ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

"CONSTRUCCIÓN DE SOFTWARE PARA REGRESIÓN EL CASO DE SELECCIÓN DEL MODELO Y PRUEBAS DE HOMOCEDASTICIDAD"

INFORME DE MATERIA DE GRADUACIÓN

Previa a la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentada por:

Macías Cabrera Sindy Victoria

Pincay Chiquito César Alfonso

Guayaquil – Ecuador

2012

AGRADECIMIENTO

A Dios por la salud brindada para que cada día hayamos podido ver un nuevo amanecer.

A nuestros padres por su apoyo, confianza y fe constantes para el cumplimiento de todas nuestras mestas.

A M. Sc Gaudencio Zurita quien nos ha brindado los conocimientos necesarios para el desarrollo de este Informe.

DEDICATORIA

Dedicamos este informe a los estudiantes del ICM y todas aquellas personas que creen en la innovación, emprendimiento y nuevas propuestas de los jóvenes de nuestro país, y a los que con su apoyo y consejos ayudaron a la culminación del mismo.

TRIBUNAL DE GRADUACIÓN

M.Sc. Gaudencio Zurita

DIRECTOR DE TESIS

Ing. Vanessa Salazar

DELEGADO

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)

Sindy Victoria Macías Cabera

Cesar Alfonso Pincay Chiquito

RESUMEN

Como propuesta de graduación se estudio la técnica de regresión lineal en su totalidad definiéndola así como Regresión Lineal Avanzada, junto con esta investigación surgió la idea de construir un software especializado dividiéndolo en varios módulos de investigación para el desarrollo del mismo; técnica que viene en diversos software estadísticos pero a nivel superficial. “ERLA” (Estadística Regresión Lineal Avanzada) llamado así por los desarrolladores es un software completo con las características básicas y avanzadas de la técnica mencionada es un programa computacional con características profesionales y que permiten su fácil entendimiento, entre las cuales se pueden mencionar cuadros de dialogo, consejos como ayuda. Menú emergente para el manejo de resultados, etc.

El desarrollo de *“ERLA”* ha sido realizado en dos plataformas informáticas estas fueron Matlab R2010a y Visual Net 2008. Este “paquete” contiene desde estadística básica como Tablas de Frecuencias, Estadísticas Descriptivas hasta Regresión de Ridge, Regresión Logística, Selección de Modelos, Puntos de Influencia y más. Siendo los indicadores de calidad de Selección de Modelos la contribución específica que se detallará en este reporte.

En el primer capítulo se consideran los principales fundamentos teóricos de la técnica, *"Regresión Lineal Simple y Múltiple"*; entre ellos la estimación de parámetros por mínimos cuadrados y máxima verosimilitud, los supuestos que se debe considerar en el modelo, contrastes de hipótesis, elaboración de la tabla ANOVA. Además se explicará el Coeficiente de Determinación, los supuestos que deben cumplir las variables explicativas y de respuesta.

En el segundo capítulo se presenta como tema específico los o indicadores de calidad de modelos de regresión con su respectiva técnicas; que permiten determinar las posibles regresiones de un conjunto de variables explicativas $X_1, X_2, \ldots, X_{p-1}$, para una variable a ser explicada *Y*. Dichos indicadores son R^2, R^2_{aj}, Criterio de Akaike, estadístico C_p de Mallows y PRESS.

En el tercer capítulo se explica paso a paso el desarrollo de ERLA como se enlazan Visual y Matlab, las funciones a utilizar y un detalle de cada uno de estos dos programas indispensables para la realización de ERLA.

INDICE GENERAL

RESUMEN I
INDICE GENERAL III
INDICE DEFIGURAS V
INDICE DE TABLAS VI
INTRODUCCIÓN VII
CAPÍTULO 1: MODELOS DE REGRESIÓN 1
1.1. Introducción 1
1.2. Regresión Lineal 2
1.3. Regresión Lineal Simple 3
1.3.1. Ilustraciones 6
1.3.2. Estimación de los Parámetros 8
1.3.3. Tipos de Estimadores 10
1.3.3.1. Estimación por Mínimos Cuadrados 12
1.3.3.2. Estimación por Máxima Verosimilitud 15
1.4. Regresión Lineal Múltiple 19
1.4.1. Representación Matricial del Modelo de Regresión Lineal Múltiple 19
1.4.2. Matriz Hat 22
1.4.3. Análisis de Varianza 24
1.4.3.1. Elaboración Tabla Anova 24
1.4.3.2. Grados de Libertad 25
1.4.3.3. Medias Cuadráticas 28
1.4.3.4. Contrastes de Hipótesis 33
CAPÍTULO 2: SELECCIÓN DE VARIABLES DE PREDICCIÓN 36
2.1. Introducción 36
2.2. Selección del Modelo 37
2.2.1. Coeficiente de Determinación (R^2) 38
2.2.2. R^2-Ajustado 40
2.2.3. Varianza Residual (s_R^2) 42
2.2.4. Estadístico C_p de Mallows 44
2.2.5. Criterio de Información Akaike (AIC) 48
2.2.6. Suma de Cuadrados de Predicción (PRESS) 50

CAPÍTULO 3: ACERCA DE ERLA 52
3.1. Introducción 52
3.2. Lenguaje y Códigos 53
3.2.1. MATLAB 53
3.2.2. VISUAL. NET 57
3.3. Conexión entre VISUAL BASIC.NET y MATLAB 59
CAPÍTULO 4: VALIDACIÓN DEL MODELO EN EL SOFTWARE “ERLA” 62
4.1. Introducción 62
4.2. Validación para el Modelo de Regresión Lineal Simple 63
4.3. Validación para el Modelo de Regresión Lineal Múltiple 70
4.4. Validación para los Indicadores de Selección de Modelos: R^2 Ajustado, C_p Mallows, Akaike Y PRESS. 74
CONCLUSIONES 79
RECOMENDACIONES 82
REFERENCIAS BIBLIOGRAFICAS 83

INDICE DE FIGURAS

Figura 1: *Relación Lineal Entre X Y Y* 4
Figura 2: *Distribución De Y_i*..... 5
Figura 3: *Representación Gráfica Del Máximo Y Mínimo De Una Función* 10
Figura 4: *Representación Gráfica De La Ecuación Ajustada*. 26
Figura 5: *Representación Gráfica Del Indicador C_p Mallows*..... 46
Figura 6: *Entorno Gráfico De Matlab*. 54
Figura 7: *Función "Regresión Lineal"*..... 55
Figura 8: *Funciones Para "Selección De Modelos" - R^2 Ajustado.* 56
Figura 9: *Programación En Visual Para "Selección De Modelos".* 58
Figura 10: *Creación De Archivos *.Dll*. 59
Figura 11: *Añadir Referencia En Visual Basic .Net.* 60
Figura 12: *Gráfica De Dispersión De Las Variables "Tensión Sistólica" Vs. "Edad"*. 66
Figura 13: *Histogramas De Frecuencias Y Diagramas De Cajas De B_0, B_1, s_{b_0} Y s_{b_1}* 69
Figura 14: *Graficas De Tendencia De Los Indicadores De Selección De Modelos*: 78

INDICE DE TABLAS

Tabla 1: *Tabla de Análisis de Varianza - Anova* 29
Tabla 2: *Tabla de Análisis de Varianza - (Anova) Forma Matricial* 32
Tabla 3: *Tensión Arterial Sistólica y Edad de 69 Pacientes* 63
Tabla 4: *Estadísticas Básicas de las Variables "Tensión Sistólica" y "Edad" Caso: "Regresión Lineal Simple"* 64
Tabla 5: *Tabla de Análisis de Varianza (Anova) de las Variables "Tensión Sistólica" y "Edad" Caso: "Regresión Lineal Simple"* 65
Tabla 6: *Estimadores de Parámetros Betas. Muestra: 30, N=69 Y E ~ N(0,1)* 67
Tabla 7: *Estadísticas Básicas de los Estimadores de los Parámetros Betas* 68
Tabla 8: *Estadísticas Básicas de las Variables "Importaciones", "Precio Relativo" y "Pib Real" Caso: "Regresión Lineal Múltiple"* 71
Tabla 9: *Tabla de Análisis de Varianza (Anova) de las Variables "Importaciones", "Precio Relativo" Y Pib Real" Caso: "Regresión Lineal Múltiple"* 72
Tabla 10: *Estimadores de Parámetros Betas. Muestra: 30, N=41 y e ~ N(0,1) Caso: "Regresión Lineal Múltiple"* 73
Tabla 11: *Estadísticas Básicas de los Estimadores de los Parámetros Betas Caso: "Regresión Lineal Múltiple"* 74
Tabla 12: *Valores de los Indicadores R^2 Ajustado, C_p Mallows, Akaike y Press – De Las 1024 combinaciones de las diez variables de explicación (Once Parámetros)* 76

INTRODUCCIÓN

En la actualidad se encuentran en el mundo un sin número de paquetes o aplicaciones estadísticas los cuales permiten efectuar el análisis descriptivo, inferencial, de un conjunto de datos. Estos paquetes para llegar al mercado pasan por un proceso de transición en el cual se llegan a corregir errores o fallas. Día tras día se busca que los programas sean cada vez más amigables a la vista del usuario, sin perder por supuesto el propósito del mismo, es por todo esto que como proyecto de graduación en las aulas del Instituto de Ciencias Matemáticas de la ESPOL, nace la idea de desarrollar un programa que cumpla con lo antes propuesto, el cual es "ERLA".

El desarrollo de "ERLA" ha sido realizado en dos plataformas informáticas estas fueron Matlab R2010a[1] y Visual Net 2008[2], lográndose una conexión basados en una estructura cliente/servidor; esta conexión en el ambiente informático es administrada por el componente conocido como Middleware[3] (COM). El middleware es un software de conectividad que ofrece un conjunto de servicios que hacen posible el funcionamiento de aplicaciones distribuidas sobre plataformas

[1] El fabricante de Matlab es MathWorks
[2] Visual Net fue creado por Microsoft
[3] Software desarrollado por Microsoft

heterogéneas y COM es el tipo de Middleware que permite la conexión específica entre las dos plataformas usadas en nuestro caso.

