ESCUELA SUPERIOR POLITECNICA DEL LITORAL

Instituto de Ciencias Matemáticas

"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

INFORME DE MATERIA DE GRADUACIÓN

Previo a la obtención del título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

Andrea Fuentes

Nathaly Rivera

Raúl Pinos

Guayaquil – Ecuador

2012

AGRADECIMIENTO

A Dios, por todas las bendiciones y oportunidades otorgadas;

A nuestras familias por su invalorable apoyo y respaldo en todo momento;

A nuestro director de Materia de Graduación M.Sc. Gaudencio Zurita por la paciencia, dedicación y apoyo brindado en la culminación de este trabajo.

DEDICATORIA

Dedicamos este trabajo a todos aquellos que creyeron en esta idea y que con su aporte directo o indirecto lograron que se plasme en realidad. Valoramos y respetamos mucho la ayuda y comprensión de todos quienes nos regalaron un poco de su tiempo, atención y dedicación.

Muchas gracias.

`

TRIBUNAL DE GRADUACIÓN

_____	_____
M.Sc. Gaudencio Zurita	M.Sc. Jorge Medina
Profesor de la Materia	**Delegado ICM**
de Graduación	

DECLARACIÓN EXPRESIVA

"La responsabilidad del contenido de esta Trabajo final de graduación de Grado, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral".

(Reglamento de Graduación de la ESPOL)

_____ _____

Raúl Alejandro Pinos Loaiza Nathaly Rivera Flores

_____

Andrea Elizabeth Fuentes Puglla

RESUMEN

Este presente trabajo se desarrolló para diseñar e implementar un software libre estadístico llamado ERLA para apoyo académico a docentes y estudiantes de la carrera de Estadística Informática del Instituto de Ciencia Matemáticas, el software fue implementado con las plataformas como son Matlab y una interfaz gráfica en .Net.

Este Software trabaja con funciones propias de MATLAB y otras funciones personalizadas para propósitos estadísticos y de ingeniería.

El software es un software especializado en la técnica de Regresión Lineal, es posible evaluar la calidad de los modelos obtenidos, realizar estimaciones de todos los modelos que se hayan generado y además seleccionar el mejor modelo considerando todas las variables que usted considere sean relevantes en el estudio.

En el primer y segundo capítulo se presentan las técnicas de Regresión Lineal Simple y Múltiple, los cuales presentan los métodos de obtener los estimadores de los parámetros como es el de Mínimos Cuadrados. Además la construcción de la Tabla de Análisis de Varianzas.

En el tercer capítulo, se presentan las familias exponenciales que permiten descomponer distribuciones exponenciales, las cuales permiten crear una función de enlace donde nace el Modelo Lineal Generalizado, luego aplicar

los métodos aplicados para estimar los parámetros y también como es el método de Newton-Raphson.

En el capítulo cuatro, se presentan las técnicas estadísticas de Regresión Logística y Poisson que son modelos no lineales, las cuales utilizan Modelos Lineales Generalizados, además contiene las distribuciones con la que se trabajan, la interpretación de los parámetros, las estimaciones de parámetros de cada uno de los modelos, la evaluación de cada uno de los modelos ya sea de la Regresión Logística y Poisson y una breve ilustración de ambas técnicas en el software ERLA.

Para finalizar, en el capítulo cinco se presenta los algoritmos creados específicamente para los módulos de Regresión Logística y Poisson y la validación de los Modelos ya mencionados, estableciendo los valores de los parámetros betas y añadiendo una variable que será $\mathrm{N}\sim(0,\sigma^2)$.

Contenido

Índice de Gráficos

Índice de Tablas

Índice de Ilustraciones y cuadros

INTRODUCCION

Previo a la obtención del título de Ingeniero en Estadística Informática, con la Materia de Graduación "Regresión Lineal Avanzada", se ha desarrollado un paquete estadístico especializado en el Análisis de la Regresión, considerando que es una de las técnicas estadísticas de mayor uso, utilización que se debe a su sencillez y amplia aplicabilidad; además lo que permite es explicar y estudiar la relación entre una o más variables de respuesta en término de un grupo de variables predictoras o de "explicación".

El desarrollo del software de Análisis de Regresión Avanzada denominado ERLA, está compuesta con diversos Módulos Específicos como son: "Regresión Ridge y Regresión Robusta", "Regresión Logística y Regresión Poisson", "Calidad de Modelos" y "Análisis de varianza de un solo factor y dos factores". Que se realizó mediante una interconexión entre el software matemático MATLAB 2010 que es un producto de The MathWork y Visual Basic.NET 2008 que es producido por Microsoft.

Lo concerniente a programación que se encarga de tomar datos ingresados por el usuario, analizarlos, aplicar algoritmos, y proporcionar información, está programado en Matlab, que es un lenguaje de programación amigable y que además permite implementar fácilmente los algoritmos simples o complejos, también está el hecho de poder importar y exportar datos e información a otros programas; fueron entre otras, las características que nos hizo decidir utilizáramos este programa como base del proyecto.

Lo que Matlab no hace es crear una interfaz gráfica amigable y sencilla que los usuarios puedan entender.

Por esta razón recurrimos a otro programa, creado por Microsoft, este es Visual Basic .NET 2008, cuyo principal características es poder relacionar todos los objetos que se incluyen en su interfaz gráfica, con comandos de programación; con este programa pudimos incluir las opciones "Abrir", "Guardar", "Importar datos", "Calculadora", "Realizar Gráficos", pero sobre todo, hacer posible incluir las librerías creadas con Matlab para poder desarrollar las operaciones de Regresión que se necesite, sin dejar de lado la simplicidad al momento de hacer las operaciones pertinentes. Entre las muchas ventajas que brindan estos programas por separados, al hacerlos trabajar en conjuntos en este Software estadístico, hemos logrado crear una forma de hacer conocer al usuario, que la Regresión no es un área difícil ni complicada de la Estadística, ya que cada paso está hecho para que el mas lego de los usuario logre comprender de inmediato los pasos requeridos para poder hacer uso de ERLA a su completa capacidad.

Este Reporte Técnico proporciona los fundamentos teóricos sobre el cual se desarrolló el módulo "Regresión Logística y Regresión Poisson". Partimos desde lo básico, desde qué es Regresión Lineal Simple, de qué trata, qué lo conforma, cómo se utiliza, cómo calculamos los estimadores de los parámetros, las hipótesis y supuestos detrás de todo, la muy útil tabla ANOVA, y lo que decide todo una vez tomada la muestra, el valor p; tratamos de ser lo más exhaustivos posible, todo para no dejar dudas, y avanzamos poco a poco, primero regresión Simple, luego modelos Polinómicos, cuándo la regresión lineal simple no es suficiente, a modelos de Regresión Lineal Múltiple cuándo hay más de una variable de explicación, como afecta esto a los modelos originales, el uso de matrices para una mejor presentación de la información a utilizarse, los nuevas hipótesis y supuestos, las modificaciones a la ANOVA y el valor p; todo esto para entender qué es Regresión.

La parte central de este trabajo es Regresión Logística y Regresión Poisson, qué es lo que las hace especiales y diferentes a la Regresión Lineal; comenzamos por lo básico, no podemos comenzar sin mencionar a la Familia Exponencial y los Modelos Lineales Generalizados, que cambian por completo el concepto de Regresión, pero no su base; que es el hecho de explicar una variable en base de otra u otras, pero al no haber una relación lineal directa, recurrimos a la Familia de Distribuciones Exponenciales, que nos permiten en gran medida resolver los problemas de regresión cuando la variable a ser explicada no tiene una distribución lineal, y por ende no cumple con los supuestos de homocedasticidad y demás, pero gracias a ellos logramos crear una forma de adaptar los modelos de regresión por medio de un enlace, pero estas nuevas funciones necesitaran de un nuevo aliado, un método numérico que se ha escogido sea el de Newton-Raphson, que permite calcular los estimadores de betas de los modelos de regresión Logística y Poisson, ya que las soluciones están expresada de manera implícita.

En este trabajo está mucho de nuestro esfuerzo y esperamos sea de utilidad para todo aquel que necesite y quiera aprender más sobre modelos de Regresión Logística y Regresión Poisson.

CAPÍTULO I

1. REGRESIÓN LINEAL

1.1. Introducción

Comúnmente en el mundo matemático, podemos relacionar dos variables entre sí, por una simple regla de correspondencia; suponiendo que Y es una variable que se explica determinísticamente por medio de X, bajo la relación Y= 2X + 3, simplemente calcularíamos el valor de Y *dado que* X = 3, tendríamos: 2(3) + 3 = 9. Todo esto dentro del mundo de los modelos matemáticos determinísticos, pero en el mundo real, las cosas no son tan sencillas.

Cuando no se conoce la relación funcional que liga a Y con X, pero podemos fijar *n* valores de X, y luego *leer*los n valores que corresponden en Y; una vez observado estos últimos valores, podremos organizarlos pareadamente, y representarlos como *n* pares a saber:

$$\binom{x_1}{y_1}; \binom{x_2}{y_2}; \dots; \binom{x_n}{y_n}$$

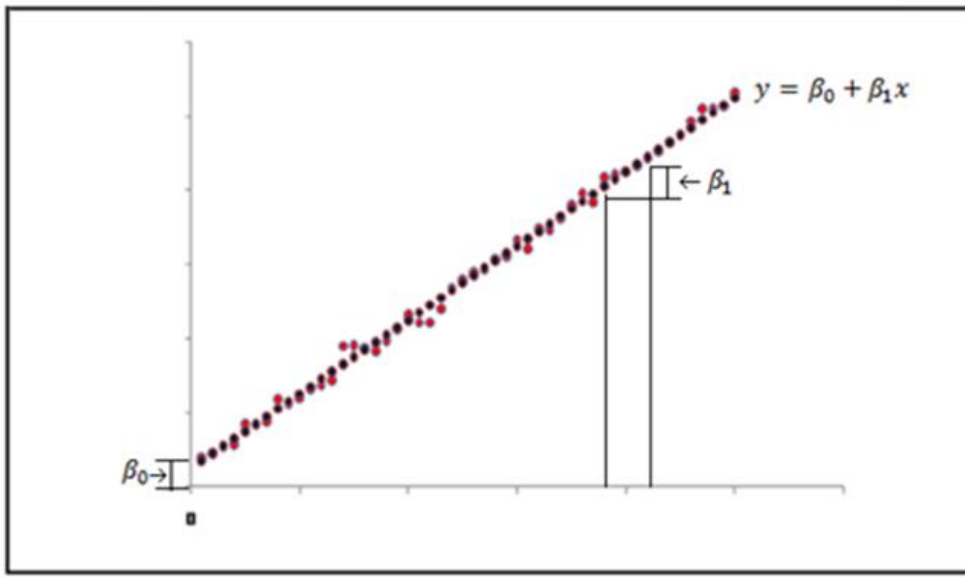
Con este tipo de datos se inicia la búsqueda de una relación funcional condicional que denominaremos *g*, que explique Y en términos de una variable X, que es en sí de lo que trata la técnica estadística denominada

Regresión. También se puede explicar Y con dos o más variables, lo cual veremos en el Capítulo II.

SI con estos datos pareados, se construye un gráfico de dispersión X vs. Y, y obtuviésemos algo semejante a una línea recta (Gráfico 1.01), sería plausible suponer que existe una relación condicional entre Y y X, y que viene dada por la ecuación:

$$y = g(x) = \beta_0 + \beta_1 x_i \qquad (1.01)$$

Gráfico 1.01: Dispersión X vs Y
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Esta ecuación es la de una recta, donde β_1 es su pendiente y β_0 es el valor que toma Y cuando la recta hace intersección con el eje vertical Y. Hasta este punto, pareciera que solo es cuestión de Calcular β_0 y β_1 en base de

los datos pareados $\binom{x_1}{y_1}; \binom{x_2}{y_2}; \ldots; \binom{x_n}{y_n}$, y en parte así es, pero también se toma en cuenta algunos aspectos propios de cada problema.

En este Capítulo además presentamos los Valores Esperados a partir del Modelo de Regresión Lineal con el Teorema de Gauss-Markov, así como la estimación de parámetros por el Criterio de Mínimos Cuadrados y Máxima Verosimilitud, como también la construcción de la denominada Tabla de Análisis de Varianza.

1.2 Regresión Lineal Simple

En Regresión Lineal Simple, tratamos de explicar Y en función de X con la asistencia de la ecuación de una recta con β_1 como la pendiente y β_0 como la intersección con el eje Y, pero una vez hecho el cálculo determinístico de Y, y tomado la lectura experimental de Y, se encuentra que no siempre coinciden, ya que hay la presencia de un error aleatorio ε_i, que nos hace reescribir la relación de Y con X:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1.02)$$

Dado este modelo donde Y es la variable a ser explicada condicionalmente por X, a quien llamaremos variable de explicación y ε_i una variable aleatoria que influencia en la observación del valor y_i de Y cuando X=x_i; vamos a trabajar con el siguiente modelo condicional y bajo los siguientes supuestos:

$$E(Y|X_i = x_i) = \beta_0 + \beta_1 x_i \quad i=1, 2, \ldots,n$$

$$E(\varepsilon_i) = 0\,; \qquad Var(\varepsilon_i) = \sigma^2\,; \qquad Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j \qquad (1.03)$$

Es un modelo de Regresión Lineal simple porque se explica la variable de respuesta Y en función de solo una variable X y los valores de β_0 y β_1 son lineales en la expresión $E(Y|X = X_i) = \beta_0 + \beta_1 x_i$ que también es denominada Función de Respuesta o Parte Determinística del modelo; los valores de σ^2, β_0, β_1 son constantes desconocidas pero estadísticamente estimables; ε_i es una variable aleatoria como fuera enunciado previamente.

1.2.1 Valores Esperados a partir del modelo de Regresión Lineal Simple (Teorema Gauss – Markov)

Como se estableció anteriormente la Relación Estadística que explica condicionalmente a Y en términos de X es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (1.04)$$

Haciendo ε_i, que el valor observado $y_i$de Y sea una Variable Aleatoria, de donde:

$$\mu_{Y_i} = E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = E[Y|X = x_i]$$

$$\mu_{Y_i} = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$$

$$\mu_{Y_i} = E(Y_i) = \beta_0 + \beta_1 x_i \qquad (1.05)$$

Como $E(\varepsilon_i) = 0$,y el Valor Esperado de una constante es la misma constante.

$$\mathrm{Var}(\mathrm{Y_i}) = \sigma^2_{\mathrm{Y_i}} = \mathrm{Var}(\beta_0 + \beta_1 \mathrm{x_i} + \varepsilon_\mathrm{i}) = \mathrm{E}[\mathrm{Y_i} - \mathrm{E}(\mathrm{Y_i})]^2$$

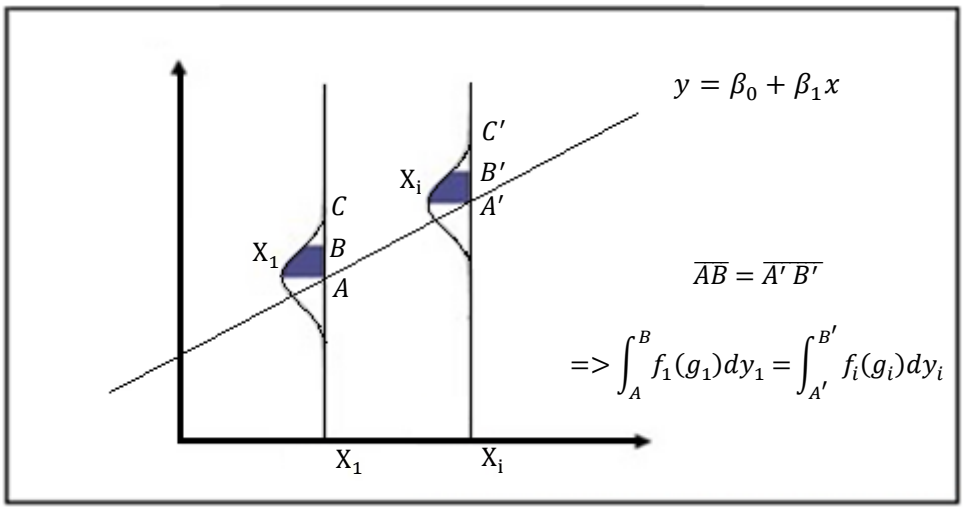
$$\mathrm{Var}(\mathrm{Y_i}) = \mathrm{E}[(\beta_0 + \beta_1 \mathrm{x_i} + \varepsilon_\mathrm{i}) - (\beta_0 + \beta_1 \mathrm{x_i})]^2 = \mathrm{E}[\varepsilon_\mathrm{i} - 0]^2 = \mathrm{E}[\varepsilon_\mathrm{i} - \mathrm{E}(\varepsilon_\mathrm{i})]^2$$

$$= \mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2 \qquad (1.06)$$

Si suponemos que el Error ε_i se distribuye normalmente, tenemos entonces que $\varepsilon_\mathrm{i} \sim \mathrm{N}(0, \sigma^2)$, siendo σ^2 constante, supuesto de homocedasticidad, lo que implica que $\mathrm{Y_i} \sim \mathrm{N}(\beta_0 + \beta_1 \mathrm{x_i}, \sigma^2)$.

