para el modelo de regresión logística eta dada por:

$$D = -2\sum \left(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i) \right) \qquad (2.88)$$

Cuando el modelo bajo investigación es verdadero se compara el valor D con el valor crítico $c^2_{(n-p)}$ de una distribución c^2 a un nivel de significación igual a a , por tanto :

Si $D > c^2_{(n-p)}$ el modelo se rechaza y

Si $D \leq c^{2}_{(n-p)}$ el modelo no se rechaza.

donde $p = k + 1$

2.6.5.5 TABLA DE CLASIFICACION

También llamada Matriz de Confusión, es una forma sencilla de evaluar el ajuste del Modelo de Regresión Logística, no es tan objetiva pero se usa como indicador de bondad de ajuste.

Es una tabla sencilla de 2x2, en el cual se muestra la distribución de los objetos que pertenecen a las categorías 1 y 2, es decir cuando $y = 0$ y cuando $y = 1$, conjuntamente con la clasificación a cualquiera de las 2 categorías de acuerdo a la probabilidad estimada.

Para interpretar se hace mediante el porcentaje de objetos bien clasificados, esto es, aquellos que mediante la probabilidad estimada permanecen en su respectiva categoría. También se interpreta mediante el porcentaje de objetos mal clasificados, esto es, aquellos que mediante la probabilidad estimada se asignan a categorías diferentes del cual fueron observados.

TABLA DE CLASIFICACION

GRUPO ACTUAL	**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 MAGINAL	$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

$\frac{n_{11}+n_{22}}{n} x100\%$ es el porcentaje de objetos bien clasificados mediante el Modelo de Regresión Logística estimado.

Por tanto, lo que se debe esperar es que este porcentaje sea lo más alto posible, a fin de concluir que el modelo obtenido clasifica bien a los objetos o individuos.

2.6.5.6 CONTRASTE DE BONDAD DE AJUSTE DE HOSMER – LEMESHOW

Este contraste evalúa la bondad de ajuste del modelo, es decir el grado en que la probabilidad predicha coincide con la observada, construyendo una tabla de contingencia a la que aplica un contraste $\div^2$. Para ello calcula los deciles de las probabilidades estimadas ($\hat{p}_i$; $i = 1,2,...n$) , $D_1, D_2, ...D_9$ y divide los datos observados en 10 categorías dadas por :

$$A_j = \{\ \hat{p}_i \in [D_{j-1}, D_j)\ /\ i \in \{1,2,...,n\}\ \} \quad ; \quad j = 1,2,..,10$$

donde $D_0 = 0$, $D_{10} = 1$.

Sean:

n_J = número de casos en Aj ; j=1,2,...,10

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

$$\overline{p}_j = \frac{1}{n_j} \sum_{i \in A_j} \hat{p}_i \quad ; \quad j = 1,2,...10$$

El estadístico del contraste viene dado por :

$$T = \sum_{j=1}^{10} \frac{(o_j - n_j \overline{p}_j)^2}{n_j \overline{p}_j (1 - \overline{p}_j)} \qquad (2.89)$$

y el p-valor del contraste es $P\left[c_8^2 \geq T_{obs}\right]$.z

2.6.6 DIAGNOSTICO DEL MODELO

Es la evaluación de la bondad de ajuste caso por caso mediante el análisis de los residuos del modelo y de su influencia en la estimación del vector de parámetros del mismo, se realiza usando:

2.6.6.1 RESIDUOS DEL MODELO

Los residuos más utilizados son los siguientes:

Residuos estandarizados.- Son el cociente entre los residuales y una estimación de la desviación estándar.

$$z_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \quad ; \quad i = 1,2,...,n \qquad (2.90)$$

Residuos studentizados.- Son el cambio en el valor de la desvianza del modelo si el caso es excluído.

$$st_i = \frac{y_i - \hat{p}_{(i)}}{\sqrt{\hat{p}_{(i)}(1-\hat{p}_{(i)})}} \quad ; \quad i = 1,2,...,n \qquad (2.91)$$

donde $\hat{p}_{(i)}$ es la estimación de p_i obtenida eliminando la observación i de la muestra.

Residuos Desvianza.- Para cada observación la desvianza se calcula :

$$d_i = \begin{cases} \sqrt{-2\log \hat{p}_i} & si\ y_i = 1 \\ \\ \sqrt{-2\log(1-\hat{p}_i)} & si\ y_i = 0 \end{cases} \quad ; j = 1,2,...,n \qquad (2.92)$$

Todos estos residuos se distribuyen aproximadamente como una N(0,1), si el modelo ajustado es correcto.