"ERLA" es un software direccionado a resolver problemas estadísticos utilizando Regresión Lineal. Este "paquete" contiene desde estadística básica como Tablas de Frecuencias, Estadísticas Descriptivas hasta Regresión de Ridge, Regresión Logística, Selección de Modelos, Puntos de Influencia y más. Siendo los indicadores de calidad de Selección de Modelos la contribución específica que se detallará en este reporte.

CAPÍTULO 1

1. MODELOS DE REGRESIÓN

1.1. Introducción

Una de las técnicas Estadísticas de mayor relevancia es Regresión Lineal; en un marco generalizado es determinar la dependencia o la relación existente entre una variable respuesta Y y una o más variables explicativas, $X_1, X_2, \ldots, X_{p-1}$.

En este capítulo se consideran los principales fundamentos teóricos de la técnica, *"Regresión Lineal Simple y Múltiple"*; entre ellos la estimación de parámetros por mínimos cuadrados y máxima verosimilitud, los supuestos que se debe considerar en el modelo, contrastes de hipótesis, elaboración de la tabla ANOVA. Además se explicará el Coeficiente de Determinación, los supuestos que deben cumplir las variables explicativas y de respuesta.

1.2. Regresión Lineal

El término *regresión* fue introducido por el científico inglés Francis Galton en su libro *"Natural Inheritance"* y se utilizó por primera vez en el estudio de variables antropométricas, al comparar la estatura de padres e hijos, resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, "regresaban" al promedio. La constatación empírica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

Esta técnica establece una relación funcional entre una variable dependiente y un conjunto de variables independientes. Un aspecto de interés sería determinar qué variables independientes explican a la dependiente. Puede existir también más de una variable dependiente, (Regresión Multivariada) caso que no consideraremos en este desarrollo.

Se pueden distinguir tres casos de acuerdo con el número de variables de explicación y al modelo que se utilice:

- **Regresión Lineal Simple**: en este caso se tiene una variable independiente, una variable dependiente y una relación rectilínea entre ellos.

- **Regresión Polinómica**: se tiene una variable dependiente y una variable de explicación, que se relacionan por un modelo polinómico.

- **Regresión Lineal Múltiple**: para este caso se tiene a una variable dependiente y varias variables de explicación o independientes.

1.3. Regresión Lineal Simple

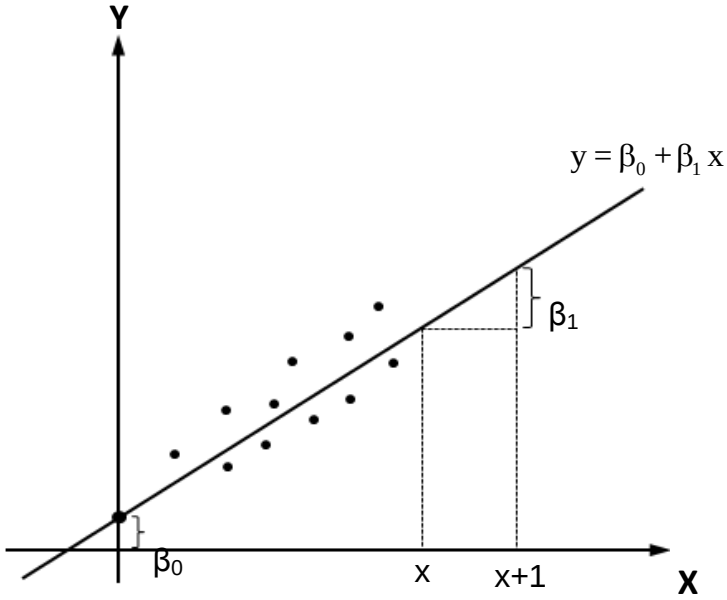
En la vida real se presentan variables de estudio, donde en diferentes ocasiones se presenta el interés de explicar una de estas variables en términos de otra. Definiendo "Y" como la variable que se quiere explicar y "X" la variable que explica a "Y" por medio de una relación funcional, que no conocemos donde experimentalmente podemos fijar ***n*** valores de "X" y leer "Y", obteniendo n valores de "Y"; existirían entonces n pares, $(x_1, y_1)^T$, $(x_2, y_2)^T$, … , $(x_n, y_n)^T$. Simplificando tendríamos vectores bivariados $x_1^T; x_2^T; \ldots; x_n^T$; donde $x_i^T \in R^2;\ i = 1, 2, \cdots n$, esto es $x^T = (x, y)$.

Recordando la expresión $y = mx + b$ que explica una recta con pendiente *m* e intersección con el eje vertical igual a *b*, se propone un modelo de la siguiente forma:

$$y = \beta_0 + \beta_1 x$$

Donde β_0 y β_1 son constantes desconocidas, pero estimables estadísticamente; β_1 es la pendiente de la recta, en tanto que β_0 es el punto de intersección con el eje de *Y*. En la Figura 1 se muestra una Relación Lineal entre X y Y.

Figura 1: Relación Lineal entre X y Y
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

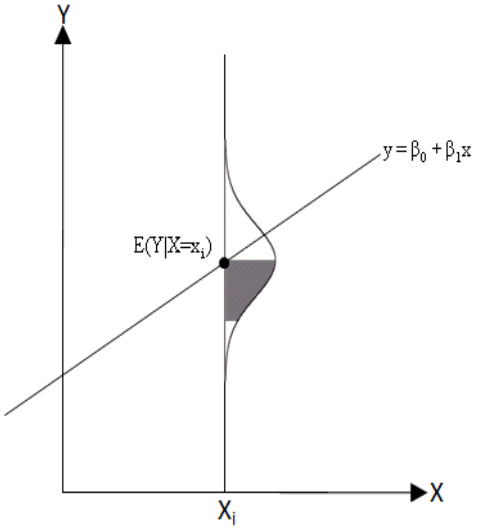
Regresión Lineal Simple es la técnica estadística con que se utiliza el modelo mencionado anteriormente. Mientras que la aproximación estadística es la siguiente; se supone que "X" explica a "Y" en

términos de una recta, esto induce a que cada valor observado de "Y" no siempre determina un punto que pertenece a la recta, es porque al efectuar la medida de "Y" una vez fijada "X" se genera un Error aleatorio "ε", de tal manera que los valores de "Y" son dados por la siguiente relación funcional, denominada Regresión Lineal Simple.

$$y = \mu + \varepsilon, \text{ donde } \mu = \beta_0 + \beta_1 x \qquad \textbf{(1.1)}$$

La distribución de los $Y_{i,}$ junto con la recta que representa la parte determinística de este modelo se la puede apreciar en la *Figura 2*

Figura 2: Distribución de Y_i
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

Entonces:

$$E\left[Y \mid X = x\right] + \varepsilon \;=\; \beta_0 + \beta_1 x + \varepsilon$$

Suponiendo se tienen *n* pares de observaciones $\begin{pmatrix} x_i \\ y_i \end{pmatrix}$ para i = 1, 2, 3,…,n, con las *n* observaciones el modelo de regresión lineal simple es el siguiente:

$$\begin{aligned} Y_i &= E\left[Y \mid X = x_i\right] + \varepsilon_i \qquad && i = 1, 2, 3, \ldots, n. \\ &= \beta_0 + \beta_1 x_i + \varepsilon_i \qquad && E\left(\varepsilon_i\right) = 0 \end{aligned} \qquad \textbf{(1.2)}$$

Siendo $\beta_0 + \beta_1 x_i$ la parte determinística del modelo ya que x_i se fija con anticipación.

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

Siendo σ^2 una constante.

1.3.1. Ilustraciones

Con la matriz de datos siguiente:

$$\begin{matrix} Y & X \end{matrix}$$
$$\begin{bmatrix} 8 & 5 \\ 1 & 3 \\ \vdots & \vdots \\ 4 & 9 \end{bmatrix}$$

La primera columna se la identifica como la variable respuesta y la segunda como la variable de explicación, por lo tanto tomando en cuenta los supuestos previamente mencionadas, del modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ para $i = 1, 2, 3, \ldots, n,$ con dos coeficientes de regresión se obtiene:

$$\begin{aligned} 8 &= \beta_0 + 5\beta_1 + \varepsilon_1 \\ 1 &= \beta_0 + 3\beta_1 + \varepsilon_2 \\ \vdots & \qquad \vdots \\ 4 &= \beta_0 + 9\beta_1 + \varepsilon_n \end{aligned}$$

Representando en forma matricial se tiene:

$$\begin{bmatrix} 8 \\ 1 \\ \vdots \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 9 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Expresando la regresión lineal simple en forma general como:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Y con esto se reduce a:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(1.3)}$$

Donde $\mathbf{Y} \in R^n$; además **X** es la Matriz de Diseño que es n x 2; $\boldsymbol{\beta} \in R^2$ es el vector de parámetros; y, $\boldsymbol{\varepsilon} \in R^n$ es el vector de errores.

1.3.2. Estimación de los Parámetros

En el modelo **(1.2)** aparecen parámetros β_0, β_1 y σ^2 a los cuales en una situación pre experimental nunca se los conoce; es aquí donde aparece la necesidad de disponer de métodos para la estimación de estos parámetros. Como métodos de estimación de los parámetros del modelo de regresión se identifican al denominado de Mínimos Cuadrados así como la estimación de Máxima verosimilitud.

Estos métodos utilizan técnicas de maximización y minimización de funciones, estas funciones pueden tener, en un determinado

intervalo, máximos y mínimos, gráficamente un máximo se presenta cuando a la izquierda de la función esta crece y a su derecha decrece y el mínimo cuando a la izquierda la función decrece y a su derecha crece; analíticamente para la determinación de máximos y mínimos podemos utilizar los siguientes criterios:

Criterio de la primera derivada:

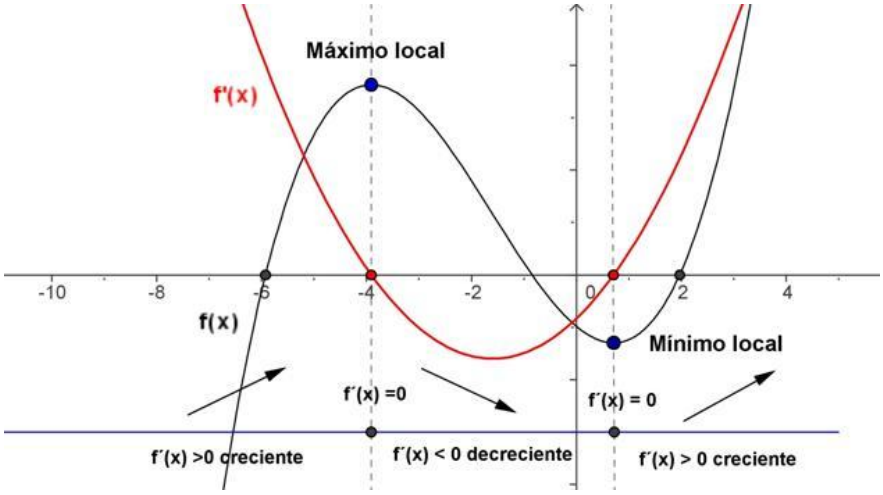
El método o teorema utilizado frecuentemente en el cálculo matemático para determinar los mínimos relativos y máximo relativos que pueden existir en una función f mediante el uso de la primera derivada, donde se observa el cambio de signo, en un intervalo abierto señalado que contiene al punto crítico sea este máximo o mínimo.