Este resultado es conocido como Teorema de Gauss-Markov, y que se ilustra en el Gráfico 1.02 donde se representa el gráfico de la distribución de Y dado que $X = x_1$.

Gráfico 1.02: Teorema Gauss-Markov
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

1.2.2 ESTIMACIÓN DE LOS PARÁMETROS

Los parámetros β_0, β_1, y σ^2 del modelo de Regresión Lineal Simple pueden ser estimados a través de diferentes criterios tales como, Mínimos Cuadrados o Máxima Verosimilitud. Vamos a estimar β_0 y β_1 en el modelo, en base a la información pareada que nos dan los datos observados, condicionando a que se cumplan los supuestos relacionados con el modelo.

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$$

$$E(\varepsilon_i) = 0; \quad VarE(\varepsilon_i) = \sigma^2; \quad cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$$

Estimación por Mínimos Cuadrados para Regresión Lineal Simple.

La estimación por Mínimos Cuadrados para Regresión Lineal Simple es una técnica de análisis numérico introducida dentro de la optimización matemática, en la que, dado un conjunto de pares $\binom{x_1}{y_1}; \binom{x_2}{y_2}; \ldots; \binom{x_n}{y_n}$, se pretende encontrar la función que mejor se aproxime a los datos, de acuerdo con el criterio de minimizar el error cuadrático, es decir intenta minimizar las suma de cuadrados de las diferencias de los errores $(Y_i - E(Y_i)) = \epsilon_i$ o entre los puntos generados por la función $Q = \sum_{i=1}^{n} {\epsilon_i}^2$. Un requisito implícito es que los errores de cada medida estén distribuidos de forma aleatoria.

Q se define como la suma cuadrática de los errores:

$$Y_i - (EY_i) = Y_i - (\beta_0 + \beta_1 x_i)$$

$$Q = \sum_{i=1}^{n} [\varepsilon_i^2] = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 = Suma\ Cuadrática\ del\ Error\ o\ Residuos$$

Tomando:

$$Q = \sum_{i=1}^{n} [\varepsilon_i^2] = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 \qquad (1.07)$$

El criterio de Mínimos Cuadrados propone que los Estimadores de β_0 y β_1 sean los valores b_0 y b_1 que minimizan Q, para un conjunto dado de n pares $\binom{x_1}{y_1}; \binom{x_2}{y_2}; \ldots; \binom{x_n}{y_n}$.

Q una función de β_0 y β_1, la minimización de Q está determinada por las igualdades:

$$\frac{\delta Q}{\delta \beta_0} = \frac{\delta Q}{\delta \beta_1} = 0 \qquad (1.08)$$

Derivando con respecto a los parámetros β_0 y β_1 e igualando a cero, se tiene un sistema de dos ecuaciones:

$$\frac{\delta Q}{\delta \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (1.09)$$

Y

$$\frac{\delta Q}{\delta \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (1.10)$$

Los valores de $\beta_0\ y\ \beta_1$ que se obtienen al resolver (1.09) y (1.10), minimizan Q, y esta minimización puede ser verificada utilizando el criterio del signo de la segunda derivada de Q.

Llamaremos a b_0 y b_1 a los Estimadores de Mínimos Cuadrados para β_0 y β_1 respectivamente, de donde el sistema de ecuaciones se convierten en:

$$-2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \tag{1.11}$$

$$-2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \tag{1.12}$$

Que al simplificar determinan las **Ecuaciones Normales** que permite obtener una estimación de punto de los parámetros del modelo, éstas son:

$$\sum_{i=1}^{n} y_i = nb_0 + \sum_{i=1}^{n} x_i \tag{1.13}$$

$$\sum_{i=1}^{n} x_1 y_i = b_0\sum_{i=1}^{n} x_i + b_1\sum_{i=1}^{n} x_i \tag{1.14}$$

A partir de las Ecuaciones Normales se puede establecer:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{1.15}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Donde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ y $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Utilizando el trabajo previo se puede calcular el Coeficiente de Correlación Muestral $\hat{\rho}_{xy}$, que es una medida de la fuerza lineal que relaciona a Y con X; de los datos observados se los puede obtener sin dificultad; determinan que:

$$\hat{\rho}_{xy} = r_{xy} = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} \tag{1.16}$$

Siendo S_{xy} y S_{xx} los valores que aparecen en (1.15) esto es:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \text{ y } S_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{1.17}$$

Se puede además probar que la pendiente de la Recta de Regresión b_1 y r_{xy}, tienen igual signo.

σ^2, es un parámetro del modelo que al mismo tiempo es la Varianza del Error y también de Y_i. En este caso, Regresión Lineal Simple, la Suma Cuadrática del Error o Suma Cuadrática de los Residuos es denotada y definida como:

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (1.18)$$

Que mide la variabilidad de los valores observados alrededor de la recta cuya ecuación es $\hat{y} = b_0 + b_1 x$. La SCE tiene $(n-2)$ grados de libertad, puesto que se pierden dos grados de libertad al estimar β_0 y β_1; por lo que la Media Cuadrática del Error o Media Cuadrática Residual del Error es:

$$MCE = \sum_{i=1}^{n} \frac{(Y_i - \hat{y}_i)^2}{n-2} = \sum_{i=1}^{n} \frac{e_i^2}{n-2} \qquad (1.19)$$

Estimación en Regresión Lineal utilizando Máxima verosimilitud

El Criterio de Máxima Verosimilitud es un procedimiento estadístico para estimación de parámetros que obviamente también es aplicable en regresión lineal. Se requiere, por ejemplo, obtener los estimadores de β_0, β_1, σ^2, bajo el supuesto que el Error es Normal con Media cero y Varianza Constante σ^2, homocedasticidad, y además que $cov(\varepsilon_i, \varepsilon_j) = 0; i \neq j$, lo que implica que las y_i son estocásticamente independientes si tienen Distribución Normal con Media $\beta_0 + \beta_1 x_i$ y varianza σ^2; en síntesis:

$$\varepsilon_i \sim N(0, \sigma^2), \text{ para } i = 1; 2; \dots; n \quad \wedge \quad y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \qquad (1.20)$$

La densidad condicional de probabilidades para la i-ésima valor de β_0, β_1 y σ^2 es:

$$f(y_i) = f_i(y_i|\beta_0,\beta_1,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2} \quad (1.21)$$

Y la densidad conjunta de $y_1, y_2, \dots, y_n$ es:

$$f(y_1, y_2, \dots, y_n|\beta_0,\beta_1,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2} \quad (1.22)$$

Donde los y_i son estocásticamente independientes, el tratamiento de esta función en términos de parámetros nos lleva a la Función de Verosimilitud en término de β_0, β_1 y σ^2; que es $L(\beta_0,\beta_1,\sigma^2|y_1,y_2,\dots,y_n)$, donde n, como ya hemos señalado, es el número de pares del tipo $\mathbf{X}^T = (x_i, y_i)$, y el logaritmo de L

$$l(\beta_0,\beta_1,\sigma^2) = \ln L(\beta_0,\beta_1,\sigma^2) = K - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad (1.23)$$

Nótese que, $K = \left(-\frac{n}{2}\right)\ln(2\pi)$, es una constante que no depende de los parámetros a ser estimados.

A partir de la derivación con respecto ha β_0, β_1 y σ^2 se obtienen los estimadores de Máxima Verosimilitud de β_0, β_1, σ^2:

$$\frac{\partial\left(l(\beta_0,\beta_1,\sigma^2)\right)}{\partial\beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) \quad (1.24)$$

$$\frac{\partial\left(l(\beta_0,\beta_1,\sigma^2)\right)}{\partial\beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) \quad (1.25)$$

$$\frac{\partial\left(l(\beta_0,\beta_1,\sigma^2)\right)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 \quad (1.26)$$

Igualando a cero las derivadas y verificando el signo de la segunda derivada, se obtienen los Estimadores de Máxima Verosimilitud de los $\beta_0,\ \beta_1,\ \sigma^2$.

Para el caso de Regresión Lineal Simple, por Mínimos Cuadrados, se puede probar que, s^2 es un Estimador insesgado de σ^2:

$$s^2 = \text{MCE} = \hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{y}_i)^2}{n-2} = \sum_{i=1}^{n} \frac{e_i^2}{n-2} \qquad (1.27)$$

Mientras que por Máxima Verosimilitud, s^2 es un estimador de σ^2, siendo:

$$s^2 = \hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{y}_i)^2}{n} = \sum_{i=1}^{n} \frac{e_i^2}{n} \qquad (1.28)$$

Estos dos estimadores de σ^2 se relacionan dela siguiente manera:

$$s^2 = \text{MCE} = \frac{n}{n-2} S^2 \qquad (1.29)$$

> ***Nótese que:*** La estimación de los parámetros, utilizamos Máxima Verosimilitud que es equivalente a la de Mínimos Cuadrados excepto para σ^2.

1.2.3 Inferencias acerca de los parámetros de regresión

1.2.3 Valores Esperados de los Estimadores de Mínimos Cuadrados

El Teorema de Gauss Markov establece que los Estimadores de Mínimos Cuadrados, b_0 y b_1, para Regresión Lineal Simple son insesgados para $\beta_0\ y\ \beta_1$ y además se puede probar que son de Mínima Varianza en el Modelo de Regresión Lineal, siendo:

$$E(b_1) = E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \beta_1 \qquad (1.30)$$

$$E(b_0) = E[\bar{y} - b_1\bar{x}] = \beta_0$$

1.2.4 Tabla de Análisis de Varianza

En la tabla de Análisis de Varianza, con las Sumas Cuadráticas se pretende medir la dispersión de un grupo de observaciones.

Suma Cuadrática Total, es la suma de cada valor condicionado de y_i, menos el Valor Promedio de los mismos, y todo esto al cuadrado.

$$\text{SCT} = \sum_{i=0}^{n} (y_i - \bar{y})^2 = \sum_{i=0}^{n} y_i^2 - n\,\bar{y}^2 \tag{1.31}$$

Suma Cuadrática de Regresión, se define como la suma de cada $\hat{y}_i$ valor estimado de Y_i, menos el Valor Promedio de Y; todo al cuadrado.

$$\text{SCR} = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2 \tag{1.32}$$

La Suma Cuadrática de los Residuos, es la función Q que construyéramos para aplicar el Criterio de Mínimos Cuadrados y así estimar los parámetros $\beta_0\ y\ \beta_1$ y a la que hemos denominado SCE o Suma Cuadrática de los Residuos.

El Coeficiente de Determinación, es una medida de calidad del modelo que estamos utilizando y se la define como:

$$R^2 = \frac{SCR}{SCT} = \frac{SCtotal - SCError}{SCT} = 1 - \frac{SCR}{SCT}; \quad 0 < R^2 < 1 \tag{1.33}$$

La Potencia de Explicación del Modelo, es definida como porcentaje

$$Potencia\ de\ Explicación\ del\ Modelo = R^2 \cdot 100\%$$

Lo deseable es que la SCE sea lo más pequeña posible con respecto a la SCT, dando evidencia que entre más pequeña es la SCE más grande será la

Potencia de Explicación del Modelo, lo cual es buen indicio acerca de la calidad del modelo.

La Media Cuadrática de Regresión es igual a la Suma Cuadrática dividida para sus correspondientes grados de libertad, así, la MCR es:

$$\frac{\mathrm{SCR}}{p-1} \tag{1.34}$$

Mientras que a Media Cuadrática de los Residuos es:

$$\frac{\mathrm{SCE}}{n-p} \tag{1.35}$$

Con la aplicación del Teorema de Cochran, SCR/σ^2 es una Variable Aleatoria con Distribución Ji-Cuadrado con $(p-1)$ grados de libertad, mientras que SCE/σ^2 es una Ji-Cuadrado con $(n-p)$ grados de libertad, para el modelo de Regresión Lineal Simple $(n-p)=(n-2)$ grados de libertad. Esto para el caso de Regresión Lineal Simple permite afirmar que el cociente $\mathrm{F}_0=\frac{\mathrm{MCR}}{\mathrm{MCE}}=\frac{\mathrm{SCR}/1}{\mathrm{SCE}/(\mathrm{n}-2)}$, es una Variable Aleatoria F con $(p-1)=1$ grados de libertad en el numerador y $(n-p),(p=2)$ grados de libertad en el denominador.

$$F_0=\frac{\mathrm{MCR}}{\mathrm{MCE}}\sim\mathrm{F}(1,\mathrm{n}-1) \tag{1.36}$$

La Tabla de Análisis de Varianza ó Tabla ANOVA, para el Modelo de Regresión Lineal Simple, véase (Tabla 1), es utilizada en Regresión para analizar estadísticamente la validez del modelo $\mathrm{y_i}=\beta_0+\beta_1\mathrm{x_i}+\varepsilon_\mathrm{i}$ y los supuestos

$\mathrm{E}(\varepsilon_\mathrm{i})=0,\ \mathrm{VarE}(\varepsilon_\mathrm{i})=\sigma^2,\ \mathrm{cov}(\varepsilon_\mathrm{i},\varepsilon_\mathrm{j})=0\ \ \text{para}\ \ \mathrm{i}\neq\mathrm{j}$. Consiste en un arreglo rectangular cuyas componentes son las Fuentes de Variación, sus Grados de Libertad, las Sumas o Medias Cuadráticas y el Estadístico de Prueba F_0.

Tabla 1: Tabla de Análisis de Varianza para un modelo de Regresión lineal
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Fuentes de Variación	Grados de Libertad	Sumas Cuadráticas	Medias Cuadráticas	Estadístico de Prueba F
REGRESION	p-1	$\sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2$	$\frac{SCR}{p-1}$	$F_0 = \frac{MCR}{MCE}$
ERROR (Residuales)	n-p	$\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$	$\frac{SCE}{n-p}$	
TOTAL	n-1	$\sum_{i=0}^{n}(y_i - \bar{y})^2$		

Autoría: Fuentes A., Pinos R., Rivera N.

Nuestra aspiración es que dado el modelo de Regresión Lineal Simple, el valor de la pendiente β_1 de la recta no sea cero, por lo que postularemos el siguiente Contraste de Hipótesis.

$$\boldsymbol{H_0}: \beta_1 = 0 \ vs. \ \boldsymbol{H_1}: \beta_1 \neq 0 \qquad (1.37)$$

Si la Hipótesis Nula H_0 fuese verdadera, entonces, $\mathrm{E(MCR)} = \mathrm{E(MCE)} = \sigma^2$, por lo que el valor del Estadístico de Prueba F, al ser cercano a uno, mostraría evidencia estadística de que la Hipótesis Nula es verdadera, es decir $\beta_1 = 0$. Caso contrario, lo cual es deseable, si F_0 es "grande", rechazaríamos H_0, Nótese que suponemos a priori que $\beta_0 \neq 0$.