2.6.6.2 MEDIDAS DE INFLUENCIA

Cuantifican la influencia que cada observación ejerce sobre la estimación del vector de parámetros o sobre las predicciones hechas a partir del mismo, de modo que, cuanto más grande son, mayor es la influencia que ejerce una observación en la estimación del modelo.

Medida de Apalancamiento (Leverage)

Se utiliza para detectar observaciones que tienen un gran impacto en los valores predichos por el modelo.

Se calcula a partir de la matriz $H = W^{1/2} X (X'WX)^{-1} X'W^{1/2}$ donde $W = diag|\hat{p}_i(1-\hat{p}_i)|$. El apalancamiento para la observación i-ésima viene dado por el elemento i-ésimo de la diagonal principal de H, h_{ii} , y toma valores entre 0 y 1 con un valor medio de p/n.

Las dos medidas siguientes miden el impacto que tiene una observación en la estimación de $\hat{a}$.

Distancia de Cook.- Mide la influencia en la estimación de $\hat{a}$.

$$COOK_i = \frac{1}{p}\left(\hat{\hat{a}} - \hat{\hat{a}}_{(i)}\right)' X'WX \left(\hat{\hat{a}} - \hat{\hat{a}}_{(i)}\right) \quad (2.93)$$

DFBETA.- Mide la influencia en la estimación de una componente de $\hat{a}$, $\hat{a}_i$

$$Dfbeta1_i = \frac{\hat{\hat{a}}_1 - \hat{\hat{a}}_{1(i)}}{std(\hat{\hat{a}}_1)} \quad (2.94)$$

donde $\hat{\hat{a}}_1, \hat{\hat{a}}_{1(i)}$ denotan las estimaciones del módelo logístico de $\hat{a}$ y $\hat{a}_{1,}$ eliminando la i-ésima observación de la muestra y $std(\hat{\hat{a}}_1)$ el error estándar en la estimación de $\hat{a}_1$.

2.6.7 VARIABLES EXPLICATIVAS CATEGORICAS – VARIABLES DUMMY

Si una de las variables explicativas es categórica, con c valores posibles, se crean c-1 variables dicotómicas como variables explicativas también

llamadas variables dummy. Estas variables cuantifican el efecto de un valor de dichas variables con respecto a un valor de referencia.

Estas variables se usan cuando los datos se muestran como categorías, las categorías pueden ser:

Nominales: La variable simplemente indica diferentes categorias, las categorías no pueden ser ordenadas en un orden particular. Ejemplo : Sexo (hombre,mujer).

Ordinales: La variable además de estar agrupada en categorías puede ser ordenada. El que una categoría este en un orden superior que otra implica que su medida representa algo mayor que la otra. Ejemplo: Clase social (baja, media, alta).

Intervalares: La variables no solo puede ser ordenada, sino que su valor mide la distancia entre categorías. Estas tienen estándares de unidades de medida.

Ejemplo: Altura, temperatura, presión sanguínea.

Cuando se tiene variables de este tipo se crean las llamadas variables dummy, si la variable tiene c categorias se usan c-1 variables ficticias o dummy. La variable indica si un dato corresponde a una categoría o no. Veamos un ejemplo de cómo se hace esto:

Supongamos que tenemos una variable clase social, codificada 1: Baja, 2:Media, y 3:Alta, entonces creamos dos variables dummy :

Clase1 : 1 si el dato corresponde a la clase social Baja, 0 si el dato no pertenece a la clase Baja.

Clase2 : 1 si el dato corresponde a la clase social Media, 0 si el dato no pertenece a la clase Media.

Como se puede ver estas nos permiten clasificar cualquier dato en una de las categorias existentes.

Supongamos que creamos las variables dummy Ind1 e Ind2 para una variable X1 de tres categorías, como sigue:

X1	Ind1	Ind2

1 2	1 0	0 1
3	0	0

En este caso la ecuación del modelo ajustado viene dada por:

$$\log\left(\frac{p(Ind1, INd2; b)}{1-p(Ind1, Ind2; b)}\right) = b_0 + b_1 Ind1 + b_2 Ind2 \qquad (2.95)$$

Sea p_i = p[y=1/X1=i]; i=1,2,3. Se tiene:

$$\frac{p_1}{1-p_1} = e^{b_0+b_1} \quad , \frac{p_2}{1-p_2} = e^{b_0+b_2} \quad , \frac{p_3}{1-p_3} = e^{b_0}$$

Se sigue que:

$$\frac{p_1}{1-p_1} \Big/ \frac{p_3}{1-p_3} = e^{b_1} \quad , \frac{p_2}{1-p_2} \Big/ \frac{p_3}{1-p_3} = e^{b_2}$$

Por lo tanto, $e^{b_i}, i = 1,2$ compara los odds ratio correspondientes a X1=1,2, frente al de la categoría de referencia X1=3.