Luego de calcular la primera derivada, la igualamos a cero $\left(f'(x)=0\right)$ y resolvemos la ecuación resultante, determinamos la segunda derivada. Las raíces de la ecuación obtenida se sustituyen en la segunda derivada. Si el resultado obtenido es positivo existe un mínimo en tal punto y si es negativo se tiene un máximo.

$$\frac{\partial f(x)}{\partial x}=0 \quad \wedge \quad \begin{cases} \dfrac{\partial^2 f(x)}{\partial^2 x}<0; & \Rightarrow \quad \text{es un Máximo} \\ \dfrac{\partial^2 f(x)}{\partial^2 x}>0; & \Rightarrow \quad \text{es un Mínimo} \end{cases}$$

En la *Figura 2* se puede observar gráficamente el criterio de la Primera derivada y de la Segunda derivada.

Figura 3: Representación Gráfica del Máximo y Mínimo de una función
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

1.3.3. Tipos de Estimadores.

- ESTIMADOR INSESGADO significa que su media o valor esperado coincide con el valor del parámetro θ desconocido, pero estadísticamente estimable, esto es: $E\left[\hat{\theta}\right]=\theta$ y por lo tanto, su sesgo $E\left[\hat{\theta}\right]-\theta=0$ por lo que; si $\hat{\theta}$ es insesgado, entonces la media cuadrática del error a ser estudiada más adelante será $MCE\left[\hat{\theta}\right]=Var\left[\hat{\theta}\right]$

- ESTIMADOR EFICIENTE: si para estimar un mismo parámetro θ, disponemos de dos estimadores insesgados, el estimador más eficiente entre los dos es el de menor varianza.

  Sea $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de un mismo parámetro θ.

  Si $\mathrm{Var}\left[\hat{\theta}_1\right] < \mathrm{Var}\left[\hat{\theta}_2\right]$ entonces $\hat{\theta}_1$ es un estimador insesgado más eficiente de θ que $\hat{\theta}_2$; y, $\hat{\theta}_2$ sigue siendo un estimador insesgado pero menos eficiente que $\hat{\theta}_1$.

- Un estimador $\hat{\theta}$ de θ es un Estimador Asintóticamente Insesgado si al aumentar el tamaño de la muestra, su media tiende a coincidir con el parámetro θ, y por lo tanto, su sesgo tiende a cero.

  Esto es $\lim_{n\to\infty} \mathrm{E}\left[\hat{\theta}\right] = \theta$.

- ESTIMADOR CONSISTENTE significa que a medida que crece el tamaño de la muestra las estimaciones que nos proporciona el estimador $\hat{\theta}$ se aproximan cada vez más al valor del parámetro θ.

  Decimos que $\hat{\theta}$ es un estimador consistente del parámetro θ si:

  $$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathrm{P}[\left|\hat{\theta} - \theta\right| > \in] = 0$$

  O lo que es equivalente:

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} P[\left|\hat{\theta} - \theta\right| < \in] = 1$$

1.3.3.1. Estimación por Mínimos Cuadrados

Este es un método de ajuste de curvas que a principios del siglo XIX sugirió el matemático francés Adrien Legendre.

Para la estimación por mínimos cuadrados se efectúa la diferencia entre el valor observado y_i y el valor esperado de y_i el cual es $\mu_i = \beta_0 + \beta_1 x_i$ con lo que se tiene $\varepsilon_i = y_i - \mu_i = y_i - \beta_0 - \beta_1 x_i$ y cuyo estimador es $\hat{\varepsilon}_i = (y_i - \hat{y}_i) = e_i$ para i=1, 2, 3,…, n, estos errores se espera sean lo más pequeños posible. Una aproximación para lograr esto, es minimizar la función

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum (y_i - \mu_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_i)^2 \qquad \textbf{(1.4)}$$

Para la minimización de esta función se aplican derivadas con respecto a los parámetros β_0 y β_1, se iguala a cero para determinar los estimadores b_0 y b_1 de β_0 y β_1 respectivamente.

Esta aproximación usa la distancia cuadrática como una medida de proximidad. Cabe mencionar que se pueden utilizar otras medidas

tales como el valor absoluto de la diferencia. Tomando las derivadas con respecto a β_0 y β_1 e igualando a cero, se tiene:

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \qquad \textbf{(1.5)}$$

$$\frac{\partial S(\beta_0,\beta_1)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \qquad \textbf{(1.6)}$$

Luego de la derivación y sustituyendo β_0 por b_0 y β_1 por b_1, se obtienen las ecuaciones

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \qquad \textbf{(1.7)}$$

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\, x_i = 0 \qquad \textbf{(1.8)}$$

A estas dos últimas igualdades se las denomina Ecuaciones Normales. Suponemos que b_0 y b_1 son la solución para β_0 y β_1 en el sistema de dos ecuaciones. Resolviendo este sistema tenemos que:

$$\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i - n b_0 = 0$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad \textbf{(1.9)}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Big/ (n-1)}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \Big/ (n-1)} \qquad \textbf{(1.10)}$$

b_0 y *b_1* son llamados estimadores de mínimos cuadrados de *"$\hat{\beta}_0$"* *y* *"$\hat{\beta}_1$"* respectivamente; los mismos que minimizan S en **(1.4)** lo cual puede ser comprobado con el criterio de la segunda derivada.

Claramente se observa que, el numerador de la expresión que determina a *"b_1"* es el estimador de la covarianza entre *"X"* y "*Y*" en tanto que el denominador es el estimador de la varianza de *"X"*.

Las características de los estimadores b_0 y b_1 por Mínimos Cuadrados de acuerdo con el Teorema de Gauss y Markov es que son insesgados y de mínima varianza.

1.3.3.2. Estimación por Máxima Verosimilitud

Sea X una variable aleatoria con función de probabilidad $f(x;\theta)$, Las muestras aleatorias simples de tamaño n, $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ tienen por distribución de probabilidad conjunta:

$$f(x_1, x_2, ..., x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta)$$

Esta función que depende de (n+1) cantidades se la considera de dos maneras:

- Fijando θ, es una función de las n cantidades x_i.
- Fijados los x_i como consecuencia de los resultados de elegir una muestra mediante un experimento aleatorio, es únicamente función de θ. A esta función de θ la denominamos *"función de verosimilitud"*.

El método de *"Máxima Verosimilitud"*, propone como un estimador el valor que maximiza la probabilidad de obtener la muestra ya disponible. Este método se basa, en la distribución del error. A tales efectos, se suele suponer que los errores aleatorios tienen una distribución Normal, con lo que $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Como consecuencia de lo anterior, se supondrá que del modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, el término aleatorio y_i sigue una distribución Normal con la siguiente función de densidad:

$$f(y_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2\sigma^2}(y_1-(\beta_0+\beta_1 x_1))^2} \qquad \textbf{(1.11)}$$

Ya que $\varepsilon_i \sim N(0, \sigma^2)$ siendo σ^2 constante, decimos que el modelo planteado es homocedástico.

La función **(1.11)** es para i = 1, por tanto, la expresión de la función de densidad conjunta para el vector $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ es la siguiente:

$$f(\mathbf{Y}) = f\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \prod_{i=1}^{n} f(y_i) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2} \qquad \textbf{(1.12)}$$

Como $\mathbf{Y}$ sigue una distribución Normal de orden n; el vector aleatorio $\mathbf{Y}$ al incluir los errores aleatorios, también tendrá distribución Normal Multivariada; pues, para que la función de densidad conjunta sea una

función de verosimilitud, el vector aleatorio ε ha de expresarse en función del vector **Y**, es decir:

$$L(\mathbf{Y};\boldsymbol{\beta},\sigma) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}} \exp^{-\frac{1}{2\boldsymbol{\sigma}^2}\left((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)} \qquad \textbf{(1.13)}$$

Siendo ahora parámetros $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$ y σ^2 constante.

Se trata, por tanto, de maximizar la función de verosimilitud L, presentada en **(1.13)**.

Para calcular el máximo de la función de verosimilitud L, es necesario determinar los valores para los cuales la derivada con respecto a **β** y σ^2 de la verosimilitud es igual a cero, pero por definición la función de verosimilitud es un producto de densidades, lo cual puede ser bastante engorroso de derivar. Por lo tanto es preferible derivar una suma, y es por esto que se substituye la función de verosimilitud por su logaritmo. Ya que la función logarítmica es una función monótona creciente, por lo que es equivalente maximizar $\log\left(L(\mathbf{Y};\boldsymbol{\beta},\sigma)\right)$ *o* $L(\mathbf{Y};\boldsymbol{\beta},\sigma)$. Una vez determinado el valor de los

estimadores de los parámetros **β** y σ^2 obtenidos de la derivación, hay que verificar con el término de la segunda derivada, que el punto en cuestión es realmente un máximo.

$$\ln\left(L(\mathbf{Y},\boldsymbol{\beta},\boldsymbol{\sigma}^2)\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(2\sigma^2\right) - \frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \qquad \textbf{(1.14)}$$

Los estimadores de máxima verosimilitud para **β** se determinan, resultando ser:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}$$

Cuya matriz de varianzas y covarianzas es:

$$\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sum_{\mathbf{b}} \qquad \textbf{(1.15)}$$

Observamos que el estimador de *"Máxima verosimilitud"* de **β** coincide con el de *"Mínimo Cuadrados"*, con lo que tendrá las mismas propiedades: insesgados y de mínima varianza, de acuerdo al Teorema de Gauss y Markov. El estimador de Máxima Verosimilitud de σ^2, en cambio, resulta diferente del Mínimo

Cuadrado y no es insesgado aunque sí es asintóticamente insesgado.