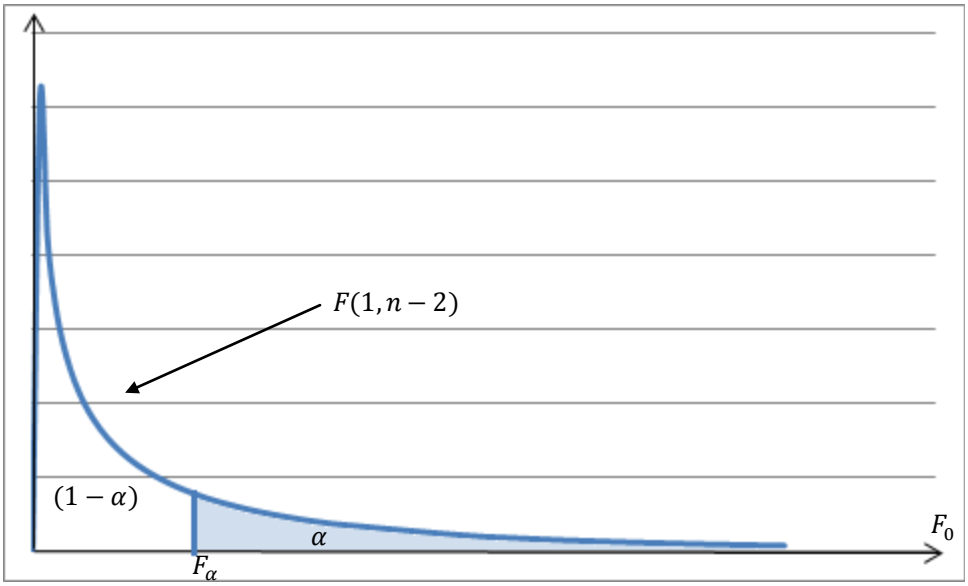
En otras palabras, con (1-α)100% de confianza, se debe rechazar H_0 en favor de H_1 si $\mathrm{F}_0 > F_{(\alpha)}$ donde $\mathrm{F}_{(\alpha)}$ es el percentil (1-α)100% de la variable

aleatoria F de Fisher con 1 grado de libertad en el numerador y $\mathrm{n}-2$ grados de libertad en el denominador, esto es:

$$\mathrm{F}=\frac{\frac{\chi^2(\upsilon_1)}{\upsilon_1}}{\frac{\chi^2(\upsilon_2)}{\upsilon_2}}\sim \mathrm{F}(\upsilon_1,\upsilon_2) \qquad (1.38)$$

$$\upsilon_1 = 1 \;\; y \;\; \upsilon_2 = n-2$$

Gráfico 1.03: $P(F \leq F_0) = 1-\alpha$
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

CAPITULO II

2. REGRESIÓN MÚLTIPLE

2.1 Introducción

Para comenzar este capítulo, hay que recordar de lo que trató Regresión Lineal Simple, que era explicar Y en términos de X, donde X es una sola variable, ahora, qué pasa cuando tenemos más de una variable que explican a Y, en estas circunstancia, nos plantemos las mismas condiciones, pero esta vez vamos a trabajar con matrices para poder denotar de una manera simplificada y formal las variables en los modelos, también con formas cuadráticas del tipo g(x) = $\mathbf{X}^{\mathrm{T}}\mathbf{AX}$. También veremos como se ve influenciada las hipótesis, supuestos y sobre todo la tabla de análisis de varianza (ANOVA).

2.2 Modelos Polinómicos

Dentro de los modelos Polinómicos alteramos un poco la forma en cómo solíamos explicar Y, en regresión lineal simple era, tomando como base que existía una relación rectilínea entre Y y X, pero cuando no es así y disponemos de una sola variable de explicación X, recurrimos a la expresión polinómica

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{x_i} + \beta_2 \mathrm{x_i}^2 + \varepsilon_\mathrm{i} \qquad (2.01)$$

De esta manera establecemos que también hay una relación cuadrática entre Y y X, si tomamos esto como cierto, se presentan los siguientes puntos.

Este modelo Polinómico de grado 2 tiene tres coeficientes (betas) y no dos como antes, ya que ahora tenemos también un término para $X^{2,}$ pero sigue siendo una sola variable de explicación, se mantienen los supuestos:

$$\mathrm{E}(\varepsilon_i) = 0 \ \mathrm{Var}(\varepsilon_i) = \sigma^2$$

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j; \qquad i, j = 1,2, \dots n(\varepsilon_i) \tag{2.02}$$

Los valores de β_0, β_1, β_2 y σ^2, son constantes desconocidas, nos basamos en un modelo Homocedástico y la Función de Respuesta sería:

$$\mu_{Y_i} = \mathrm{E}(Y_i) = \mathrm{E}(\beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \varepsilon_i)$$

$$= \mathrm{E}(\beta_0) + \mathrm{E}(\beta_1 x_i) + \mathrm{E}(\beta_2 {x_i}^2) + \mathrm{E}(\varepsilon_i) \tag{2.03}$$

La función de condicionamiento quedaría:

$$\mathrm{E}(Y_i | X = x_i) = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 \tag{2.04}$$

Similar a la Regresión Lineal Simple, ahora demos el siguiente paso, que es estimar β_0, β_1 y β_2 utilizando el Criterio de Mínimos Cuadrados, donde b_0, b_1 y b_2 minimizarán Q.

$$Q = \sum_{i=1}^{n} \left[Y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 \right) \right]^2 \tag{2.05}$$

Y Obtenemos:

$$\frac{\delta Q}{\delta \beta_0} = \frac{\delta Q}{\delta \beta_1} = \frac{\delta Q}{\delta \beta_2} = 0 \tag{2.06}$$

Y las igualdades de (2.06) nos conducirán a tres Ecuaciones Normales, que son:

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} {x_i}^2 \tag{2.07}$$

$$\sum_{i=1}^{n} x_1 y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2 + b_2 \sum_{i=1}^{n} {x_i}^3 \tag{2.08}$$

$$\sum_{i=1}^{n} x_2 y_i = b_0 \sum_{i=1}^{n} {x_i}^2 + b_1 \sum_{i=1}^{n} {x_i}^3 + b_2 \sum_{i=1}^{n} {x_i}^4 \tag{2.09}$$

Y si seguimos así, al ser un sistema lineal en b_0, b_1 y b_2, y de ser consistente, lograremos determinar los estimadores de β_0, β_1 y β_2.

2.3 Modelos de Regresión Lineal Múltiple

Cuando hablamos de Regresión Múltiple esto significa que existe más de una Variable de Explicación, por lo que consideraremos un modelo con p términos y $(p-1)$ variables de Explicación, suponiendo información de n casos, esto es: $i = 1, 2, \dots, n$.

El Modelo Lineal para el i-ésimo caso es el siguiente,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \tag{2.10}$$

Expresado el modelo de la forma matricial para n observaciones de Y y $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{\mathrm{p}-1}$ es:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2.11}$$

Que es el denominado Modelo Lineal General

Donde $\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $\boldsymbol{Y} \in \mathrm{R}^{\mathrm{n}}$

$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix}$ es la Matriz de Diseño del modelo y $\boldsymbol{X} \in \mathrm{M}_{\mathrm{n}\times\mathrm{p}}$

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ es el Vector de Estimadores, siendo $\boldsymbol{\beta} \in \mathrm{R}^{\mathrm{p}}$

$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ es denominado Vector de Errores, $\boldsymbol{\beta} \in \mathrm{R}^{\mathrm{n}}$

Entonces el Modelo, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, es expresado como:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.12}$$

Además debemos tener en cuenta que la Matriz $\boldsymbol{\Sigma}_{\varepsilon}$ de Varianzas y Covarianzas del Error es: $\boldsymbol{\Sigma}_{\varepsilon} = \sigma^2\mathbf{I}$, donde I es la Matriz identidad $n \times n$, y que los errores son independientes.

Siendo:

$$\boldsymbol{\Sigma}_{\varepsilon} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I} \in \mathbf{M}_{\mathrm{nxn}} \qquad (2.13)$$

Bajo los supuestos: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$, y $\boldsymbol{\Sigma}_{\varepsilon} = \sigma^2\mathbf{I}$

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}; \quad \mathbf{0} \in R^n \; ; \quad \sigma^2\boldsymbol{I} \in M_{nxn}$$

2.4 Estimación de los Parámetros

En el Modelo de Regresión Múltiple debemos estimar los p coeficientes $\beta_0, \beta_1, \ldots, \beta_{p-1}$, siendo el modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \; ; \quad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$\Sigma_{\boldsymbol{\varepsilon}} = \sigma^2\mathbf{I}$; $\Sigma_{\boldsymbol{\varepsilon}}$ es la Matriz de varianzas y covarianza

La estimación de los parámetros al igual que en los casos previos se la realiza bajo el Criterio de Regresión Lineal Múltiple de Mínimos Cuadrados de forma similar como lo hicimos en la *Sección 1.3.1*.

El Criterio de Mínimos Cuadrados propone que los Estimadores de los p parámetros $\beta_0, \beta_1, \dots, \beta_{p-1}$ del modelo, sean los valores $\mathrm{b}_0, \mathrm{b}_1, \dots, \mathrm{b}_{\mathrm{p}-1}$ que minimizan **Q**(2.05).

2.4.1 Estimación por Mínimos Cuadrados

En forma Matricial, deseamos encontrar el un vector de los estimadores de Mínimos Cuadrados, $\boldsymbol{\beta}$, que minimice:

$$\mathrm{L} = \sum_{\mathrm{i}=0}^{\mathrm{n}} \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \qquad (2.14)$$

L se puede expresar como:

$$L = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$$

$$\mathrm{L} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} \qquad (2.15)$$

Dado que $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ es una matriz (1x1), o un escalar, y su transpuesta $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$ es el mismo escalar. Los estimadores de Mínimos Cuadrados deben satisfacer

$$\left.\frac{\partial \mathrm{L}}{\partial \beta}\right|_{\hat{\beta}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \qquad (2.16)$$

Que se simplifica a:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \qquad (2.17)$$

Ésta es la forma matricial de las Ecuaciones Normales de Mínimos Cuadrados; para resolver estas ecuaciones, multiplicamos a ambos lados por la inversa de $\boldsymbol{X^T X}$, bajo el supuesto que $\boldsymbol{X^T X}$ no es singular, esto es, que $(\boldsymbol{X^T X})^{-1}$ existe, de tal modo que el estimador de Mínimos Cuadrados de $\boldsymbol{\beta}$ es:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} \qquad (2.18)$$

Se puede probar que esto también es válido para Regresión Lineal Simple, donde $p = 2$.

2.5 Inferencias acerca de los parámetros de regresión

Llamando al modelo $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, donde $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ y con Matriz de Covarianza $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\varepsilon}}} = \sigma^2\mathbf{I}$. Donde $\mathbf{X}$ es una matriz con rango p. Suponiendo Normalidad e independencia de los errores, el modelo implica que $\mathrm{E}[\mathbf{Y}] = \mathbf{X\beta}$ y $\mathrm{V}(\mathbf{Y}) = \sigma^2\mathbf{I}$. El estimador por el Criterio de Mínimos Cuadrados del vector de parámetros $\boldsymbol{\beta}$, es $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Da como resultado que sus estimadores obtenidos son insesgados, lo que significa que $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, puesto que:

$$E[\widehat{\boldsymbol{\beta}}] = E[(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}] = \boldsymbol{\beta} \qquad (2.19)$$

A demás $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}} = (\boldsymbol{X^T X})^{-1}\sigma^2$, pudiendo además estimar $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}}$ de la siguiente manera: $\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}} = \mathrm{MCE}(\boldsymbol{X^T X})^{-1}$, puesto que es estimador $\hat{\sigma}^2 = \mathrm{MCE}$.

2.6 Tabla de Análisis de Varianza para Regresión Múltiple

Para el Modelo de Regresión Múltiple o cualquier Modelo Lineal, con la notación usual tenemos:

Suma Cuadrática Total, para cualquier modelo y en su forma Matricial es:

$$\mathrm{SCT} = \sum_{i=0}^{n} (Y_i - \bar{y})^2 = \sum_{i=0}^{n} Y_i^2 - \frac{\left(\sum_{i=0}^{n} Y_i^2\right)^2}{n} \tag{2.20}$$

Expresando de forma matricial las expresiones:

$$\sum_{i=0}^{n} \mathrm{Y}_i^2 = \mathbf{Y}^{\mathrm{T}}\mathbf{Y}; \qquad \frac{\left(\sum_{i=0}^{n} Y_i^2\right)^2}{n} = \frac{1}{\mathrm{n}}\mathbf{Y}^{\mathrm{T}}\mathbf{J}\mathbf{Y}$$

Se puede probar que:

$$\mathrm{SCT} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{J}\mathbf{Y} \tag{2.21}$$

Donde J es una Matriz cuadrada nxn, cuyos elementos son todos 1.

Suma Cuadrática de Regresión, de igual manera:

$$\mathrm{SCT} = \mathrm{SCR} + \mathrm{SCE}$$

$$\mathrm{SCR} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{J}\mathbf{Y} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{2.22}$$

Suma Cuadrática del Error o *Suma Cuadrática de los Residuos.*

$$\mathrm{SCE} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{b}^{T}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{2.23}$$

Estas tres Sumas Cuadráticas podemos expresarlas de la siguiente manera, tal como lo hace Zurita [14].

$$\mathrm{SCT} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{J}\mathbf{Y}$$

$$\mathrm{SCT} = \mathbf{Y}^{\mathrm{T}}\left[\mathbf{I} - \left(\frac{1}{\mathrm{n}}\right)\mathbf{J}\right]\mathbf{Y}$$

$$\mathrm{SCT} = \mathbf{Y}^{\mathrm{T}}\mathbf{A}\mathbf{Y}$$

$$\mathrm{SCE} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{b}^{T}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\mathrm{SCE} = \mathbf{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{H}]\mathbf{Y}$$

$$\mathrm{SCE} = \mathbf{Y}^{\mathrm{T}}\mathbf{B}\mathbf{Y}$$

$$\mathrm{SCR} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{J}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \mathbf{b}^{T}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\mathrm{SCR} = \mathbf{Y}^{\mathrm{T}}\left[\mathbf{H} - \left(\frac{1}{\mathrm{n}}\right)\mathbf{J}\right]\mathbf{Y}$$

$$\mathrm{SCR} = \mathbf{Y}^{\mathrm{T}}\mathbf{C}\mathbf{Y} \qquad (2.24)$$

Siendo $\mathbf{H} \in S_{\mathrm{n}\times\mathrm{n}}$ la denominada *Matriz Hat* : $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}^{\mathrm{T}}$

Esta matriz sirve para visualizar los valores estimados de $\mathbf{Y}$ como combinaciones lineales de los valores observados de $\mathbf{Y}$, que se muestran en la siguiente forma:

$$\begin{bmatrix}\hat{\mathrm{Y}}_1\\ \hat{\mathrm{Y}}_2\\ \vdots\\ \hat{\mathrm{Y}}_{\mathrm{n}}\end{bmatrix} = \hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \begin{bmatrix}\mathrm{h}_{11} & \mathrm{h}_{12} & \mathrm{h}_{13} & \cdots & \mathrm{h}_{1\mathrm{n}}\\ \mathrm{h}_{21} & \mathrm{h}_{22} & \mathrm{h}_{23} & \cdots & \mathrm{h}_{2\mathrm{n}}\\ \vdots & \vdots & \vdots & \cdots & \vdots\\ \mathrm{h}_{\mathrm{n}1} & \mathrm{h}_{\mathrm{n}2} & \mathrm{h}_{\mathrm{n}3} & \cdots & \mathrm{h}_{\mathrm{nn}}\end{bmatrix}\begin{bmatrix}\mathrm{Y}_1\\ \mathrm{Y}_2\\ \vdots\\ \mathrm{Y}_{\mathrm{n}}\end{bmatrix} \qquad (2.25)$$

Además se puede probar que la matriz $\mathbf{H}$ es idempotente, es decir que: $\mathbf{HH} = \mathbf{H^2} = \mathbf{H}$.

La versión matricial de la Tabla de Análisis de Varianza para un Modelos de Regresión Lineal Múltiple se representa en la Tabla 2

Tabla 1.01: Análisis de varianza Regresión Múltiple
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Fuentes de Variación	Grados de Libertad	Sumas Cuadráticas	Medias Cuadráticas	Estadístico de Prueba F
REGRESION	$p-1$	$\boldsymbol{Y^TY}-\frac{1}{n}\boldsymbol{Y^TJY}-\boldsymbol{Y^TY}-\boldsymbol{b^TX^TY}$	$\frac{SCR}{p-1}$	$F_0=\frac{MCR}{MCE}$
ERROR (Residuales)	$n-p$	$\boldsymbol{Y^TY}-\boldsymbol{b^TX^TY}$	$\frac{SCE}{n-p}$	
TOTAL	$n-1$	$\boldsymbol{Y^TY}-\frac{1}{n}\boldsymbol{Y^TJY}$		

Autoría: Fuentes A., Pinos R., Rivera N.

Para medir la calidad del modelo que estamos utilizando hacemos uso del *Coeficiente de Determinación*, un valor pequeño de R^2 es indicio de Independencia entre **X** y **Y**:

$$R^2=\frac{SCR}{SCT};\quad 0<R^2<1 \tag{2.26}$$

La Potencia de Explicación del Modelo, es definida como porcentaje

$$Potencia\ de\ Explicación\ del\ Modelo = R^2.100\%$$

Planteamos el siguiente Contraste de Hipótesis, para verificar si existe evidencia de que al menos uno de los $(p-1)$ coeficientes, que hemos propuesto en $\boldsymbol{H_0}$ es realmente distinto de cero;

$$\boldsymbol{H_0}: \beta_i = 0\ ;\ \ i=0,\dots,p-1$$

$$vs.$$

$$\boldsymbol{H_1}: Al\ menos\ un\ \beta_i \neq 0;\ \ i=0,\dots,p-1 \tag{2.27}$$

Si la Hipótesis Nula del contraste es rechazada, como es la expectativa del investigador, habría que buscar cual o cuales de los parámetros (betas) no es cero, puesto que esos términos serían los que aportan de manera significativa a explicar $\mathbf{Y}$.