1.4. Regresión Lineal Múltiple

En el modelo de regresión lineal múltiple se mantienen los mismos supuestos planteados para el caso de regresión lineal simple, para este se consideran (p-1) variables de explicación, y se lo define como sigue:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p\text{-}1} x_{i\,p\text{-}1} + \varepsilon_i, \quad i = 1, 2, 3, \ldots, n \qquad \textbf{(1.16)}$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases}$$

Siendo σ^2 constante, lo que indica homocedasticidad.

1.4.1. Representación Matricial del Modelo de Regresión Lineal Múltiple

El modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i$ para i=1, 2, 3,…, n, con *p* parámetros ó (p-1) variables de explicación, se lo puede representar matricialmente de la siguiente manera:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Quedando como en el caso previo

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

Donde $\mathbf{Y} \in \mathrm{R}^{\mathrm{n}}$ es el vector de la variable a ser explicada, $\mathbf{X} \in \mathrm{M}_{\mathrm{nxn}}$ es la matriz de diseño**,** $\boldsymbol{\beta} \in \mathrm{R}^{\mathrm{p}}$ es el vector de parámetros y $\boldsymbol{\varepsilon} \in \mathrm{R}^{\mathrm{n}}$ el vector de errores.

Para la estimación de los parámetros, Betas, se puede utilizar Mínimos Cuadrados o de Máxima Verosimilitud. Para el caso de Regresión Lineal Simple utilizando mínimos cuadrados se realizaba la derivación de la diferencia:

$$\mathrm{S}(\beta_0,\beta_1) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_i^2 = \sum \left(\mathrm{y_i} - \mu_{\mathrm{i}}\right)^2 = \sum \left(\mathrm{y_i} - \beta_0 - \beta_1 \mathrm{x_i}\right)^2$$

En este caso se tendrá:

$$\begin{aligned} \mathrm{S}(\beta_0, \beta_1, \dots, \beta_{\mathrm{p-1}}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \varepsilon_i^2 &= \sum \left(y_i - \mu_i\right)^2 \\ &= \sum \left(\mathrm{y_i} - \beta_0 - \beta_1 \mathrm{x_{i\,1}} - \dots - \beta_{\mathrm{p-1}} \mathrm{x_{i\,p-1}}\right)^2 \end{aligned}$$

Luego se determinan las derivadas con respecto a cada “beta” e igualando a cero, y se tiene:

$$\frac{\partial S(\beta_0,\beta_1,\ \ldots,\ \beta_{p-1})}{\partial \beta_0} = 0$$
$$\frac{\partial S(\beta_0,\beta_1,\ \ldots,\ \beta_{p-1})}{\partial \beta_1} = 0$$
$$\vdots$$
$$\frac{\partial S(\beta_0,\beta_1,\ \ldots,\ \beta_{p-1})}{\partial \beta_{p-1}} = 0$$

Es conveniente llevar estas *“ecuaciones normales”* a la forma matricial para mayor facilidad de cálculo.

Según el modelo de regresión lineal simple en el que solo se estiman dos parámetros, las ecuaciones normales serían:

$$\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) = 0 \ \rightarrow\ \sum_{i=1}^{n} y_i = n\,b_0 - b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right) x_i = 0 \quad \rightarrow \quad \sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

La forma matricial de este sistema de dos ecuaciones es la siguiente:

$$\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad \textbf{(1.17)}$$

De esto se tiene que

$$\mathbf{X^T Y = (X^T X)^{-1} b}$$

Determinando $\mathbf{b}$, se tiene

$$\mathbf{b = (X^T X)^{-1} X^T Y} \quad \textbf{(1.18)}$$

La ecuación **(1.18)** se la generaliza para la estimación de los p betas del modelo **(1.16)**. Un punto a considerar es que debe existir la inversa del producto de las matrices $\mathbf{X}^T$ con **X,** otra de las características es que ($\mathbf{X^T X}$) es simétrica y permite estimar la matriz de varianzas y covarianzas $\boldsymbol{\Sigma_b}$ de los estimadores b_o, b_1,…,b_{p-1}, por lo que se supone ésta es no singular, es decir su determinante es distinto de cero.

1.4.2. Matriz Hat

La "*Matriz Hat*", *"H"*, relaciona los valores ajustados $\hat{Y}_i$ con los valores observados Y_i, lo cual indica la influencia que cada valor observado

tiene sobre cada valor ajustado. Pues bien, suponiendo un modelo de regresión lineal, se tiene que

$$\hat{\mathbf{Y}} = \mathbf{X}b$$

Considerando la ecuación **(1.18)**, se obtiene:

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathbf{T}}\mathbf{Y} \qquad \textbf{(1.19)}$$

Llamaremos matriz *"Hat"* a:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}} \qquad \textbf{(1.20)}$$

Por lo tanto la expresión **(1.19)** se reduce a:

$$\hat{\mathbf{Y}} = \mathbf{HY} \qquad \textbf{(1.21)}$$

El vector de residuales se lo define

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

En términos de la matriz *"Hat"* los residuales serían

$$\mathbf{e} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y} \qquad \textbf{(1.22)}$$

La matriz *"Hat"* tiene aplicaciones prácticas en el análisis de regresión, tales como "apalancamiento" y "distancia Cook", que se ocupan de la identificación de observaciones que tienen un gran efecto sobre los resultados de una regresión, como veremos posteriormente en este trabajo.

1.4.3. Análisis de Varianza

El *Análisis de Varianza* es una aproximación para la evaluación del grado de fortaleza de la relación de regresión lineal.

En este análisis se realizan contrastes de hipótesis para los betas, se determinan los residuos, el coeficiente de determinación y la elaboración de la Tabla de Análisis de Varianza (ANOVA).

1.4.3.1. Elaboración Tabla Anova

La validez de los valores estimados en el modelo está dada por el ajuste del modelo, ajuste que se mide a través de indicadores de calidad a ser estudiados en el *Capítulo 2*.

La tabla de Análisis de Varianza (Tabla ANOVA), utilizada en Regresión para analizar estadísticamente la validez del modelo y los

supuestos alrededor del mismo, es un arreglo matricial, constituido en sus filas las descripciones consideradas por la fuente de variación tales como la de regresión, la del error y la total; y en sus columnas formadas por: la fuente de variación, los grados de libertad, las sumas cuadráticas, las medias cuadráticas y el valor del estadístico de prueba con distribución F de Fisher, estos parámetros serán explicados a continuación.

FUENTES DE VARIACION

La tabla ANOVA está conformada por tres fuentes de variación: la de "*Regresión"* que presenta los valores que se estudian explícitamente para las variables del modelo. La del *'' Error",* para estudiar los datos de los errores y la *"Total"* que presenta toda la información respecto al modelo completo.

1.4.3.2. Grados de Libertad

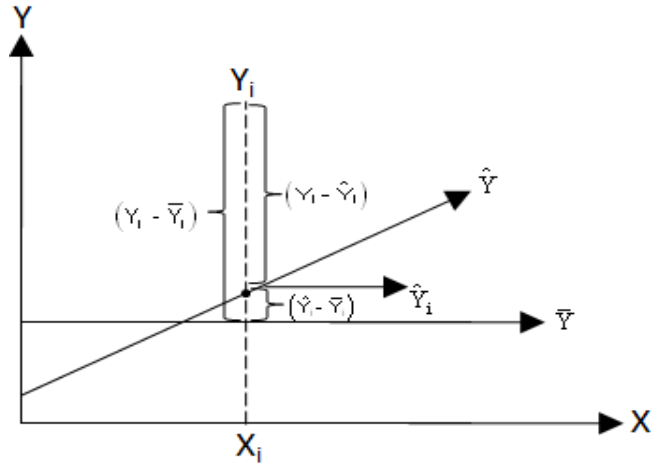
En Estadística, grados de libertad es un estimador del número de categorías independientes en una prueba particular o experimento estadístico. En la tabla ANOVA se presentan varias consideraciones de grados de libertad.

La fuente de variación de Regresión tiene (p-1) grados de libertad donde *p* es el número de variables y se le resta 1 por la variable dependiente "Y". Para el Error es similar ya que ésta se ve influenciada por el número de observaciones "n" y el número de variables "p", los grados de libertad son (n-p). En el caso de la fuente de variación Total es la suma de la de Regresión y Error que es (n-1) donde n sigue siendo el número de observaciones.

SUMAS CUADRATICAS

La *"Figura 3"*, explica un modelo ajustado a un dato. Para un valor "x_i" de X se ha tomado el correspondiente valor de "y_i" de Y.

Figura 4: Representación Gráfica de la ecuación ajustada.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

La distancia que hay entre el valor observado y la media de los valores observados de y $\left|y_i - \bar{y}\right|$ denominada distancia total, puede descomponerse en dos partes que son: la distancia entre el valor observado y el estimado por la regresión $\left(y_i - \hat{y}_i\right)$; y, la distancia entre el valor estimado y el promedio $\left(\hat{y}_i - \bar{y}\right)$ también llamada distancia de regresión, es decir:

$$\left(y_i - \bar{y}\right) = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{y}\right)$$

Como se tienen observaciones para cada caso se presenta la misma situación, por lo tanto se toma la suma de estas distancias al cuadrado. La variabilidad entre las *"y_i's"* usualmente se lo mide por las desviaciones de la media $y_i - \bar{y}$. Así, una medida de la variación total alrededor de la media está previsto por la suma cuadrática total SCT, la cual es $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$. Pues bien mediante esta suma cuadrática se establece lo siguiente:

$$\left(y_i - \bar{y}\right) = \left(y_i - \hat{y}_i + \hat{y}_i - \bar{y}\right)$$

Sumado y restado el valor estimado se tiene

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i + \hat{y}_i - \bar{y}\right)^2$$

Agrupando de la siguiente manera

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{y}_i - \bar{y}\right) + \left(y_i - \hat{y}_i\right)\right]^2$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{y}_i - \bar{y}\right)^2 + 2\left(\hat{y}_i - \bar{y}\right)\left(y_i - \hat{y}_i\right) + \left(y_i - \hat{y}_i\right)^2\right]$$

Quedando finalmente

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

SUMA CUADRÁTICA TOTAL	**SUMA CUADRÁTICA DE REGRESIÓN**	**SUMA CUADRÁTICA DE ERROR**

De estas sumas cuadráticas SCT=SCR+SCE, la del error (SCE) es la que se desearía fuera lo más pequeña posible.