El estadístico de Prueba F_0, es definido de la misma forma que en Regresión Lineal Simple, como lo vimos $\mathrm{F}_0 = \frac{\mathrm{MCR}}{\mathrm{MCE}} = \frac{\mathrm{SCR}/(\mathrm{p}-1)}{\mathrm{SCE}/(\mathrm{n}-\mathrm{p})}$, que es una Variable Aleatoria F con $(p-1)$ grados de libertad en el numerador y $(n-p)$ grados de libertad en el denominador.

$$\mathrm{F}_0 = \frac{\mathrm{MCR}}{\mathrm{MCE}} \sim \mathrm{F}(\mathrm{p}-1, \mathrm{n}-\mathrm{p}) \qquad (2.28)$$

Se puede probar que bajo los supuestos de Normalidad e independencia del error ϵ_i, la Variable Aleatoria $\frac{\mathrm{b_i}-\beta_\mathrm{i}}{\mathrm{s_{b_i}}}$ tiene distribución T de Student con $(n-p)$ grados de libertad. $\frac{\mathrm{b_i}-\beta_\mathrm{i}}{\mathrm{s_{b_i}}} \sim \mathrm{T}_{\mathrm{n}-\mathrm{p}}$.

Con este resultado si $\boldsymbol{H_0}$ es rechazado en (2.29), proponemos los (p-1) contrastes:

$$\boldsymbol{H_0}: \beta_\mathrm{i} = 0 \ \ vs. \ \boldsymbol{H_1}: \beta_\mathrm{i} \neq 0 \qquad\qquad i = 1,2,3,\dots,(p-1)$$

Siendo el estadístico de prueba $\mathrm{T} = \frac{\mathrm{b_i}}{\mathrm{s_{b_i}}}$

Se rechaza la Hipótesis Nula $\boldsymbol{H_0}$ a favor a la Hipótesis Alterna $\boldsymbol{H_1}$, con $(1-\alpha)100\%$ de confianza sí:

$$|\mathrm{T}| > t_{\left(\frac{\alpha}{2}, n-p\right)}$$

Siendo $t_{\left(\frac{\alpha}{2}, n-p\right)}$ el percentil $\left(1-\frac{\alpha}{2}\right)100$ de la distribución T con $(n-p)$ grados de libertad.

CAPITULO III

3. MODELO DE REGRESIÓN NO LINEAL

3.1 Introducción

En este capítulo se presentan las familias exponenciales que permiten descomponer las distribuciones exponenciales tales como Normal, Poisson, Binomial, en términos de funciones lineales de tal manera que se crea un "enlace" mediante una relación algebraica.

El Modelo Lineal Generalizado, nace cuando las variables de Y y X no están relacionadas de una manera directa y utilizando las familias exponenciales se creó una función de "enlace", la cual permite utilizar los mismos métodos que fueron aplicados para calcular los estimadores de beta, como mínimos cuadrados y máxima verosimilitud, pero en este caso las ecuaciones no tienen solución explicita, sino una solución implícita lo que hace que se necesite un método numérico.

Existen algunos métodos que permiten resolver esta situación entre los cuales se encuentran el método de Newton-Raphson y el de Gauss-Jordan, siendo el método escogido el Newton-Raphson que es de rápida convergencia y sencilla programación.

3.2 Familia de Funciones Exponenciales

La familia exponencial es una clase de distribuciones de probabilidad cuya formulación matemática comparten cierta forma. Esta forma especial es escogida por interés matemático, que confiere a las distribuciones de esta familia una serie de propiedades algebraicas y estadísticas. Incluye distribuciones, sean estas continuas o discretas como la normal, binomial, etc.

El concepto de la familia exponencial fue introducido por E. J. G. Pitman [16], G. Darmois [17], and B. O. Koopman [18] en 1935.

En sí hay varias expresiones para definir las familias exponenciales, aunque todas responden a una definición general que pasamos a presentar.

Considérese una variable aleatoria Y cuya distribución de probabilidades depende de un parámetro θ. La distribución pertenece a las familias exponenciales si puede ser escrita de la forma.

$$f(x;\theta) = h(x)g(\theta)e^{[\eta(\theta)T(x)]} \qquad (3.01)$$

Donde $h, g, \eta\ y\ T$ son funciones conocidas, Nótese la simetría entre x y θ. Esto se enfatiza si la ecuación (3.01) es reescrita como:

$$f(x;\theta) = \exp[\eta(\theta)T(x) + c(\theta) + d(x)] \qquad (3.02)$$

Donde $h(x) = exp[d(x)]$ y $g(\theta) = \exp[c(\theta)]$.

Si $T(x) = x$, la distribución se dice que está en su Forma Canoníca (esto es, estándar), y $\eta(\theta)$ es llamada el parámetro natural de la distribución.

A $\eta(\theta)$ se lo conoce como Parámetro Natural, que nos proporciona en sí el "enlace" que se utilizará más adelante; $\eta(\theta)$ especifica los parámetros necesarios para dicha distribución.

$g(\theta)$ es el factor de "normalización", que asegura que $p(x|\theta)$ siga siendo una distribución de probabilidad

$T(x)$ es el estadístico suficiente de la "información".

$h(x)$ es una base de medida no negativa, que es generalmente 1.

Si hay otras variables en la función, además del parámetro de interés θ, son relegadas como parámetros ruido formando parte de las funciones η, T, c y d, Muchas distribuciones bien conocidas pertenecen a la familia exponenciales. Por ejemplo, Poisson, Normal, Binomial que pueden ser escritas en su forma canónica, véase Tabla 3.

Tabla 3: Distribuciones de la familia exponencial.
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Distribución	Parámetro Natural	c	d
Poisson	$\log \theta$	$-\theta$	$-\log x!$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$	$\frac{x^2}{2\sigma^2}$
Binomial	$\log\left(\frac{\pi}{1-\pi}\right)$	$n\log(1-\pi)$	$\log\binom{n}{y}$

Autoría: Fuentes A., Pinos R., Rivera N.

A continuación se ofrecen algunas ilustraciones de la representación de algunas familias de funciones de densidad de acuerdo con el formalismo de miembro de la familia de exponenciales.

Distribución Binomial

Como miembro de la familia exponencial consideremos la variable aleatoria Bernoulli. Su función de probabilidad es:

$$p(x;\theta) = \theta^{x}(1-\theta)^{1-x}$$

$$= \exp\{\, x\, log\theta + (1-x)\log(1-x)\}$$

$$= \exp\left\{x\, log\, \frac{\theta}{1-\theta} + \log(1-\theta)\right\} \qquad (3.03)$$

Se define:

Parámetro natural $\eta(\theta) = log\, \frac{\theta}{1-\theta}$

Factor de normalización $g(\eta(\theta)) = \log\left(1 + \exp\left(\eta(\theta)\right)\right)$

$$= \log\left(1 + \frac{\theta}{1-\theta}\right)$$

$$= -\log(1-\theta\,) \quad (3.04)$$

Estimador suficiente de la distribución $T(x) = \mathrm{x}$

Base de medidas $h(x) = 1$

Distribución Poisson

Para la distribución Poisson se hace algo similar al descomponerlo en una familia exponencial, su función de probabilidad es:

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Para llevar esta expresión a su forma de familia exponencial es cuestión de un poco de algebra:

$$\frac{e^{-\lambda}\lambda^{x}}{x!} = exp\left[\log\left(\frac{e^{-\lambda}\lambda^{x}}{x!}\right)\right]$$

$$= \exp\left[\log\left(e^{-\lambda}\right) + \log(\lambda^{x}) - \log\left(y!\right)\right]$$

$$= exp[-\lambda + ylog(\lambda) - \log\left(y!\right)]$$

Se define:

Parámetro natural $\eta(\theta) = log\,\lambda$

Factor de normalización $g(\eta(\theta))$ = λ

Estimador suficiente de la distribución $T(x) = -\log\left(y!\right)$

Base de medidas $h(x) = 1$

Distribución Normal

Tenemos una distribución $N(\theta, 1)$, la función de densidad puede ser escrita según (3.1) de la siguiente manera:

$$f(x;\theta)\frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2} = \underbrace{\frac{1}{\sqrt{2\pi}}e^{-\theta^2/2}}_{g(\theta)} * \underbrace{\left(e^{-x^2/2}\right)}_{h(x)} * e^{\theta x} \tag{3.05}$$

Y $\eta(\theta) = \theta$, $T(x) = \mathrm{x}$, quedando:

$$g(\theta) = \frac{1}{\sqrt{2\pi}}e^{-\theta^2/2}, \qquad h(x) = e^{-x^2/2}$$

$$f(x;\theta) = h(x)g(\theta)e^{\theta x} \tag{3.06}$$

3.3 Modelo Lineal Generalizado

Un Modelo Lineal Generalizado es una generalización de la Regresión Lineal para poder responder a otros tipos de modelos además de los lineales siempre y cuando la distribución de la respuesta sea miembro de las familias exponenciales.

Vamos a suponer que se trata de predecir la variable Y de un grupo de variable X. En un modelo lineal con parámetros β, suponemos que:

$$\mathrm{E}\big(\mathrm{Y}(\mathbf{X})\big) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} \qquad (3.07)$$

La generalización se obtiene al suponer que $\mathrm{E}\big(\mathrm{Y}(\mathbf{X})\big)$ no es igual a la combinación lineal $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$, pero que está relacionado con este, por medio de una función de acuerdo a la naturaleza de Y. Formalmente el modelo lineal Generalizado consiste en 3 componentes:

1) El "componente aleatorio" Y(variable de respuesta), que tiene distribución de las familias exponenciales con un parámetro canónico η que determina la forma de la respuesta, por ejemplo, Poisson. Nótese que se necesita poder escribir la distribución de la familia exponencial en su forma canónica.
2) El ''componente sistemático'' que especifica que las covariables $\boldsymbol{X}$ sean parte del modelo por la combinación lineal $\boldsymbol{X^T\beta}$ y dado que estamos en la familia exponenciales, ellos definen el parámetro natural η.

3) Una función diferenciable y monótona g, que conecta el componente sistemático con el parámetro $\boldsymbol{\theta}$.

$$\mathrm{g}(\boldsymbol{\theta}) = \boldsymbol{X^T \beta} \tag{3.08}$$

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\theta} = \mathrm{g}^{-1}(\boldsymbol{X^T \beta}) \tag{3.09}$$

g es llamada la función de enlace y es la inversa de la función de respuesta. Dado $\boldsymbol{X^T \beta} = \eta$, la función de respuesta es la misma que la función de asignación entre el parámetro natural y el parámetro $\eta(\boldsymbol{\theta})$

Ejemplo:

Para el caso de la denominada "Regresión Logística", que ampliaremos en el capítulo 4, se utiliza la distribución Bernoulli como variable de respuesta, que como verificamos en líneas previas, tiene como función de enlace:

$$\mathrm{g}(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta}) = \log\left(\frac{\boldsymbol{\theta}}{1-\boldsymbol{\theta}}\right) \tag{3.10}$$

La función de respuesta es:

$$\mathrm{g}^{-1}(\eta) = \frac{1}{1+\exp(\eta)} = \frac{1}{1+\exp(\boldsymbol{X^T \beta})} \tag{3.11}$$

3.3.1 Distribuciones y Funciones de enlace

Como se insinuó en el ejemplo previo, el Modelo Lineal General con variable de respuesta Y está linealmente asociado a los valores de la variable de explicación X por:

$$\mathrm{Y} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} \tag{3.12}$$

Mientras que la relación en el Modelo Lineal Generalizado se define por:

$$\mathrm{Y} = \mathrm{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) \tag{3.13}$$

Siendo $\mathrm{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ una función, la función inversa de $\mathrm{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ es $\mathrm{f}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$ que es denominada “función de enlace”. Se obtiene:

$$\mathrm{f}(\varpi_{\mathrm{y}}) = \mathrm{g}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}) \tag{3.14}$$

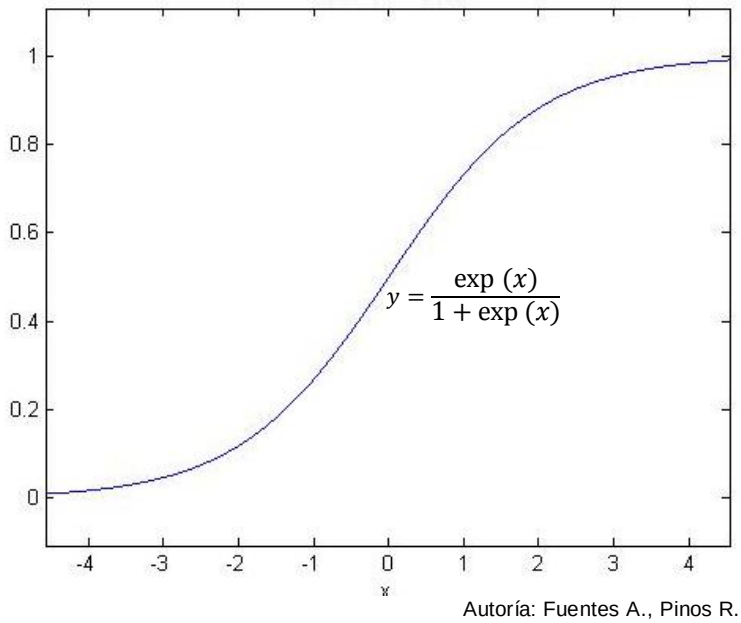
Donde ϖ_{y} representa al valor esperado de Y.

Varias funciones de enlace pueden ser escogidas dependiendo de la distribución de los valores de la variable de respuesta que hemos denominado Y.

Para diferenciar los modelos lineales generalizados, vamos a graficar algunas funciones de respuesta generalmente utilizados y ver la relación que hay entre las variables implicadas.

Para $f(x) = \frac{\exp(x)}{1+\exp(x)}$, que es el parámetro natural de la distribución Bernoulli.

Gráfico 3.01: Función de respuesta $f(x) = \frac{\exp(x)}{1+\exp(x)}$
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

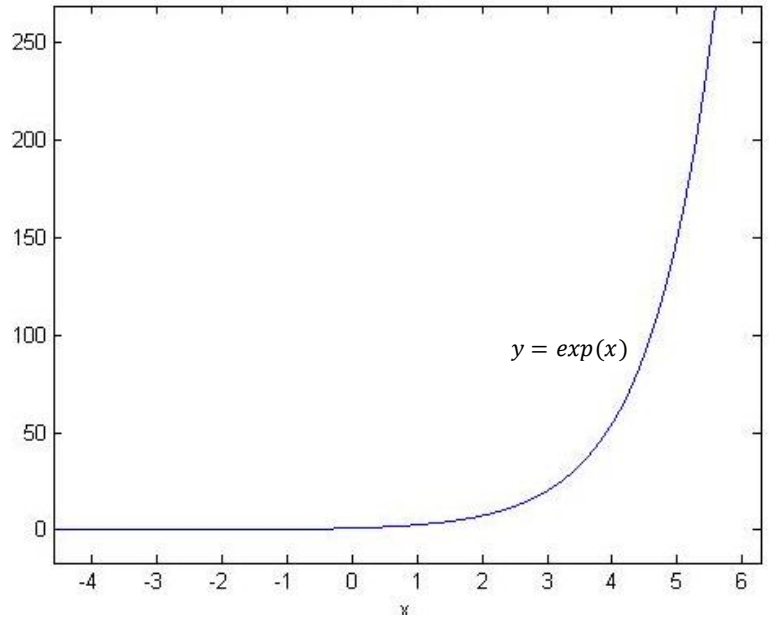


Autoría: Fuentes A., Pinos R., Rivera N.

Se puede observar en el grafico 3.01, que los valores de Y se encuentran entre 0 y 1, lo cual es ideal para este modelo, donde la variable a ser explicada, toma valores 0 y 1, que permitirá al modelo calcular la probabilidad de ocurrencia en un valor especifico de X.

Para $f(x) = \exp(x)$, que es el parámetro natural de la distribución Poisson.

Gráfico 3.02: Función de respuesta $y = exp(x)$
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Como se puede observar en el Grafico 3.02, está la función de enlace que se utiliza en la Regresión Poisson, que a diferencia del Grafico 3.01, los valores de la variable Y van de 0 a infinito, de esta manera se podrá modelar valores de Y enteros, esto es 0, 1, 2,…, y así calcular que valor tomara Y en cada valor especifico de X.