1.4.3.3. Medias Cuadráticas

Las medias cuadráticas son un cociente, entre las sumas cuadráticas y sus grados de libertad. La media cuadrática del error es el estimador de la varianza del error y por lo tanto de las y_i. Adicionalmente a esto tenemos el valor F_0 el cual es definido como:

$$F_0 = \frac{MCR}{MCE} \qquad \textbf{(1.24)}$$

Se puede probar que bajo supuestos de normalidad e independencia que el estadístico F_0 es un cociente de variables aleatorias Ji cuadrado independientes, por lo que su distribución es Fisher, $F(\nu_1, \nu_2)$ donde $\nu_1 = (p-1)$ son los grados de libertad del numerador y $\nu_2 = (n-p)$ los grados de libertad del denominador. La *"Tabla 1"* presenta lo que es una Tabla de Análisis de Varianza (ANOVA).

Tabla 1: Tabla de Análisis de Varianza - ANOVA.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	p-1	$SCR=\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MCR=SCR/p-1	$\frac{MCR}{MCE}$
Error	n-p	$SCE=\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MCE=SCE/n-p	
Total	n-1	$SCT=\sum_{i=1}^{n}(y_i - \bar{y})^2$		

Autoría: Macías S. – Pincay C.

Usando la expresión de los estimadores de betas **(1.18)** con respecto al modelo $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ se tiene que:

$$\hat{\mathbf{Y}} = \mathbf{Xb}$$

(1.25)

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\ \mathbf{X}^T\mathbf{Y}$$

Resolviendo algebraicamente la expresión

$$SCT=\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$$

se llega

$$SCT=\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}$$

la expresión anterior queda como sigue:

$$\sum_{i=1}^{n}y_i^2=\begin{bmatrix}y_1 & y_2 & \cdots & y_n\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}=\mathbf{y}'\mathbf{y}$$

Dicho esto, la expresión

$$\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}=\frac{1}{n}\mathbf{y}^{T}\mathbf{J}\,\mathbf{y}$$

donde **J** es una matriz de 1's de dimensión "m x n", siendo *m* el número de fila y *n* el de columnas, por lo tanto

$$SCT = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{\mathrm{n}}\mathbf{y}^{\mathrm{T}}\mathbf{J}\,\mathbf{y} \qquad \textbf{(1.26)}$$

Para la $\mathrm{SCE} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}\right)^2$ se obtiene lo siguiente:

$$\begin{aligned} SCE &= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\,\mathbf{y} \\ &= \mathbf{y}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{H}\right)\mathbf{y} \end{aligned} \qquad \textbf{(1.27)}$$

De estas dos y de acuerdo con la ecuación SCT=SCR+SCE se obtiene:

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

$$\mathrm{SCR} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \frac{1}{\mathrm{n}}\mathbf{y}^{\mathrm{T}}\mathbf{J}\,\mathbf{y} - \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\,\mathbf{y}\right)$$

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

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

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

Por lo que:

$$SCR = \mathbf{y}^{\mathrm{T}}\left(\mathbf{H} - \frac{1}{\mathrm{n}}\mathbf{J}\right)\mathbf{y} \qquad \textbf{(1.28)}$$

La “Tabla 2” muestra la tabla de Análisis de Varianza (ANOVA) con las sumas cuadráticas expresadas en forma Matricial, esto a partir de las ecuaciones **(1.26), (1.27)** y **(1.28).**

Tabla 2: Tabla de Análisis de Varianza - (ANOVA) Forma Matricial.
“Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad”

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	p-1	$SCR = \mathbf{y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$	MCR=SCR/p-1	$F_o = \frac{\text{MCR}}{\text{MCE}}$
Error	n-p	$SCE = \mathbf{y}'\left(\mathbf{I} - \mathbf{H}\right)\mathbf{y}$	MCE=SCE/n-p	
Total	n-1	$SCT = \mathbf{y}'\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\,\mathbf{y}$		

Autoría: Macías S. – Pincay C.

Junto con la Tabla ANOVA se determina la calidad del modelo con indicadores que expresan cuan eficiente es el modelo de regresión lineal o múltiple según sea el caso. Para esto si la SCE=0, lo cual sería el modelo perfecto, ya que eso implicaría que la variable o variables independientes *“X´s”* explican perfectamente a *“Y”*, es decir SCT=SCR y para el caso del Coeficiente de Determinación (R^2) que será tratado en su momento se tendría que R^2=1, nótese que este cociente por la forma que se lo define, no puede ser menor que cero ni mayor que uno, ya que SCR≤SCT; cabe mencionar que este no es

el único indicador de eficiencia del modelo, existen otros tales como el R^2 ajustado, el de Akaike, el C_p Mallows que serán explicados y analizados en el capítulo siguiente. La denominada potencia de explicación del modelo, es definida como R^2 x 100.

1.4.3.4. Contrastes de Hipótesis

Para conocer si el modelo de regresión propuesto mide en realidad la relación lineal existente, es de sumo interés realizar una prueba que ofrezca la evidencia estadística para justificar el modelo. Por esto, sea el caso del modelo de regresión lineal simple en que se tiene a los parámetros $\beta_0, \ldots \beta_{p-1}$, se esperaría que el β_1 que es el coeficiente de la única variable de explicación sea distinto de cero, ya que de no ser así el modelo sería una recta constante, para el caso de regresión múltiple sería de igual forma, por lo tanto para comprobar estadísticamente se realiza el contraste de hipótesis correspondiente, que es el siguiente.

$$\mathbf{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

vs

$$\mathbf{H}_0 : \text{Al menos uno de los betas es distinto de cero}$$

En vista de que $F = \frac{MCR}{MCE}$ tiene distribución $F_{(p-1,\, n-p)}$, con $(1-\alpha)100\%$ de confianza se debe rechazar H_0 a favor de H_1, si el estadístico F_0 en

(1.24) es mayor que el percentil $(1-\alpha)100$ de $\mathrm{F}(\nu_1, \nu_2)$ con $\nu_1 = (p-1)$ grados de libertad en el numerador y $\nu_2 = (n-p)$ grados de libertad en el denominador.

$$F_0 = \frac{MCR}{MCE} > F_{(\alpha,\, p-1,\, n-p)}$$

Una vez que $\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ ha sido rechazada, si es que esto ocurre, se realiza la prueba individual para determinar cuáles de los betas son distintos que cero y por lo tanto que variables aportan al modelo. El contraste de hipótesis para cada beta será:

$$\begin{aligned} &\mathrm{H}_0 : \beta_i = 0 \\ &\quad vs \qquad para\ i = 1, 2, \ldots, p-1 \\ &\mathrm{H}_1 : \beta_i \neq 0 \end{aligned}$$

Donde se utiliza como estadístico de prueba $t = \frac{b_i - \beta_i}{s_{b_i}}$ que tiene distribución de t-Student con (n-p) grados de libertad por lo que con $(1-\alpha)100\%$ de confianza rechazar H_0 a favor de H_1, si el valor absoluto del estadístico t es mayor que el percentil $(1-\alpha)$ con (n-p) grados de libertad.

$$\left|\frac{b_i - \beta_i}{s_{b_i}}\right| > t_{(\alpha/2,\, n-p)}$$

Se obtiene de igual manera el coeficiente de determinación R^2, por lo general la potencia de explicación del modelo debería ser mayor que 80%, para considerar que el modelo de regresión utilizado es aceptable.

CAPÍTULO 2

2. SELECCIÓN DE VARIABLES DE PREDICCIÓN

2.1. Introducción

Antes de iniciar el análisis de regresión, se realiza una investigación básica a las variables objeto de estudio, todo esto con el fin de observar el comportamiento y las fortalezas de la relación entre ellas. Dicho de otra manera, se realiza el análisis descriptivo y determinamos las correlaciones entre dichas variables, para de esta manera observar qué variables son las que aportarían en proporción significativa a los modelos de regresión.

Ante esto nos vemos obligados a realizar empíricamente la selección de las variables explicativas, aquellas combinaciones de variables que de acuerdo con la matriz de correlación determinamos tienen mayor fortaleza con la variable respuesta. Existen métodos de selección de las variables explicativas, pero no son comunes en los softwares estadísticos más usuales.

Como tema específico en este capítulo se detallarán las técnicas que permiten determinar las posibles regresiones de un conjunto de variables explicativas $X_1, X_2, \ldots, X_{p-1}$, para una variable a ser explicada *Y*. Dichas técnicas, son las que utilizan R^2, R^2_{aj}, Criterio de Akaike, estadístico C_p de Mallows y PRESS.

2.2. Selección del Modelo

Para decidir entre dos o más subconjuntos de variables explicativas en el estudio de un modelo de regresión múltiple es interesante disponer de indicadores que midan la bondad del ajuste del modelo construido. Se supone que el número de variables explicativas que pueden haber en el modelo es (p -1), el número de observaciones es n; y, si se ajusta un modelo de regresión lineal con estas variables explicativas, el número de parámetros del modelo es p. Entonces se definen las siguientes medidas de bondad de ajuste: R^2; R^2_{aj}; Criterio de Akaike; Estadístico C_p de Mallows; y, PRESS.

2.2.1. Coeficiente de Determinación (R^2)

R^2, definido en la sección anterior. Como:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Este indicador intenta medir la calidad del modelo utilizado y aumenta al ir introduciendo nuevas variables en el modelo. Se denota R_j^2 j=1,...,p-1, el máximo valor posible de R^2 cuando en el modelo hay "*j*" variables explicativas, se verifica $R_{j-1}^2 < R_j^2$, (R_j^2 es monótona creciente) y las diferencias $R_j^2 - R_{j-1}^2$ decrecen. En base a esto, al crecer "*j*" un criterio sería considerar un número pequeño que por conveniencia es denotado por "Δ" y elegir el modelo con "*j*" más pequeño y tal que $R_{p-1}^2 - R_j^2 < \Delta$; siendo R_{p-1}^2 el coeficiente de determinación del modelo con las "p-1" variables explicativas.