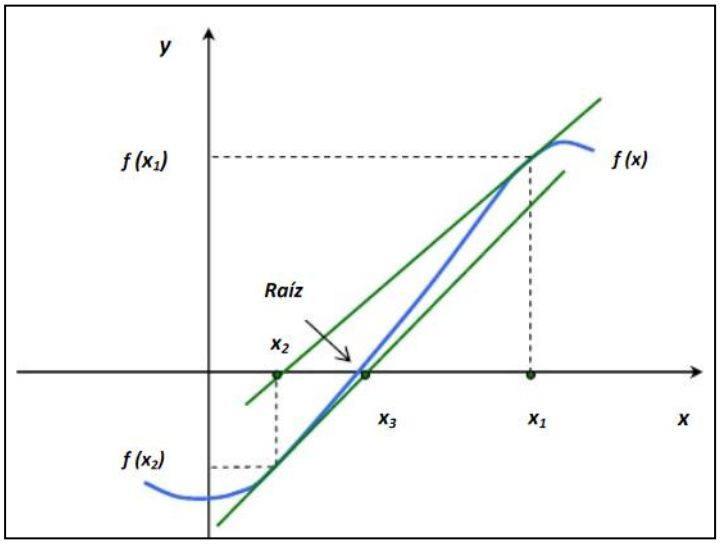
3.4 Método de Newton-Raphson para determinación de mínimo de una función

El Método de Newton-Raphson es un procedimiento numérico; se utiliza para encontrar raíces de una función o ecuaciones por aproximaciones sucesivas usando la tangente, que no es otra cosa que comenzar con un valor cercano a cero, y después ir determinando las rectas tangentes a la función que se nos plantea, hasta que encontremos uno que se aproxime lo suficiente a la raíz.

Veámoslo ayudados por un gráfico:

Pensemos en una función f cuya regla de correspondencia es $f(x)$ y queremos hallar una de sus raíces, si existe. Para ello, escogemos un valor x_1, "cercano" a la raíz de la función, y trazamos una recta tangente que incluirá el punto x_2, Calculamos $f(x_2)$, este punto, nos dará un nuevo valor x_3, que es más cercano a la raíz que queremos calcular.

Gráfico 3.03: Newton-Raphson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Para encontrar el valor dex_2, se tomará la ecuación punto pendiente.

$$f(x_2) - f(x_1) = m(x_2 - x_1) \tag{3.15}$$

Para que x_2 sea una raíz de$f$, $f(x_2)$ tendrá que ser igual a 0, para mayor comprensión, reemplazamos $f(x_2)$ por "y"; el enunciado quiere decir, hacemos $y = 0$ para poder hallar x_2:

$$-f(x_1) = m(x_2 - x_1) \tag{3.16}$$

Ahora tomamos "m" como $f'(x_2)$, al ser la pendiente de la recta tangente a la función en el punto x_2, $f(x_2)$ nos dará una mejor aproximación:

$$-f(x_1) = f'(x_1)(x_2 - x_1) \rightarrow\ x_2 - x_1 = \frac{f(x_1)}{f'(x_1)} \tag{3.17}$$

Ponemos la ecuación en función de x_2:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \tag{3.18}$$

Al generalizar de manera inductiva, quedará:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \qquad k \in \mathbb{Z}^+ \tag{3.19}$$

La ecuación 3.19 es la que se conoce como Ecuación de Newton-Raphson.

Esta no es la única forma de llegar a deducir el algoritmo de Newton-Raphson, hay un método alternativo, que es la función $f(x)$ en serie de Taylor, para un entorno del punto x_n:

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + (x - x_n)^2 \frac{f''(x_n)}{2!} + \cdots$$

Si se trunca el desarrollo a partir del término de grado 2, y evaluamos en x_{n+1}:

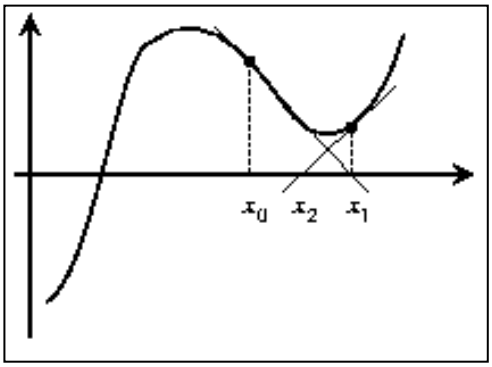
$$f(x_{n+1}) = f(x) + f'(x_n)(x - x_n) \tag{3.20}$$

Si además se acepta que x_{n+1} tiende a la raíz, se ha de cumplir que $f(x_{n+1}) \cong 0$, luego, sustituyendo en la expresión anterior, obtenemos el algoritmo.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{3.21}$$

Un inconveniente de este metodo, es de la existencia de falsas raices de la funcion, que no hacen que $f(x) = 0$

Gráfico 3.04: Inconvenientes del Método de Newton-Raphson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Para resolver este inconveniente, tenemos que incluir en el algoritmo la segunda derivada de la función, que nos asegurará que la raíz que buscamos sea cuando $f(x)$ es igual a 0, y lo logramos gracias al método de Taylor, dándole desarrollo hasta el grado 2.

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + (x - x_n)^2 \frac{f''(x_n)}{2!} \qquad (3.22)$$

Que al ponerlo de manera matricial queda:

$$\mathrm{f}(\beta) \cong \mathrm{f}(\beta^*) + \mathbf{g}(\beta^*)^{\mathrm{T}}(\beta - \beta^*) + \frac{1}{2}(\beta - \beta^*)^{\mathrm{T}}\mathrm{G}(\beta^*)(\beta - \beta^*) \qquad (3.23)$$

Donde $\mathbf{g}(\beta)$ es la columna del vector de la primera derivada of $\mathrm{f}(\beta)$ con respecto a β, este vector tiene elementos de $\frac{\partial \mathrm{f}(\beta)}{\partial \beta_1}, \frac{\partial \mathrm{f}(\beta)}{\partial \beta_2}, \ldots, \frac{\partial \mathrm{f}(\beta)}{\partial \beta_\mathrm{p}}$. El vector $[\mathbf{g}(\beta)]^{\mathbf{T}}$ es la transpuesta de $\mathrm{g}(\beta)$, y la notación $[\mathbf{g}(\beta^*)]^{\mathbf{T}}$ expresa el hecho de que el vector de las derivadas se evalúa en $\beta = \beta^*$. $\mathbf{G}(\beta)$, la segunda derivada es denotada como $\mathbf{G}(\beta^*)$ indica que las derivadas se evalúan en $\beta = \beta^*$. La Matriz de Segundas derivadas es llamada MATRIZ HESSIANA.

Diferenciando la ecuación anterior con respecto a los elementos de los rendimientos de β.

$$g(\beta) \cong g(\beta^*) + G(\beta^*)(\beta - \beta^*) \tag{3.24}$$

Como el vector de las primeras derivadas de $g(\beta^*) = 0$ en el óptimo β^*. Dejando en términos de β^* nos lleva a:

$$\beta^* \cong \beta - [G(\beta^*)] - \frac{1}{2}g(\beta) \tag{3.25}$$

Ejemplo:

Encuentre el o los valores de x que satisfacen la siguiente ecuación:

$$f(x) = e^x - \pi x = 0$$

Para resolver este problema por el método de Newton-Raphson se puede aplicar directamente con la función tal y como está. Se comienza calculando la primera derivada de $f(x)$.

$$f(x) = e^x - \pi x$$

$$f'(x) = e^x - \pi$$

Se toma $x_0 = 0.5$ por ser un valor pequeño y sencillo de calcular en la función y en su derivada.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{e^{x_0} - \pi x_0}{e^{x_0} - \pi}$$

$$evaluando\ para\ x_0 = \frac{1}{2} \Longrightarrow 0.5 - \frac{e^{0.5} - 0.5\pi}{e^{0.5} - \pi} = 0.5522$$

Las iteraciones realizadas se muestran en la Tabla 4:

Tabla4: Iteraciones-Newton Raphson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Iteración	x_i	x_{i+1}	error=$x_{i+1} - x_i$
1	0.5	0.5522	0.0522
2	0.5522	0.5538	0.0016
3	0.5538	0.5538	0.0000

Autoría: Fuentes A., Pinos R., Rivera N.

En los resultados se observa la rápida convergencia hacia el valor de la raíz. En la tercera iteración el resultado tiene cuatro decimales que coincide con la segunda iteración.

ahora que hemos presentado este procedimiento numérico, en secciones venideras podremos determinar las raíces de varias ecuaciones derivadas de procesos de estimación de parámetros que no están presentados de forma explícita. Pasamos a explicar más de la relación entre las familias exponenciales con la Regresión Logística y con Regresión Poisson.

3.5 Función de enlace para Regresión Logística

Considerando el caso en el cual $Y_1, Y_2, \dots, Y_n$ son Bernoulli (Independientes con Probabilidad de éxito $\theta = \pi$.

$$E(Y_i) = \pi_i = P(Y_i = 1) \tag{3.26}$$

La relación entre x_i y π es

$$\ln\frac{\pi(x)}{1-\pi(x)} = \boldsymbol{X^T\beta}$$

$$f_p(\boldsymbol{x}) = p^y(1-p)^{1-y} = (1-p)exp\left(y\ln\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right)\right) \qquad (3.27)$$

Donde $\ln\frac{\pi(x)}{1-\pi(x)}$ es un parámetro "natural" de la familia exponencial y se lo usa como "enlace"

$$g(\pi) = \ln\frac{\pi}{1-\pi}$$

Entonces

$$\pi(x) = \frac{e^{\boldsymbol{X^T\beta}}}{\boldsymbol{1}+e^{\boldsymbol{X^T\beta}}};\ \pi_i \in (0,1) \qquad (3.28)$$

3.6 Función de Enlace para Regresión Poisson

La Función de enlace con notación g es estándar en el Modelo Lineal Generalizado. Para el Modelo de Regresión Logística, la función de enlace es $g(\mu) = \ln\frac{\mu}{1-\mu}$ que responde a una Distribución Bernoulli. En el Modelo de Regresión Poisson la función de enlace es el logaritmo $g(\mu) = \ln\mu$, que responde a una distribución Poisson. Estas funciones de enlace son funciones monótonas de μ, esto es, $g(\mu) \geq g(\mu^*)$ para $\mu > \mu^*$.

La distribución de Poisson escritas por las probabilidades

$$P(Y_i = 1) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0,1,2,\ldots \tag{3.29}$$

Su media y varianza está dada por $\mu > 0$.

Si y es Poisson con parámetro μ, $E(y) = var(y) = \mu$

La media de la Distribución Poisson puede depender de las variables explicativas, pero la relación no puede ser Lineal porque esto podría conducir a valores negativos para μ. sin embargo la función de enlace

$$g(\mu) = \ln(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \tag{3.30}$$

Satisface la restricción de No Negatividad.

CAPITULO IV

4. REGRESIÓN LOGÍSTICA Y REGRESIÓN POISSON

4.1 Introducción

En este capítulo presentamos el módulo específico en el que hemos centrado nuestro trabajo en el paquete estadístico ERLA, que es Regresión Logística y Regresión Poisson; hemos explicado ya lo fundamental que nos permitirá entender y aplicar este tipo poco convencional de Regresión, pues utilizaremos Modelos Lineales Generalizados.

4.2 Regresión Logística

La regresión logística es un modelo no lineal mediante el cual se puede determinar la relación entre una variable de respuesta Y que es binaria y una o más variables de explicación $X_1, X_2, \ldots, X_k$, que son variables continuas.

A continuación se presenta la variable aleatoria X a la que se denominamos Distribución Logística con parámetro θ, su densidad es,

$$f(x) = \frac{e^{-(x-\theta)}}{(1 + e^{-(x-\theta)})^2} \text{ ; con soporte S = R; } \theta \in R \qquad (4.01)$$

Para el caso cuando θ es cero se lo llama Distribución Logística, la cual es:

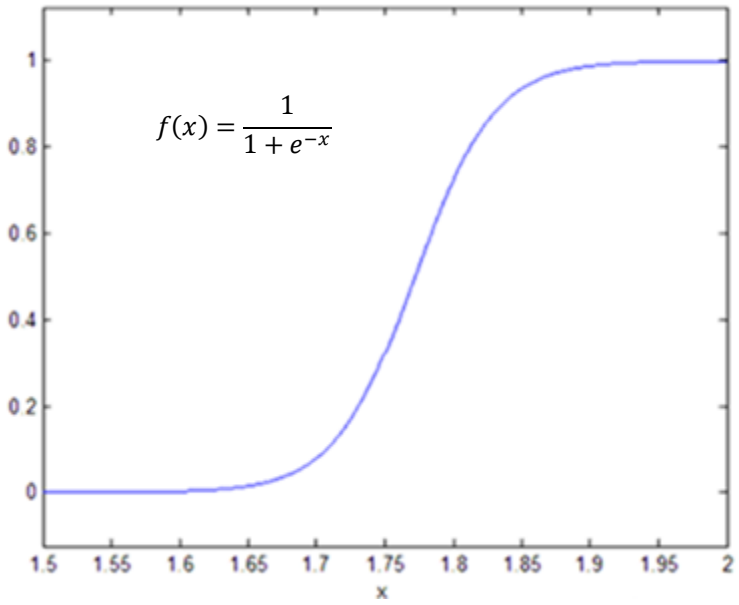
$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2} \text{ ; S = R} \qquad (4.02)$$

Su Distribución Acumulada F(x) = P(X≤ x) es;

$$\mathrm{P}(\mathrm{X} \leq \mathrm{x}) = F(x) = \frac{1}{1 + e^{-x}}, x \in \mathrm{R} \qquad (4.03)$$

La representación gráfica de f(x) se presenta en el Gráfico 4.01

Gráfico 4.01: Distribución Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Como pueden apreciar f es una curva que se extiende sobre R y cuyo dominio en el intervalo real que va desde cero hasta uno; la curva presentada es monótona creciente.

4.2.1 INTERPRETACIÓN DE LOS PARÁMETROS

Recordando las familias exponenciales en el Capítulo 3, las que permiten que la distribución de Bernoulli sea definida en términos lineales:

$$p(x;\theta) = \theta^{x}(1-\theta)^{1-x}$$

$$= \exp\{\, x\, log\theta + (1-x)\log(1-x)\}$$

$$= \exp\left\{x \log \frac{\theta}{1-\theta} + \log(1-\theta)\right\} \qquad (4.04)$$

Con este resultado y junto con lo que los modelos lineales generalizados definen, tomamos la "función de enlace" de la distribución.

$$g(\theta) = \log\left(\frac{\theta}{1-\theta}\right) = \boldsymbol{X^T\beta} \qquad (4.05)$$

Y se obtiene la función de respuesta al invertir la función de enlace:

$$\mathrm{g}^{-1}(\theta) = \frac{1}{1+\exp(\theta)} \qquad (4.06)$$

Reemplazando $\theta = \mathbf{X}^T\boldsymbol{\beta}$, se obtiene la función de respuesta de la regresión logística

$$\mathrm{Y} = \mathrm{f}\left(\mathbf{X^T\beta}\right) = \frac{1}{1+\exp(\mathbf{X^T\beta})} \qquad (4.07)$$

4.2.2 Estimación de parámetros en un modelo de Regresión Logística

En Regresión Logística la estimación de los coeficientes del modelo y de sus errores estándar se recurre al método de Máxima Verosimilitud, es decir, estimaciones que hagan máxima, la probabilidad de obtener Y proporcionados por los datos de la muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de los coeficientes en la Regresión Lineal Simple o Múltiple que efectuáramos en los capítulos 1 y 2 de este trabajo. Para el cálculo de estimaciones máximo–verosímiles en Regresión Logística, ya que no se obtienen expresiones explícitas para los valores de "los betas" incluidos en el modelo y por tanto debe recurrirse a métodos iterativos, como lo hemos enunciado, usaremos el método de Newton–Raphson (Capítulo 3).

Utilizar Métodos Numéricos por ser procesos iterativos puede llevarnos a cálculos tediosos, hace necesario que se recurra al uso de rutinas de programación de computadoras. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de sus errores estándar y de las covarianzas entre las variables de explicación del modelo.