Puesto que a medida que se introducen variables en el modelo, la potencia de explicación aumenta y además R^2 tiene el inconveniente de no considerar el número de variables explicativas, lo que hace que tienda a sobre ajustar y utilizar demasiadas variables.

El R^2_{p-1} es el coeficiente de determinación R^2 para un modelo con (*p-1*) variables de explicación "*p*" coeficientes de regresión, en líneas previas se dijo que:

$$R^2_{p-1} = \frac{SCR_{p-1}}{SCT}$$

Debido a que la SCT = SCR + SCE, manipulando algebraicamente se obtiene:

$$R^2_{p-1} = 1 - \frac{SCE_{p-1}}{SCT}$$

Donde SCE_{p-1} es la Suma Cuadrática del Error para el modelo con (p-1) variables de explicación, y $SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ es la Suma Cuadrática Total que es la misma para todos los modelos donde "*p-1*" no cambia.

Es preferible tener modelos con R^2_{p-1} de mayor tamaño. Habrá varios modelos con "*p-1*" variables y cada uno tendrá un Coeficiente de Determinación (R^2) diferente. Esto tendría sentido para seleccionar el mejor o los mejores R^2 de los modelos de "*p-1*" variables.

2.2.2. R^2-Ajustado

El R^2_{adj} ajustado, tiene como principal importancia determinar la variabilidad explicada por las variables explicativas, con respecto a la variable respuesta cuando se introduce una variable adicional al modelo.

El Coeficiente de Determinación Ajustado (R^2_{adj}) se define: por los grados de libertad asociados a la sumas cuadráticas; la SCE y la SCT son ajustados por *(n-p-1)* y por *(n-1)* que son sus grados de libertad respectivamente.

En términos de sumatorias R^2_{adj} se define por la expresión

$$R^2_{adj} = 1 - \frac{\frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Simplificando

$$R^2_{adj} = 1 - \frac{\frac{1}{n-(p+1)}SCE}{\frac{1}{n-1}SCT}$$

Quedando R^2_{adj} en términos del Coeficiente de Determinación R^2, definido por la siguiente expresión

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p-1)}\left(1-R^2\right)$$

Dicha expresión en términos de varianzas se tiene que:

$$R^2_{adj} = 1 - \frac{s^2}{SCT/(n-1)} = 1 - \frac{s^2}{s^2_y}$$

Donde $s^2 = \frac{SCE}{(n-p-1)}$ es la Media Cuadrática de los Residuos, y s^2_y es la varianza de la muestra, sin ningún ajuste por variables de regresión. La ecuación anterior muestra que R^2_{adj} no aumenta necesariamente con una variable de explicación más. Si no hay mejoría en R^2_{adj} por la adición de una variable, que el término $\frac{(n-1)}{(n-p-1)}$ en realidad baja el R^2_{adj}. Por esta razón, se postula que el R^2 ajustado es una mejor medida que R^2 para la selección del modelo.

$$R^2_{adj} = 1 - \frac{(n-1)}{n-(p+1)}\left(1-R^2\right) \Leftrightarrow R^2_{adj} \leq R^2$$

2.2.3. Varianza Residual (s_R^2)

Para cada valor x_i de X, se obtiene una diferencia (el residuo) entre el valor observado de Y y el correspondiente valor teórico obtenido en el modelo de regresión. Por lo tanto se define la VARIANZA RESIDUAL como la media de todos los residuos elevados al cuadrado:

$$s_R^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = MCE$$

Donde MCE es la media cuadrática del error; un buen criterio de selección de variables explicativas es elegir el subconjunto de *"j"* variables que minimice el valor de MCE, siendo esta la varianza residual obtenida con el modelo de *"j"* variables de explicación.

Teniendo en cuenta que:

$$R_{adj}^2 = 1 - \frac{1}{s_y^2} MCE$$

Se puede deducir que:

$$R_{adj(p-1)}^2 > R_{adj.\,j}^2 \Leftrightarrow MCE_{p-1} < MCE_j$$

Por lo tanto el criterio de minimizar la varianza residual es equivalente al criterio de maximizar el coeficiente de determinación ajustado.

El R^2_{adj} representa la reducción (proporcional) en la varianza residual obtenidos por el modelo de regresión. Es así que en el momento de considerar la selección del mejor modelo, no solo se deben observar los indicadores sino que además el valor de la varianza residual la cual s^2_R. Es conveniente enfatizar que la varianza residual no se la considera como un indicador de selección de modelos, sino más bien como una guía para así determinar cuál de los indicadores es el que más conviene en el estudio de Regresión.

Se ha mencionado anteriormente que habrá más de un modelo fijo para (p-1) variables de explicación, en lugar de examinar todos estos modelos, se fijará la atención al mejor, por ejemplo, los mejores tres o cuatro modelos con mayores valores de $R^2_{adj.}$ y menores valores de s^2_R.

2.2.4. Estadístico C_p de Mallows

Los criterios previos se basan en la Suma Cuadrática del Error *"SCE"*, ahora se explicará un criterio que toma en cuenta la Media Cuadrática del Error (MCE, es decir la varianza del error) en la selección del modelo, lo que conlleva a que si se omite una variable explicativa importante que influya en la predicción, los estimadores de los coeficientes de regresión serían sesgados, es decir $E\left(\hat{\beta}_i\right) \neq \hat{\beta}_i$ lo cual indica que el objetivo de este indicador es minimizar la MCE, C_P de Mallows está definido como:

$$C_p = \frac{SCR_p}{s^2} - (n - 2p)$$

Donde p es el número de parámetros en un modelo de Regresión Lineal Múltiple, con (p – 1) el número de variables explicativas, s^2 es la varianza del error con todas las variables y SCE_p es la suma cuadrática del error al ir ajustando el modelo con *p* parámetros.

Para interpretar este estadístico, se define el Error Cuadrático Medio de predicción *"ECMP"* para los puntos observados cuando se utiliza un modelo con *"p"* parámetros como:

$$\mathrm{ECMP_p} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{y}_{p,i} - m_{p,i}\right)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\hat{y}_{p,i} - \mathrm{E}\left(\hat{y}_{p,i}\right) + \mathrm{E}\left(\hat{y}_{p,i}\right) - m_{p,i}\right)^2$$

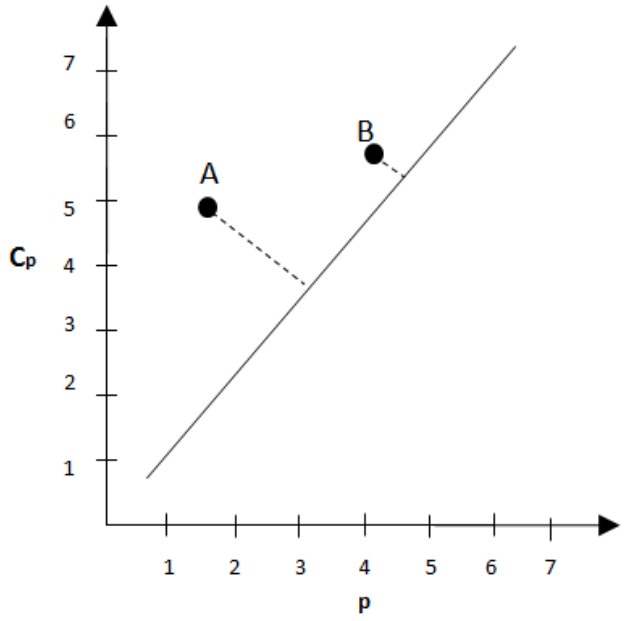
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{var}\left(\hat{y}_{p,i}\right) + Sesgo^2\left(\hat{y}_{p,i}\right)$$

Donde $\hat{y}_{p,i}$ es el valor ajustado cuando se utiliza el modelo con p parámetros y $\mathrm{m_{p,i}} = \mathrm{E}\left[\mathrm{y} \mid \mathrm{X} = \mathrm{x_{p,i}}\right]$ siendo un buen criterio de selección del modelo el de elegir el modelo que tenga el ECMP (Error Cuadrático Medio de Predicción) mínimo.

También se puede probar que en los modelos sin sesgo $C_p = p$. Por lo tanto, aquellos subconjuntos de "p-1" variables explicativas que tengan un $C_p \cong p = j + 1$ son los mejores. Se puede construir una gráfica de C_p para los diferentes subconjuntos que se quieren analizar frente a p. Y se considerarán buenos a aquellos subconjuntos que tienen C_p pequeño que $C_p = p$.

En la *"Figura 4"* se puede observar el gráfico C_p para dos puntos de variables explicativas y se observa que el punto A tiene un sesgo mucho mayor que el del subconjunto B, pero éste tiene menor C_p.

Figura 5: Representación Gráfica del Indicador C_P Mallows.
" Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

En estadística, C_p Mallows, llamado así por Colin Mallows, se utiliza a menudo como una regla de identificación para diversas formas de regresión paso a paso. Un punto a considerar es la colinealidad la cual en el análisis de regresión consiste en que las variables de explicación del modelo están relacionadas constituyendo así una combinación lineal. Este inconveniente resulta ser muy frecuente en los modelos de regresión. A menudo muchas de las variables independientes se esperaría que tengan efectos que son altamente correlacionados y no se puede estimar por separado. Cuando hay

demasiadas variables explicativas muchas de ellas cuyos coeficientes deben ser estimados, se han incluido en un modelo de regresión que se dice que está "sobre-ajustado." El peor caso es cuando el número de parámetros a estimar es mayor que el número de observaciones, por lo que no pueden ser estimadas en absoluto. El estadístico *"C_p"* se puede utilizar en la selección de un modelo reducido sin problema, tanto tiempo como *"S^2"* Error cuadrático Medio, es distinto de cero, lo que permite calcular *"C_p"*.

El modelo con parámetros *p*. Denotemos el error cuadrático medio de este modelo por *"S^2"*. Nosotros suponemos que el modelo más grande da una descripción adecuada, y por lo tanto $\mathrm{E}(S^2)=\sigma^2$.