Para aplicar el método de Máxima Verosimilitud en Regresión Logística se trabaja con que cada observación Y_i de la muestra sigue la distribución de Bernoulli, suponiendo independencia de las n observaciones, donde la densidad de probabilidades conjuntas, dado $\boldsymbol{\beta}$, de $y_1, y_2, \ldots, y_n$ está dada por:

$$\mathrm{p}\,(y_1, y_2 \ldots, y_n | \boldsymbol{\beta}) = \prod_{i=1}^{n} [{\pi_i}^{y_i}]\,[1 - \pi_i]^{1-y_i} \qquad (4.08)$$

Entonces la función de verosimilitud está dada por:

$$\mathrm{L}\,(\boldsymbol{\beta} \mid y_1, y_2, \ldots, y_n) \;=\; \sum_{i=1}^{n} y_i \ln \pi_i \;+\; \sum_{i=1}^{n} (1 - y_i) \ln\,(\,1 - \pi_i) \qquad (4.09)$$

Las condiciones son las siguientes:

La variable Y, que es la variable dependiente, al ser n veces observada, condicionado a valores de X, genera una matriz de n filas y 1 columna:

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}$$

Además, un conjunto de p variables, que podemos expresar como una matriz de n filas y p columnas. Sin embargo, dado que el modelo contiene una constante, ésta se expresa como una columna adicional en la que todos sus elementos son 1. Por tanto la matriz **X** queda como una matriz con n filas y (p+1) columnas, de la forma:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & \ldots & x_{1,p} \\ 1 & x_{2,1} & \ldots & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n-1,1} & \ldots & \ldots & x_{n-1,p} \\ 1 & x_{n,1} & \ldots & \ldots & x_{n,p} \end{pmatrix}$$

Y por último un conjunto de coeficientes de regresión **β**, uno para cada variable de explicación, incluida la variable creada para la constante β_0, con 1 columna y (p+1) filas.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Si derivamos (4.09) para cada uno de los parámetros betas:

$$\frac{\ln L\,(\boldsymbol{\beta} \mid y_1, y_2, \ldots, y_n)}{\partial \beta_1} = x_{1i}(\boldsymbol{X^T\beta} - \boldsymbol{\pi})$$

$$\frac{\ln L\,(\boldsymbol{\beta} \mid y_1, y_2, \ldots, y_n)}{\partial \beta_2} = x_{2i}(\boldsymbol{X^T\beta} - \boldsymbol{\pi})$$

$$\vdots$$

$$\frac{\ln L\,(\boldsymbol{\beta} \mid y_1, y_2, \ldots, y_n)}{\partial \beta_p} = x_{p_i}(\boldsymbol{X^T\beta} - \boldsymbol{\pi})$$

Como se aprecia en cada una de las derivadas parciales de cada parámetro en $\boldsymbol{\beta}$, se observa que cada β_i se encuentra implícito en la ecuación correspondiente por lo que se concluye que no se obtiene una respuesta directa, recurriéndose, como ya lo anunciáramos, a métodos numéricos que calculan el valor de las raíces en ecuaciones implícitas. En el Capítulo 3 se menciono el método de Newton Raphson, el cual da solución numérica al problema.

Para poder aplicar el método de Newton Raphson falta calcular la matriz Hessiana, la cual se obtiene de derivar el vector de las derivadas parciales de β, que matricialmente se escribe:

$$\boldsymbol{U(\beta)} = \frac{\boldsymbol{\partial LL(\beta)}}{\boldsymbol{\partial \beta}} = \boldsymbol{X^T} \cdot (\boldsymbol{Y} - \boldsymbol{\pi}) \qquad (4.10)$$

Y al derivar por segunda vez la función de verosimilitud se encuentra la matriz Hessiana $H(\boldsymbol{\beta})$, que se denota y define como:

$$\boldsymbol{H(\beta)} = \left[\left[\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta \partial \beta'}\right]\right] = -\boldsymbol{X^T} \cdot \boldsymbol{W} \cdot \boldsymbol{X} \qquad (4.11)$$

Siendo $\boldsymbol{W}$ una matriz diagonal, que queda de la forma siguiente:

$$\boldsymbol{W} = \begin{bmatrix} \pi_1(1-\pi_1) & 0 & \cdots & 0 \\ 0 & \pi_2(1-\pi_2) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n(1-\pi_n) \end{bmatrix} \qquad (4.12)$$

Donde,

$$\pi_i = \frac{1}{1 + e^{-\sum_{j=1}^{n+1} \beta_j x_{ij}}}, \qquad i = 1, 2, \ldots n \qquad (4.13)$$

Luego de obtener las derivadas de la función de verosimilitud de la ecuación de regresión, se llega a concluir que para las iteraciones se presenta lo siguiente:

$$\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}}_{t-1} + (\boldsymbol{X}^T \cdot \boldsymbol{W}_{t-1} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^T \cdot (\boldsymbol{Y} - \pi_t) \qquad (4.14)$$

Desde este punto, empiezan los cálculos iterativos, que dada su complejidad, es necesario un programa computacional, por lo que se ha desarrollado el software estadístico ERLA para que ingrese los datos y se obtengan los resultados correspondientes de una manera fácil y rápida.

Se puede ilustrar este método dando un ejemplo, se toma el caso de la creación de un nuevo insecticida para combatir escarabajos en las manzanas, el estudio consistió en la cantidad X de insecticida en miligramos disueltos en un litro de agua y la cantidad de escarabajos, cada solución logra matar; como se muestra en la Tabla 4.01:

Tabla 4.01: Ejemplo-Insecticida-Distribución Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Dosis	Número de insectos	Número de muertos	Probabilidades
1.6907	59	6	0.10
1.7242	60	13	0.21
1.7552	62	18	0.29
1.7842	56	28	0.5
1.8113	63	52	0.82
1.8369	59	53	0.89
1.8610	62	61	0.98
1.8839	60	60	1

Autoría: Fuentes A., Pinos R., Rivera N.

Tomamos como variable X la dosis de insecticida, y la variable Y los éxitos y fracasos para cada dosis, esto es, Ingresamos 59 observaciones con x=1.6907, donde 6 serán Y=1 y 53 serán Y=0, y así con las siguientes observaciones; de esta manera podemos ingresar los datos al programa, generando el siguiente modelo

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}$$

Al ingresar los datos en el programa ERLA, se muestra el resultado final mas no el cálculo del método numérico de Newton-Raphson hace en las diferentes iteraciones, de tal manera que se ilustra la forma como converge las estimaciones del valor deseable de acuerdo al método numérico, como podemos observar en la Tabla 4.02:

Tabla 4.02: Iteraciones con el Método de Newton – Raphson, ejemplo insecticida
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Iteraciones	β_0	β_1	l(β)
0	0.000	0.000	-333.4038
1	-37.8564	21.3374	-200.0098
2	-53.8532	33.8442	-187.27
3	-59.9652	34.2648	-186.24
4	-60.7078	34.2703	-186.23
5	-60.7175	34.2703	-186.23
6	-60.7175	34.2703	-186.23
7	-60.7175	34.2703	-186.23
8	-60.7175	34.2703	-186.23
9	-60.7175	34.2703	-186.23
10	-60.7175	34.2703	-186.23

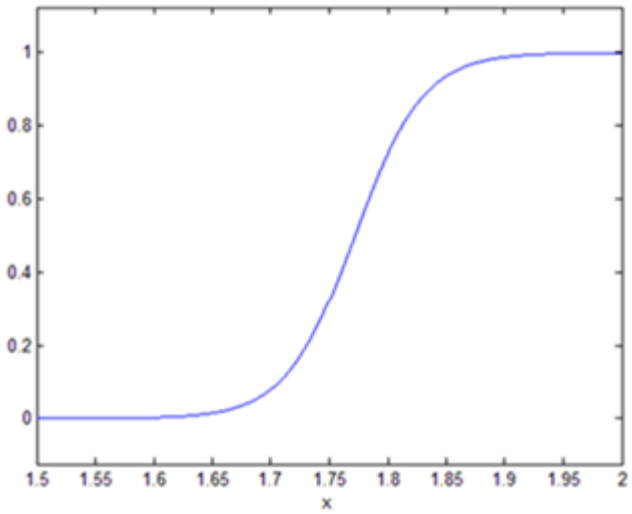
Autoría: Fuentes A., Pinos R., Rivera N.

Se puede observar que los valores de β_0 y β_1 se estabilizan en la quinta iteración, luego de ésta, los valores no cambian con una precisión de 4 decimales, por lo que podemos tomar los valores de los estimadores de beta como $\beta_0 = -60.7175 \; y \; \beta_1 = 34.2703$

Si graficamos la función de la distribución que tienen la probabilidad del insecticida de matar a los escarabajos en función de los miligramos del compuesto, con los estimadores de betas calculados, obtenemos el Gráfico 4.02:

Gráfico 4.02: Modelo de Regresión Logística – Estimaciones de β_0 y β_1
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

$$\frac{1}{1+e^{-(-60.7175+34.2703)}}$$



Autoría: Fuentes A., Pinos R., Rivera N.

4.2.2 Evaluación de los Modelos de la Regresión Logística

El siguiente paso será comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo. Para ello podemos emplear dos métodos, el del Estadístico de Wald y el del Estadístico G de Verosimilitud:

1. **El estadístico de Wald**. Se utiliza el denominado estadístico W de Wald que se define como:

$$W = \frac{\hat{\beta}_j^{\,2}}{var(\hat{\beta}_j^{\,2})}$$

Que tiene aproximadamente una distribución χ^2 con $(p-1)$ grados de libertad.

Para el caso multivariado, $\boldsymbol{W}$ se lo expresa como la expresión matricial:

$$\boldsymbol{W} = \widehat{\boldsymbol{\beta}}'^{[\Sigma(\hat{\beta})]^{-1}\hat{\beta}} = \widehat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})\widehat{\boldsymbol{\beta}} \qquad (4.15)$$

Se hace el siguiente contraste de hipótesis:

$$\boldsymbol{H_0}: \beta_j = 0, \qquad j = 1,2,..p$$

Vs.

$$\boldsymbol{H_1}: \beta_j \neq 0, \qquad j = 1,2,..p$$

Como regla general rechazamos $\boldsymbol{H_0}$ a favor de $\boldsymbol{H_1}$ si el nivel de significancia de la muestra (valor p) es menor que 0.05, definiendo la Región Critica como: con $(1-\alpha)100\%$ de confianza se rechaza $\boldsymbol{H_0}$ a favor de $\boldsymbol{H_1}$ si W > $\chi^2{}_{1-\alpha}(p-1)$.

2. **El estadístico G de la razón de verosimilitud.**

Otra opción para verificar estadísticamente el valor de los parámetros $\beta_1, \beta_2, \ldots$ es utilizar el denominado estadístico G de la Razón de Verosimilitud, cuya definición se bosqueja a continuación:

Se trata de comparar el modelo que resulta de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una distribución χ^2con 1 grado de libertad (no se supone normalidad). La ausencia de significación implica que el modelo sin la covariable eliminada no desmejora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no es significativa en el mismo. Esta prueba no supone distribución alguna, por lo que es la más recomendada.

Es más una método de “prueba y error”, que compara diferentes modelos donde se sustituyen las variables que se emplean, por lo que en si no tiene un contraste de hipótesis.

4.3 Regresión Poisson

La Regresión Poisson es una técnica estadística en lo que se utiliza un modelo no lineal que pertenece a la categoría del análisis de datos de recuento. En estos casos, la variable dependiente toma más de dos valores discretos dígase: 0,1,2,3…, no negativos.

A igual que el capítulo anterior partimos de una "Función de Enlace" para la Regresión Poisson.

$$\mu = e^{\beta_0+\beta_1 x_1+\cdots+\beta_{p-1}x_{p-1}} \tag{4.16}$$

Siguiendo a Greence (1999), se tiene que y_i es la realización de una variable aleatoria Y_i, que sigue una distribución de Poisson, con parámetros λ_i, que está relacionada con las variables explicativas X. Así, $\text{Prob}(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$, donde y_i= 0,1,2…, al tiempo queλ_i = exp($\boldsymbol{\beta}^{\boldsymbol{T}} x_i$), y por lo tanto, ln λ_i= $\boldsymbol{\beta}^{\boldsymbol{T}} x_i$

Una característica de este tipo de distribución es:

$$E(y_i|x_i) = \text{Var}(y_i|x_i) = \lambda_i$$

Y sus efectos marginales, al igual que pasaba en el modelo de regresión logística depende de los valores de las variables explicativas, ya que:

$$\frac{\partial E(y_i|x_i)}{\partial x_i} = \lambda_i \boldsymbol{\beta} \tag{4.17}$$

4.3.1 Los Modelos de Regresión de Poisson

Siendo y Poisson, la variable dependiente a explicar es, por tanto, una variable discreta ordinal.

Ejemplos:

El número de llamadas que recibe una central telefónica en una hora.

El número de accidentes que sufre un conductor durante un año.

El número de veces que un cliente compra una misma marca en un año.

4.3.2 Interpretación de los Parámetros

El incremento esperado en el parámetro λ_i cuando X_j cambia una unidad es:

$$\frac{\partial E[y_i|\boldsymbol{X}_i]}{\partial X_j} = e^{\boldsymbol{X}_i\boldsymbol{\beta}}\beta_j = \lambda_i\beta_j \qquad (4.18)$$

Cuando se dispongan de estimaciones de los parámetros este valor se puede calcular para cualquier vector de datos **X.**

En la práctica es habitual realizar únicamente interpretaciones del signo de los parámetros estimados, que indica la dirección en que se mueve el valor de λ_i cuando aumenta la variable explicativa correspondiente X_j.

4.3.3 Estimación De los parámetros

El método, ya varias veces utilizado en este trabajo, es el de **Máxima Verosimilitud**. La función de verosimilitud de obtiene a partir de:

$$L = \prod_{i=1}^{n} P\left[Y_i \middle| X = x_i\right] = \prod_{i=1}^{n} \frac{\lambda_i^{\,y_i}}{y_i!} e^{-\lambda_i} \tag{4.19}$$

Donde tomando logaritmos:

$$ln\, L = \sum_{i=1}^{n} [y_i \ln \lambda_i - \ln y_i ! - \lambda_i] \tag{4.20}$$

Sustituyendo $\ln \lambda_i$ por el modelo logarítmico-lineal tenemos:

$$\ln L = \sum_{i=1}^{n} \left[y_i \left(\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} \right) - \ln y_i ! - \lambda_i \right] \tag{4.21}$$

Al igual que en Regresión Logística, al derivar, se obtiene un sistema de ecuaciones implícitas, el cual no tienen solución explicita, por ello se utiliza el método numérico de Newton Raphson, como ya se explicó anteriormente, se muestra el cálculo directamente.

Como el vector de las primeras derivadas de $g(\beta^*) = 0$ en el optimo β^*. Dejando en términos de β^* nos lleva a:

$$\beta^* \cong \beta - [G(\beta^*)] - \frac{1}{2} g(\beta) \tag{4.22}$$

Para estos se aplicara el método Newton-Raphson para varias variables como se vio en la sección 3.4, utilizando la ecuación 3.21. Para poder aplicar el método falta de calcular la Matriz Hessiana, la cual como se indicara

posteriormente, se obtiene de derivar el vector de las derivadas parciales de β, que matricialmente se escribe:

$$\boldsymbol{U(\beta)} = \left[\frac{\partial LL(\beta)}{\partial \beta}\right] = \boldsymbol{X^T} \cdot (\boldsymbol{Y} - \boldsymbol{\pi}) \qquad (3.23)$$

Y al derivar por segunda vez la función de verosimilitud se encuentra la matriz Hessiana, que se escribe:

$$\boldsymbol{H(\beta)} = \left[\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'}\right] = -\boldsymbol{X^T} \cdot \boldsymbol{W} \cdot \boldsymbol{X}$$

Siendo:

$$\boldsymbol{W} = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_n \end{bmatrix}$$

$$\boldsymbol{\pi_i} = \mathbf{e}^{\sum_{j=1}^{n+1} \beta_j X_{ij}} \qquad i = 1, 2, \ldots n \qquad (4.24)$$

Luego de obtener las derivadas de la función de verosimilitud de la ecuación de regresión, se concluye que para la i-esima iteración que:

$$\widehat{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}}_{t-1} + (\boldsymbol{X}^T \cdot \boldsymbol{W}_{t-1} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^T \cdot (\boldsymbol{Y} - \boldsymbol{\pi}) \qquad (4.25)$$

Con esto construye la Regresión de Poisson con el software estadístico ERLA para que ingrese los datos y se obtengan los resultados correspondientes.

4.3.4 Evaluación de los modelos de Poisson

Para la evaluación del modelo de regresión de Poisson se realiza la prueba de estadístico de Wald la cual consiste en la estimación de los parámetros del θ se compara con el valor propuesto θ_0, con la diferencia entre los dos estará aproximadamente normal. El cuadrado de la diferencia se compara típicamente a distribución ji-ajustada.

1. **El estadístico de Wald**. Se utiliza el denominado estadístico W de Wald que se define como:

$$W = \frac{\hat{\beta}_j^{\,2}}{var(\hat{\beta}_j^{\,2})}$$

Que tiene aproximadamente una distribución χ^2 con $p-1$ grados de libertad.