Deteniéndose especialmente un modelo candidato con $q = p\text{-}1$ variables explicativas, p ≤ q y p escrito como parámetros $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Cuando $\mathbf{X}_1$ contiene **1** (la columna de unos) y los vectores (p-1) variables explicativas. Si este modelo más pequeño ya es adecuado, entonces:

$$\frac{\mathrm{SCE}_p}{\sigma^2} \sim \chi^2_{n\text{-}p\text{-}1}$$

Los modelos bajo supuestos de normalidad e independencia estocástica, que se consideran más opcionales son aquellos con pocas variables y $C_p \cong p$. Una vez se haya encontrado ese modelo, no hay necesidad de emplear un modelo más complicado que involucra a más de (*p-1)* variables.

Se concluye que el mejor modelo es aquel que no tiene falta de ajuste (“underfitting”) ni alto sobreajuste (“overfitting”) en los datos.

Falta de ajuste, se da cuando el estimado del valor predicho de la variable de respuesta tiene *alto sesgo y poca varianza*,

Sobreajuste, se da cuando *la varianza* del estimado del valor predicho es *alta*, pero el *sesgo es bajo*.

2.2.5. Criterio de Información Akaike (AIC)

El indicador AIC derivado del denominado Criterio de Información Akaike, otra medida de bondad de ajuste y de un modelo de Regresión; fue desarrollado por el científico Japonés Hirotsugu Akaike y publicado por primera vez bajo el nombre de “criterio de

información”, se basa en la entropía de la información, el cual ofrece una medida relativa de la pérdida de información cuando un determinado modelo se utiliza para describir la realidad.

El AIC no es una prueba del modelo en el sentido de las pruebas de hipótesis, sino que proporciona un medio para la comparación entre modelos, un criterio para la selección del modelo.

Dado un conjunto de datos, varios posibles modelos pueden ser clasificados de acuerdo a su AIC, los modelos con valores más pequeños de la AIC son los preferidos.

Así se define el AIC como:

$$AIC_p = n\left[\ln\left(\frac{SCE_p}{n}\right)\right] + 2(p+1)$$

El primer término en la expresión anterior es, como en la C_p de Mallows, una medida de bondad de ajuste (disminuye al crecer el de la estimación por máxima de la verosimilitud); el segundo penaliza el número de parámetros.

El segundo término, 2(p+1), representa una función que aumenta, con el número de parámetros estimados.

2.2.6. Suma de Cuadrados de Predicción (PRESS)

Este indicador de calidad de los modelos de regresión fue propuesto por Allen en 1974, de una combinación de todas las regresiones posibles, basado en el análisis de residuales y validación cruzada, la cual consiste en estimar los modelos con una muestra (muestra de entrenamiento o aprendizaje) y evaluarlos examinando su comportamiento en la predicción de otra diferente (muestra de validación). Supongamos que hay *p* parámetros en el modelo y que tenemos *"n"* observaciones disponibles para estimar los parámetros del modelo, en cada paso se deja de lado la i-ésima observación del conjunto de datos y se calculan todas las regresiones posibles; se calcula la predicción y el residual correspondiente para la observación que no fue incluida, el cual es llamado el residual "*PRESS".*

Se puede expresar esta medida:

$$e_i = \frac{e_i}{1 - h_{ii}}$$

como una función de los residuales ordinarios $e_i = y_i - \hat{y}_i$ y los términos de apalancamiento h_{ij} del modelo de regresión original.

Siendo $1 - h_{ii}$ parte de la Suma cuadrática del error, visto en el capítulo anterior.

Donde la medida de Sumas Cuadradas de Predicción *"PRESS"* para el modelo de regresión que contiene *"p"* parámetros se define por:

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)}^2$$

O equivalente a

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

En conclusión se dice que el mejor modelo entre varios es aquel que tiene el menor valor del índice *"PRESS"*.

CAPÍTULO 3

3. ACERCA DE ERLA

3.1. Introducción

ERLA es un software desarrollado para ser implementado en Microsoft Windows, para el cual se utilizó Visual Basic.NET y Matlab.

La utilización básica de estos dos programas es Visual Basic.NET para la presentación de la interfaces de interacción con el usuario y Matlab para el desarrollo de las funciones matemáticas y estadísticas.

En este capítulo se explica paso a paso el desarrollo de ERLA como se enlazan Visual y Matlab, las funciones a utilizar y un detalle de cada uno de estos dos programas indispensables para la realización de ERLA.

3.2. Lenguaje y Códigos

3.2.1. MATLAB

MATLAB (Laboratorio de Matrices) es un programa interactivo de uso general. Es un instrumento computacional simple, versátil y de gran poder para aplicaciones numéricas, simbólicas y gráficas que contiene una gran cantidad de funciones predefinidas para aplicaciones en ciencias e ingeniería. Los objetos básicos con los cuales opera MATLAB son matrices.

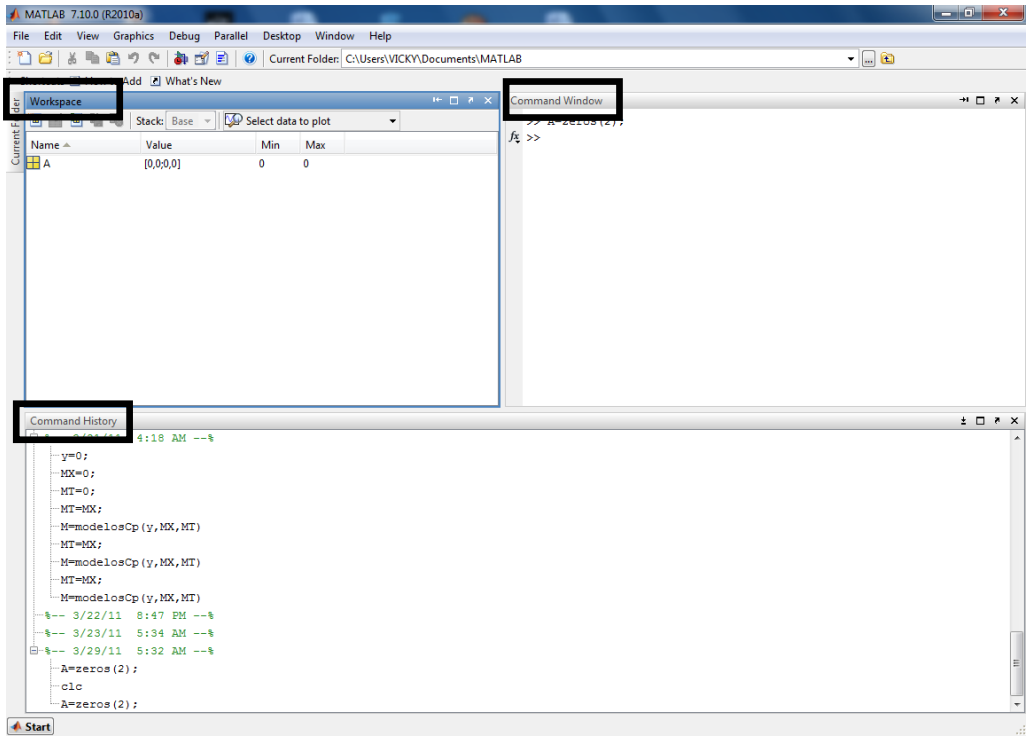
El entorno de MATLAB está organizado mediante ventanas. Las principales son:

Command Window Es la ventana de comandos para interactuar.

Command History Contiene el registro de los comandos que han sido ingresados.

Workspace Contiene la descripción de las variables usadas en cada sección.

Figura 6: Entorno Gráfico de MATLAB.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

El símbolo ">>" indica que el programa está listo para recibir las instrucciones.

MATLAB es un programa de "cálculo numérico" orientado a matrices tal como es lo requerido en la aplicación de las técnicas estadísticas desarrolladas en ERLA. El algoritmo utilizado para construir la Función "Regresión Lineal" se presenta en la *Figura 7*.

Figura 7: Función "Regresión Lineal".
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

```
function R1=RegressionCoefficients(y,MX)
%El primer argumento debe ser la variable a ser explicada
%El segundo argumento debe ser la matriz con variables de
explicación
%Devuelve una matriz con las inferencias sobre los betas
paramat long g;
d=size(MX);
n=d(1);
p=d(2)+1;
j=ones(n,1);
X=[j,MX];
I=eye(n);
J=ones(n);
A=inv(X'*X);
H=X*A*X';
SCE=y'*(I-H)*y;
MCE=SCE/(n-p);
b=A*X'*y;
Sb=MCE*A;
R1=zeros(p,4);
para i=1:p
   R1(i,1)=b(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-p);
   R1(i,4)=(1-R1(i,4))*2;
 fin
```

Autoría: Macías S. – Pincay C.

Con esta función se obtienen los coeficientes de Regresión Lineal, los argumentos de entrada o datos de entrada son la variable a ser explicada y la matriz con las variables de explicación. Los resultados obtenidos luego de la ejecución de dicha función son los coeficientes de para los estimadores de los parámetros del vector β.

Figura 8: Funciones para "Selección de Modelos" - R^2 Ajustado.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

```
función M=modelosR2(y,MX)
t1=size(MX);
v=t1(2);
SCT=R2Ajustado2_SCT(y,MX);
para i=1:v
   c(i)=nchoosek(v,i);
fin
p=1;
i=1;
k=c(1);
t=0;
si v==1
   M(t+1)=R2 Ajustado2(y,MX,SCT);
   M=M';
Si no
   mientras i<v
      cc=1;
      vr=combinacion(v,i,'c');
      para j=p:k
         M(j)=R2 Ajustado2(y,MX(:,vr(cc,:)),SCT);
         t=j;
         cc=cc+1;
      fin
   p=t+1;
   i=i+1;
   k=t+c(i);
   fin
   vr=combinator(v,v,'c');
   M(t+1)=R2 Ajustado2(y,MX,SCT);
   M=M';
Fin
```

Esta función tiene como argumentos la variable dependiente ***y*** y la matriz de datos ***MX***. Posteriormente se realiza un bucle, para obtener todas las combinaciones posibles entre las variables explicativas (***MX)***.

Se ejecuta otra función llamada R2Ajustado2, previamente diseñada por el usuario y finalmente se almacena en un vector llamado **M**, para luego ser usado en Visual Net.