Para el caso multivariado W se lo expresa como la expresión matricial:

$$W = \hat{\beta}'^{[\Sigma(\beta)]^{-1}\hat{\beta}} = \hat{\beta}'(X'VX)\hat{\beta} \qquad (4.26)$$

Para los fines pertinentes, se propone el siguiente contraste de hipótesis:

$$\boldsymbol{H_0}: \beta_j = 0, \qquad j = 1,2,..p$$

Vs.

$$\boldsymbol{H_1}: \beta_j \neq 0, \qquad j = 1,2,..p$$

Con $(1-\alpha)100\%$ de confianza se rechaza $\boldsymbol{H_0}$ a favor de $\boldsymbol{H_1}$ si:

$W > \chi^2{}_{\alpha}(p-1)$, o en situaciones post experimentales, si el nivel de significancia del la muestra (valor p) es menor a 0.01

2. **El estadístico G de la razón de verosimilitud.**

Otra opción para verificar estadísticamente el valor de los parámetros $\beta_1, \beta_2, \ldots$ es utilizar el denominado estadístico G de la Razón de Verosimilitud, que se lo define de la siguiente manera:

Como se indicó en líneas previas, trata de comparar cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. En este caso cada estadístico G sigue una distribución χ^2con 1 grado de libertad (no se supone normalidad). La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no es significativa en el mismo. Esta prueba no supone ninguna distribución alguna, por lo que es la más recomendada.

4.3.5 Regresión Poisson con ERLA

Ilustramos los redultados del software diseñado, con un ejemplo basado en datos ecuatorianos, relacionados con el éxito de apareamiento de caballos de acuerdo a su edad***,*** los datos corresponden a la hacienda Glorieta,

ubicada en el Km. 58 vía a Guayaquil - Salinas. Veremos cómo se ejecuta dentro de ERLA con Regresión Poisson, basándonos en la teoría descrita. Con los datos proporcionados en la Tabla 4.03

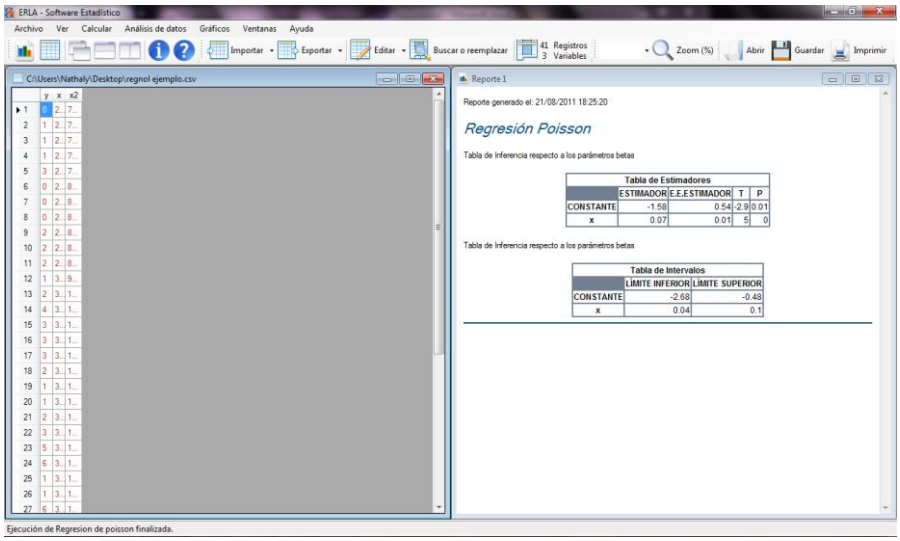
Tabla 4.03: Ejemplo Reproducción-caballos-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Caballo	Edad	Numero de éxitos		Caballo	Edad	Numero de éxitos
1	3	0		20	10	1
2	4	1		21	10	2
3	4	1		22	10	3
4	4	1		23	12	5
5	4	3		24	12	6
6	5	0		25	13	1
7	5	0		26	13	1
8	5	0		27	13	6
9	5	2		28	14	2
10	5	2		29	15	1
11	5	2		30	17	3
12	6	1		31	18	4
13	8	2		32	19	0
14	9	4		33	19	2
15	9	3		34	19	3
16	9	3		35	19	4
17	9	3		36	19	9
18	9	2		37	20	3
19	10	1		38	21	5

Autoría: Fuentes A., Pinos R., Rivera N.

Para el ingreso de los datos puede revisarse el manual de usuario, donde se describe paso a paso el uso del software estadístico.

Ilustración4.01: Éxito de apareamiento de los caballos ERLA-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

La ilustración 4.01, observamos la ventana de ERLA donde al ejecutar el ejemplo antes mencionado nos devuelve los estimadores de los betas y los intervalos de confianza.

Bajo el modelo de $Y = \beta_0 + \beta_1 X$, ya que solo tenemos una variable de explicación y una variable a ser explicada, claro está, que podríamos agregar una segunda variable de explicación y hacer un modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ pero bajo las condiciones actuales tenemos:

$$\beta_0 = -1.58, \ \ \beta_1 = 0.07$$

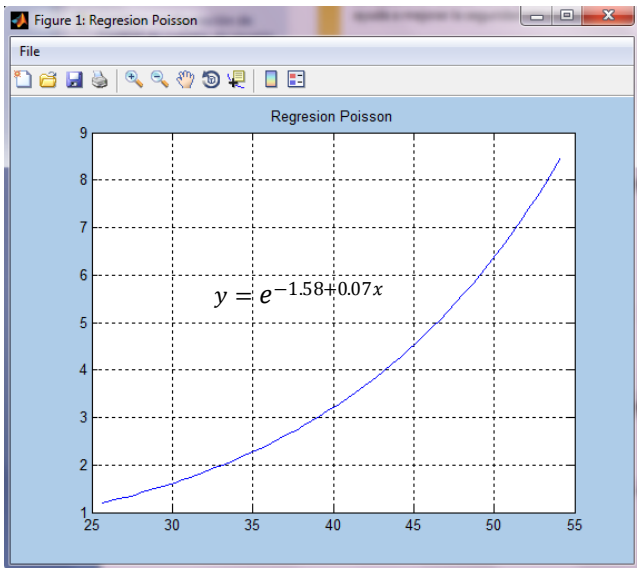
En la Tabla 4.04 aparecen las cotas superiores e inferiores para los intervalos para $\beta_0 \ y \ \beta_1$, con 95% de confianza:

Tabla 4.04: Intervalos de confianza de los Betas (con 95% de confianza)
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

betas	Límite Inferior	Límite Superior
β_0	- 2.68	-0.48
β_1	0.04	0.1

En la ilustración 4.02 se puede apreciar el grafico que también se genera después de mostrar los valores de los estimadores de beta.

Ilustración4.02: Gráfico del éxito de apareamiento de los elefantes-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

CAPITULO V

5. PROGRAMACIÓN Y VALIDACION

5.1Introducción

En este Capítulo presentamos los algoritmos creados específicamente para los módulos de Regresión Logística y Poisson, con sustento teórico en los Capítulo IV y V además se realizara la validación de los Modelos ya mencionados, estableciendo los valores de los parámetros betas y añadiendo una variable que será $\mathrm{N}\sim(0,\sigma^2)$.

5.2Regresión Logística

5.2.1 Validación del Modelo de Regresión Logística

De acuerdo al modelo de Regresión Logística la función de "enlace" es:

$\pi(\boldsymbol{x_i}) = \frac{1}{1+e^{(x_i^T\beta)}}$ dónde:

$$\boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip-1}$$

$$\pi(\boldsymbol{x_i}) = \frac{1}{1+e^{(\beta_0+\beta_1 x)}} \qquad (5.01)$$

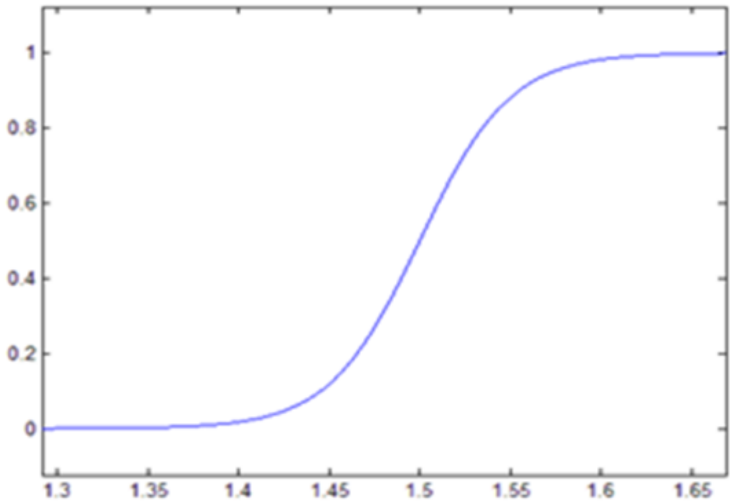
Se establece los valores para el modelo inicial con:

$$\beta_0 = 60,\ \beta_1 = 40;$$

Con lo que obtenemos $Y = \frac{1}{(1+e^{(60-40x)})}$.

La grafica presentada a este modelo determinístico es:

Gráfico 5.01: Modelo determinístico de Regresión Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Tomando para los valores de x desde x=1,4 hasta x=1,6,

Los Errores que se agregaron al modelo determinístico para darle variabilidad fueron de diferentes tipos para simular lo que se encuentra en la realidad:

$$\varepsilon \sim N(0\ ,0.01) \qquad \varepsilon \sim N(0\ ,0.005) \qquad \varepsilon \sim N(0\ ,0.015)$$

Para cada muestra que se hizo, se tomaron 11 valores del error y agregarle a cada una de las agrupaciones de datos con las que estamos trabajando para este ejemplo, cada agrupación consta de 100 observaciones (esto es, que para cada agrupación hay p datos que son uno, y 100-p datos que son cero), ya que recordemos que estamos calculando probabilidad y tenemos que ponerlos en datos numéricos.

Haciendo uso de las agrupaciones, *una primera réplica* que se realizó con $\boldsymbol{\varepsilon \sim N(0\ ,\ 0.01)}$, resulto:

Tabla 5: Primera réplica de la Validación del Modelo con $\boldsymbol{\varepsilon \sim N(0\ ,\ 0.01)}$,
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

X (Agrupacion)	$\boldsymbol{Y = \frac{1}{(1+e^{(60-40x)})}}$	***Error~*** $\boldsymbol{N(0,\sigma^2)}$	$\boldsymbol{Y+\varepsilon \sim N(0,\sigma^2)\ (Probabilidad: p)}$	***Y toma valor de 1*** $\boldsymbol{= p * n}$	***Y toma valores de 0*** $\boldsymbol{=(p-1)*n}$
1,40	0,0180	-0,0096	0,0084	1	99
1,42	0,0392	-0,0078	0,0314	3	97
1,44	0,0832	-0,0106	0,0726	7	93
1,46	0,1680	0,0050	0,1730	17	83
1,48	0,3100	-0,0039	0,3062	31	69
1,5	0,5000	0,0276	0,5276	53	47
1,52	0,6900	-0,0280	0,6619	66	34
1,54	0,8320	-0,0008	0,8313	83	17
1,56	0,9168	0,0084	0,9253	93	7
1,58	0,9608	-0,0170	0,9439	94	6
1,60	0,9820	-0,0054	0,9766	98	2

Autoría: Fuentes A., Pinos R., Rivera N.

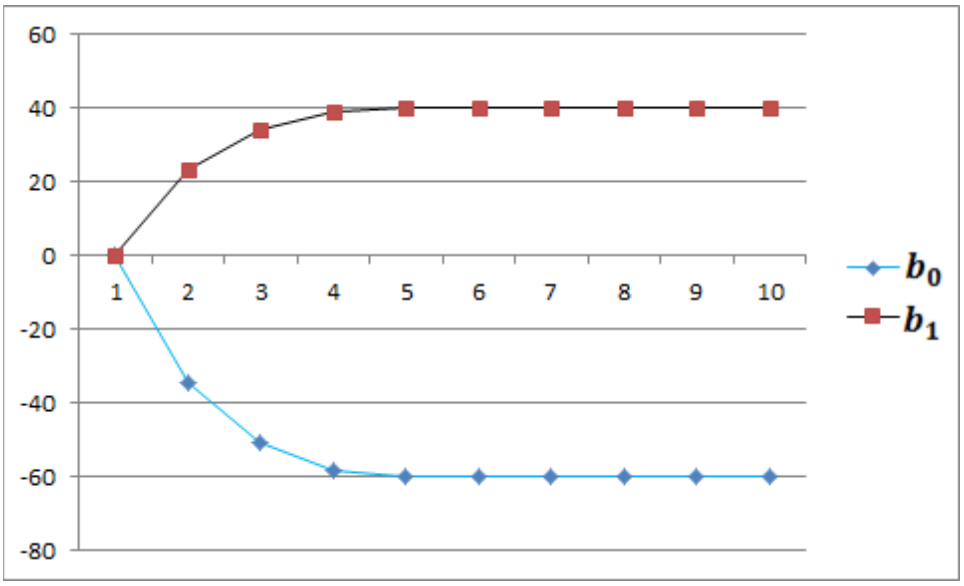
Ingresando los datos de la primera réplica al programa de la regresión logística nos da los siguientes betas estimados:

$$b_0 = -60.67$$

$$b_0 = 39.94$$

La forma en cómo el programa calcula los betas es de forma iterativa, esto es, calcula un beta y luego según este calcula uno mejor, y asi hasta que la diferencia este dentro de los parámetros aceptados, como va evolucionando el valor de los betas desde su valor inicial cero, se lo puede observar en el Gráfico 5.02.

Gráfico 5.02: Comportamiento de los Betas Estimados
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Así los betas estimados de las 10 diferentes iteraciones las podemos observar en la Tabla 6.

Tabla 5.01: Betas estimados-Regresión Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Iteraciones	b_0	b_1
1	0	0
2	-34,67	23,11
3	-51,08	34,04
4	-58,62	39,07
5	-59,90	39,92
6	-59,93	39,94
7	-59,93	39,94
8	-59,93	39,94
9	-59,93	39,94
10	-59,93	39,94

Autoría: Fuentes A., Pinos R., Rivera N.

Se puede apreciar, a partir de la 6ta iteración, los valores de los betas se han estabilizado en un valor concreto, que no cambia con las siguientes iteraciones.

A pesar de utilizar diferentes errores ($\varepsilon \sim N(0\,, 0.01)$, $\varepsilon \sim N(0\,, 0.005), \varepsilon \sim N(0\,, 0.015)$) podemos ver que nuestro programa nos han generado modelos con betas que convergen a los valores de $\beta_0\ y\ \beta_1$ determinados en un inicio

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -60 \\ +40 \end{bmatrix}$$

5.2.2 PROGRAMACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA

Se ha realizado una función en Matlab Reglogcontr.m, la cual toma las variables a ser explicada y la(s) variable(s) de explicación, el cual recibe

valores de “y” y “x” donde “y” representa el vector de la variable a ser explicada y “x” es la matriz que contiene a la variables de explicación del modelo que permita el cálculo de los valores estimados de los betas y además los intervalos de confianza.

Cuadro 1: Programación para los estimadores de los Betas-Regresión Logística
“Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson”

```
function R1 = reglogcontr(y,x,b0)
[n,ppp]=size(x);
beta=b0;
dife=1;
pp=zeros(1,n);
w=zeros(n);
x=[ones(n,1),x];
whiledife>0.0001
bini=beta;
for i=1:n
suma=x(i,:)*beta;
pp(i)=1/(1+exp(-suma));
end
    p=pp';
for i=1:n
w(i,i)=p(i)*(1-p(i));
end
beta=bini+(inv(x'*w*x))*x'*(y-p);
dife=sum(abs(beta-bini));
end
Sb=inv(x'*w*x);
R1=zeros(ppp,4);
for i=1:ppp+1
R1(i,1)=beta(i);
R1(i,2)=sqrt(Sb(i,i));
R1(i,3)=R1(i,1)/R1(i,2);
R1(i,4)=abs(R1(i,3));
R1(i,4)=tcdf(R1(i,4),n-ppp);
R1(i,4)=(1-R1(i,4))*2;
end
```

Autoría: Fuentes A., Pinos R., Rivera N.

Al ejecutar esta programación desde ERLA, los valores que se muestran son el Estimador de los Betas, el Error estimado del estimador, el estadístico de Prueba T de Student y el “valor P” en el siguiente formato, utilizaremos el ejemplo anterior, el de la Tabla 5, los resultados están en la Tabla 7.

Tabla 5.02: Tabla de Estimadores-Regresión Logística
“Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson”

beta	Estimador	E.E. Estimador	T	P
β_0	-59.93	5.1807	-11.7179	0.00
β_1	39.94	2.9121	11.7681	0.00

Todo esto se base en un modelo de:

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

La función en Matlab Reglogbeta.m, es la programación para determinar los intervalos de confianza de los betas, la programación la podemos hallar en el Cuadro 2:

Cuadro 2: Programación para los Intervalos de confianza para b_0 y b_1-Regresión Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

```
function B = reglogbeta(y,x,b0)
[n,ppp]=size(x);
beta=b0;
dife=1;
pp=zeros(1,n);
w=zeros(n);
x=[ones(n,1),x];
whiledife>0.0001
bini=beta;
for i=1:n
suma=x(i,:)*beta;
pp(i)=1/(1+exp(-suma));
end
    p=pp';
for i=1:n
w(i,i)=p(i)*(1-p(i));
end
beta=bini+(inv(x'*w*x))*x'*(y-p);
dife=sum(abs(beta-bini));
end
Sb=inv(x'*w*x);
B=zeros(ppp,2);
for be=1:ppp+1
vbeta=sqrt(Sb(be,be));
%conf=input('ingrese   el   valor   de
alpha: ');
conf=0.975;
tt=TINV(conf,n-ppp);
%el calculo de la T con la confianza y
el n-p
B(be,1)=beta(be)-vbeta*tt;
B(be,2)=beta(be)+vbeta*tt;
end
```

Siguiendo con el mismo ejercicio, al ingresar en el software ERLA, se muestra la Tabla 8 con los valores de los intervalos de confianza para los Betas.

Tabla 5.03: Intervalos de los Betas-Regresión Logística
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Beta	Límite Inferior	Límite Superior
β_0	-70.8971	-50.5378
β_1	28.5482	39.9924

5.3 Regresión Poisson

5.3.1 Validación del Modelo de Regresión Poisson

De acuerdo al modelo de regresión Poisson de probabilidades $\pi(\boldsymbol{x_i}) = e^{(x_i'\beta)}$ donde:

$$\boldsymbol{X^T\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip-1}$$

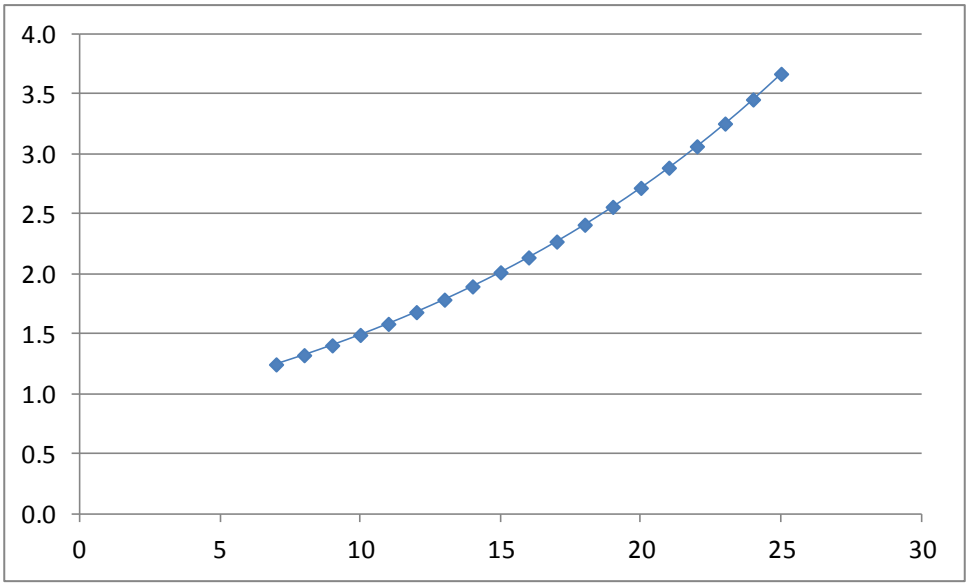
$$\pi(\boldsymbol{x_i}) = e^{(\beta_0 + \beta_1 x)}$$

Se establece los valores para el modelo inicial con

$\beta_0 = -0.2,$

$\beta_1 = 0.06;$ Entonces $Y = e^{(-0.2+0.06*x)}$.

Gráfico 5.03: Modelo determinístico, Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

Tomando valores de X desde 7 hasta 25,

El error incluido ϵ a la muestra para simular aleatoriedad tiene distribución:

$$\epsilon \sim N(0\ ,1)$$

A continuación la Tabla 9 representa como se obtuvo los datos a ingresar en el software ERLA:

Tabla 5.04: Muestra-Modelo determinístico-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

X	$Y = e^{(-1.5+0.06*x)}$	$Error \sim N(0,1)$	$Y + \varepsilon \sim N(0,1)$	Valor entero
7	1,24607673	1,059	2,30507673	2
8	1,32312981	-0,7575	0,56562981	0
9	1,40494759	-0,15595	1,24899759	1
10	1,4918247	0,93402	2,4258447	2
11	1,58407398	-0,99819	0,58588398	0
12	1,68202765	1,57244	3,25446765	3
13	1,78603843	-1,06016	0,72587843	0
14	1,89648088	-0,88481	1,01167088	1
15	2,01375271	-1,02125	0,99250271	0
16	2,13827622	-1,13474	1,00353622	1
17	2,27049984	0,58773	2,85822984	2
18	2,41089971	-0,66836	1,74253971	1
19	2,55998142	-0,28647	2,27351142	2
20	2,71828183	0,56757	3,28585183	3
21	2,88637099	-1,36348	1,52289099	1
22	3,0648542	0,34913	3,4139842	3
23	3,2543742	0,40724	3,6616142	3
24	3,45561346	-0,09489	3,36072346	3
25	3,66929667	-0,08449	3,58480667	3

Autoría: Fuentes A., Pinos R., Rivera N.

La última columna es el valor entero de la suma entre el valor calculado y el error, ya que recordemos que la variable Y se caracteriza por estar conformada por números enteros.

Al ingresar estos datos en el programa obtenemos los siguientes estimadores de betas: $b_0 = -0.7614 \quad y \quad b_1 = 0.0732$

Los betas estimados de las 10 réplicas las podemos observar en la Tabla 12

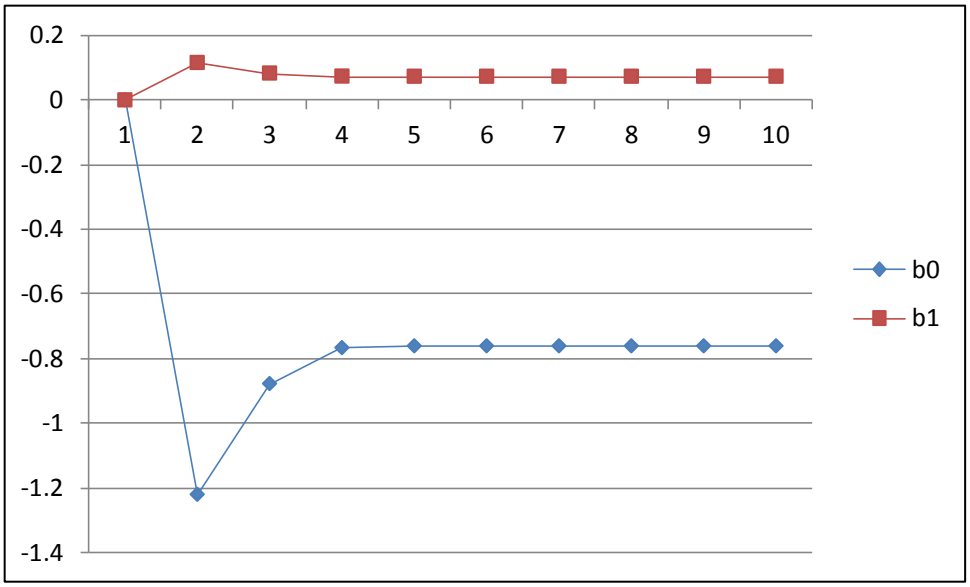
Tabla 5.05: Réplicas Betas Estimados-Modelo determinístico, Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

Iteracion	b_0	b_1
1	0	0
2	-1.2211	0.1158
3	-0.8783	0.0831
4	-0.7670	0.0737
5	-0.7614	0.0732
6	-0.7614	0.0732
7	-0.7614	0.0732
8	-0.7614	0.0732
9	-0.7614	0.0732
10	-0.7614	0.0732

Autoría: Fuentes A., Pinos R., Rivera N.

A partir de la 6ta iteración las estimaciones tienden a los betas inicialmente supuestos en el planteamiento del modelo, podemos apreciar su convergencia en el Grafico 5.04

Gráfico 5.04: Comportamiento estimado de los betas-Validación Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"



Autoría: Fuentes A., Pinos R., Rivera N.

A pesar de utilizar diferentes errores que le agregamos a la muestra podemos ver que el programa genera estimadores de los parámetros (betas) que convergen a los valores inicialmente propuestos, esto es:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -0.2 \\ +0.06 \end{bmatrix}$$

5.3.2 Programación del Modelo de Regresión Poisson

De igual manera como la programación de Regresión logística, se desarrolló la función Regpoicontr, Cuadro 3, para estimar los parámetros betas,

Cuadro 3: Programación para los estimadores de los Betas-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

```
function R1=regpoicontr(y,x,b0)
[n,ppp]=size(x);
beta=b0;
dife=1;
pp=zeros(1,n);
w=zeros(n);
x=[ones(n,1),x];
whiledife>0.0001
bini=beta;
for i=1:n
suma=x(i,:)*beta;
pp(i)=exp(suma);
end
    p=pp';
for i=1:n
w(i,i)=p(i);
end
beta=bini+(inv(x'*w*x))*x'*(y-p);
dife=sum(abs(beta-bini));
end
Sb=inv(x'*w*x);
R1=zeros(ppp,4);
for i=1:ppp+1
R1(i,1)=beta(i);
R1(i,2)=sqrt(Sb(i,i));
R1(i,3)=R1(i,1)/R1(i,2);
R1(i,4)=abs(R1(i,3));
R1(i,4)=tcdf(R1(i,4),n-ppp);
R1(i,4)=(1-R1(i,4))*2;
end
```

Autoría: Fuentes A., Pinos R., Rivera N.

La programación se ejecuta bajo el modelo:

$$y = e^{(\beta_0 - \beta_1 x)}$$

Al ejecutar la programación en el ejemplo determinístico visto recientemente, se presenta la tabla de estimadores de los parámetros betas, con el Modelo de Regresión Poisson, junto al error estándar de cada beta, el valor T de student y el valor p, para poder comprobar si el beta es significativo o no. (Tabla 11).

Tabla 5.06: Tabla de Estimadores- Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

betas	Estimador	E.E. Estimador	T	P
β_0	-0.7614	0.6486	-1.1739	0.2557
β_1	0.0732	0.0344	2.1303	0.0472

Autoría: Fuentes A., Pinos R., Rivera N.

Para encontrar los intervalos de confianza de cada parámetro beta se desarrolló la función Regpoibeta, visto en el Cuadro 4

Cuadro 4: Programación para los Intervalos de confianza-Regresión Poisson
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

```
function B=regpoibeta(y,x,b0)
[n,ppp]=size(x);
beta=b0;
dife=1;
pp=zeros(1,n);
w=zeros(n);
x=[ones(n,1),x];
whiledife>0.0001
bini=beta;
for i=1:n
suma=x(i,:)*beta;
pp(i)=exp(suma);
end
    p=pp';
for i=1:n
w(i,i)=p(i);
end
beta=bini+(inv(x'*w*x))*x'*(y-p);
dife=sum(abs(beta-bini));
end
Sb=inv(x'*w*x);
B=zeros(ppp,2);
for be=1:ppp+1
vbeta=sqrt(Sb(be,be));
%conf=input('ingrese el valor de
alpha: ');
conf=0.975;
tt=tinv(conf,n-ppp);
%el calculo de la T con la
confianza y el n-p
B(be,1)=beta(be)-vbeta*tt;
B(be,2)=beta(be)+vbeta*tt;
end
```

Autoría: Fuentes A., Pinos R., Rivera N.

Al ejecutar la función Regpoibeta con ERLA con los mismos datos que se utilizó anteriormente, obtenemos los intervalos de confianza que podemos ver en la Tabla 12, igual que antes, estos son calculados con un 95% de confianza.

Tabla 5.07: Intervalos de los Betas -Regresión Poisson-ERLA
"Software Estadístico para Regresión. El caso de Regresión Logística y Regresión Poisson"

betas	Límite Inferior	Límite Superior
β_0	- 2.1240	0.6012
β_1	0.0010	0.1455

Autoría: Fuentes A.. Pinos R.. Rivera N.

Conclusiones y Recomendaciones

El desarrollo del presente Proyecto de Materia de graduación ha permitido obtener las siguientes conclusiones y recomendaciones:

Conclusiones

- Se ha obtenido un software libre estadístico, aplicado al pre-grado de la carrera de Estadística informática sobre programación y estadística, que ayudará a que los usuarios hacer más factible obtener resultados.

- Se ha integrado, métodos, funciones y herramientas de ingeniería de software en el desarrollo del software estadístico.

- El sistema informático desarrollado ERLA permite tener un manejo y análisis de datos para tomar decisiones

- El sistema informático desarrollado ERLA permite desarrollar estadística descriptiva, inferencial y mutivariada clara y concisa.

- Se ha integrado armónicamente la tecno ciencia, en este caso la ingeniería de software y la informática, una combinación entre tecnología y educación.

Recomendaciones

- Se recomienda dar actualizaciones en el software para el análisis de datos para tomas de decisiones.

- Recomiendo que el presente software sirva de base para la realización de otros software que permiten realizar más técnicas estadísticas multivariadas.

- Se recomienda el trabajo disciplinario para la consecución de este tipo de proyectos, para que el software de igual forma disciplinario.

- El uso de Matlab y Visual Studio 2011 es una buena opción para la realización de software libre por su versatilidad y entorno amigable que presenta.

BIBLIOGRAFIA

[1] ***Abraham, B. y Ledolter, J.*** (2006), Introduction to Regression Modeling, Editorial Thomson Book/Cole.

[2] ***Cassella, G y Berger, R.*** (2002), Statistical Inference, Segunda Edición, Editorial Thomson Book/Cole 2002.

[3] ***Freeman, H.*** (1979), Introducción a la inferencia estadística, Instituto Tecnológico de Massachusetts, Editorial Trillas México.

[4] ***http://www.monografias.com/trabajos27/regresion-simple/regresion-simple.shtml,*** actualizado al 2005 y consultado a Enero del 2011

[5] ***http://www.scribd.com/doc/29771741/Regresion-multiple***. Actualizado el 4 de Diciembre del 2010 y consultado a Diciembre del 2010

[6] ***http://www.mathtools.net/MATLAB/Statistics/index.html*.** Actualizado a Marzo del 2010 y consultado a Junio del 2010.

[7]***http://www.maths.lth.se/matstat/stixbox/Contents.html.*** STIXBOX, Caja de Herramientas para Matlab, Versión 1.29, 10 de Mayo del 2000, consultado en Junio del 2010.

[8]***http://www.virtual.unal.edu.co/cursos/ciencias/2001091/html/capitulo_7/leccion-07-02.html***, Universidad Nacional de Colombia. Consultado Febrero del 2011.

[9] http:// **es.wikipedia.org.** Mínimos Cuadrados, Categoría: Optimización | Análisis de Regresión | Álgebra Lineal. Actualizado al 2010 y consultado en Julio del 2010.

[10] ***Montalvo, D.*** (2000), Tesis de Grado "Análisis estadístico de la producción arrocera en el Ecuador", Escuela Superior Politécnica del Litoral.

[11] ***Moral, I***. Modelos de Regresión: Lineal Simple y Regresión Logística, Capítulo 14.

[12]***Sosa, W***, Introducción a los Modelos de Regresión, Universidad de San Andrés, Argentina.

[13] ***Zurita, G.*** (2010) Probabilidad y Estadística, Fundamentos y Aplicaciones, Segunda Edición, Instituto de Ciencias Matemáticas ESPOL, Guayaquil, Ecuador.

[14] ***Andersen, E (September 1970).*** ≪Sufficiency and Exponential Families for Discrete Sample Spaces≫.Journal of the American Statistical Association 65 (331): pp. 1248‐1255.

[15]***Pitman, E. (1936).*** ≪Sufficient statistics and intrinsic accuracy≫.Proc. Camb. phil. Soc. 32: pp. 567‐579.

[16]***Darmois, G. (1935).*** ≪Sur les lois de probabilites a estimationexhaustive≫. C.R. Acad. sci. Paris 200: pp. 1265‐1266.

[17] ***Koopman, B (1936).*** ≪On distribution admitting a sufficient statistic≫.Trans. Amer. math. Soc. 39: pp. 399‐409