Autoría: Macías S. – Pincay C.

La descripción de la función "modelosR2(y,MX)" detallada en la Figura 8, para R2 Ajustado, es la misma para la función "modelosAIC(y,MX)" que se refiere al indicador Akaike, "modelosCp(y,MX,MT)" para Cp Mallows y "modelosPRESS(y,MX)" para PRESS. Todas estas funciones siguen la misma estructura.

3.2.2. VISUAL. NET

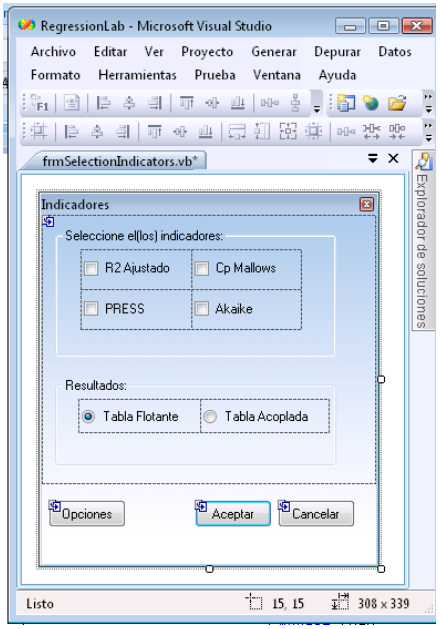
Microsoft Visual Studio es un entorno de desarrollo integrado (IDE) para sistemas operativos Windows. Soporta varios lenguajes de programación tales como Visual C++, Visual C#, Visual J#, ASP.NET y Visual Basic .NET.

Visual Studio permite a los desarrolladores crear aplicaciones, sitios y aplicaciones web, así como servicios, además de que intercomuniquen entre estaciones de trabajo, páginas web y dispositivos móviles.

Para el caso de ERLA, el funcionamiento en este entorno se presenta en la ***"Figura 9".*** En el primer recuadro se tiene la interfaz gráfica del formulario de Selección de Modelos, en el segundo está el Pseudocódigo de Programación y en el último recuadro están las funciones con las cuales se realiza la comunicación con las funciones previamente creadas en Matlab.

Figura 9: Programación en Visual para "Selección de Modelos".
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

Interfaz Gráfica
"Modelos de selección"



Pseudocodigos De Programación
"Modelos de selección"

```
Public Class frmSelectionIndicators

    Private Sub btnAceptar_Click(ByVal sfiner As
System.Object, ByVal e As System.EventArgs) Handles
btnAceptar.Click
        frmModelSelection.R2Aj = R2Ajus.Seleccionar
        frmModelSelection.Cp = CPM.Seleccionar
        frmModelSelection.AIC = AK.Seleccionar
        frmModelSelection.PR = PRS.Seleccionar
        frmModelSelection.OPA = opcTablaA.Seleccionar
       frmModelSelection.OPF = opcTablaF.Seleccionar
       frmModelSelection.btnAceptar.Enabled = True
    Fin Sub

    Private Sub frmSelectionIndicators_Load(ByVal sfiner
As System.Object, ByVal e As System.EventArgs)
Handles MyBase.Load
        R2Ajus.Seleccionar = False
        CPM.Seleccionar = False
        AK.Seleccionar = False
        PRS.Seleccionar = False
        opcTablaA.Seleccionar = True
    Fin Sub

Fin Class
```

Funciones en Visual Net para la comunicación con Matlab
"Modelos de selección"

```
Public Function VSAkaike(ByVal Y As MWNumericArray, ByVal X As MWNumericArray) As MWArray
        mwa = mva.modelosAIC(Y, X)
        Return mf.RoundTo(mwa, prec)
    Fin Function
    Public Function VSR2Ajustado(ByVal Y As MWNumericArray, ByVal X As MWNumericArray) As
MWArray
        mwa = mva.modelosR2(Y, X)
        Return mf.RoundTo(mwa, prec)
    Fin Function
    Public Function VSMallows(ByVal Y As MWNumericArray, ByVal X As MWNumericArray, ByVal XT As
MWNumericArray) As MWArray
        mwa = mva.modelosCp(Y, X, XT)
        Return mf.RoundTo(mwa, prec)
    Fin Function
    Public Function VSPRESS(ByVal Y As MWNumericArray, ByVal X As MWNumericArray) As MWArray
        mwa = mva.modelosPRESS(Y, X)
        Return mf.RoundTo(mwa, prec)
    Fin Function
```

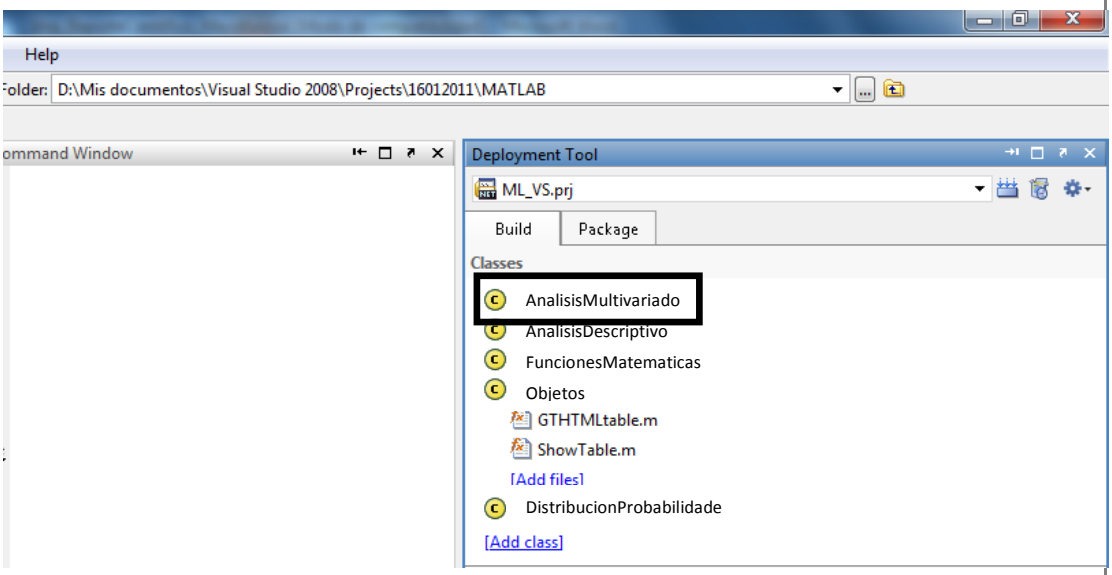
Autoría: Macías S. – Pincay C.

3.3. Conexión entre VISUAL BASIC.NET y MATLAB

La conexión entre estos dos programas comienza en Matlab con la creación de las librerías respectivas, ya que ésta es la base para la creación de las funciones que proporcionaran los resultados esperados.

Para ello inicialmente se crean funciones (***ver Figura 8 o 9)***, para luego de las comprobaciones respectivas de dichas funciones, se crean librerías (archivos *.dll), dichos archivos son un comprimido de las funciones creadas previamente, en la ***"Figura 10"***, se observa la creación de las librerías.

Figura 10: Creación de Archivos *.dll.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

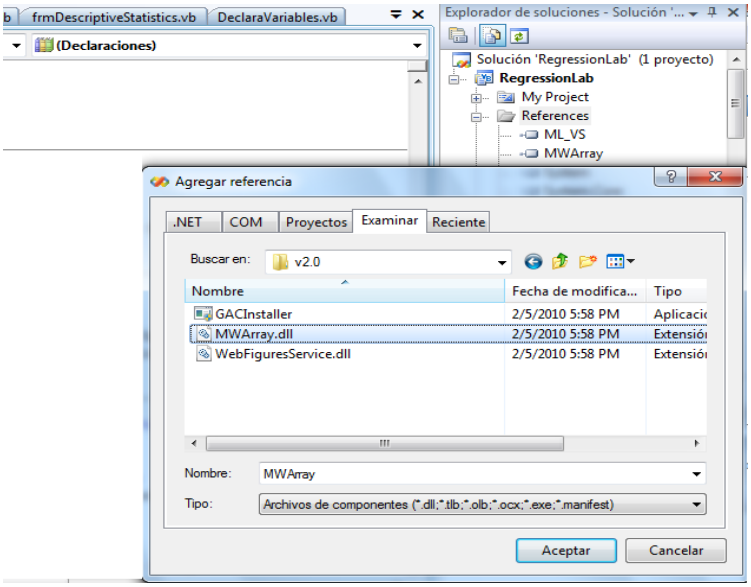


Autoría: Macías S. – Pincay C.

En la opción ***"Classes"*** se van creando las categorías dentro de las cuales se quiera organizar las funciones, para este caso se tienen las clases de Análisis Multivariado, Análisis Descriptivo, Funciones Matemáticas, Objetos para mostrar y Distribución de Probabilidades, luego se procede a compilar estos archivos, presionando el botón  y con esto se crean las librerías y archivo *.prj ***(Nombre Proyecto)***.

Ya desde Visual Basic.NET, se añade una referencia hacia la librería principal de Matlab MWArray.dll, para con esto poder acceder a las funciones creadas en Matlab convertidas en librerías.

Figura 11: Añadir Referencia en Visual Basic .NET.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

El proyecto desarrollado en Visual Studio.NET se lo compila para luego poder tener un archivo ejecutable ***(*.exe)***, con el cual este software podrá ser instalado en sistemas operativos Windows.

CAPÍTULO 4

4. VALIDACIÓN DEL MODELO EN EL SOFTWARE “ERLA”

4.1. Introducción

Una de las etapas que se deben llevar a cabo en el desarrollo de un nuevo software es la validación o comprobación de sus resultados, mediante pruebas de las funcionalidades.

En este capítulo se efectuará pruebas para el modelo regresión simple, múltiple, y para los indicadores de selección de modelos, vistos en el Capítulo 2. Para dicha validación se consideraran tres casos: Pruebas de Tensión Sistólica, Importaciones de cierto producto y el caso de una Central Hidroeléctrica. Cada caso será detallado en las secciones posteriores.

En estas pruebas se realizan simulaciones para el mismo número de observaciones en cada caso, y se obtendrá de una cantidad determinada de simulaciones los estimadores respectivos.

4.2. Validación para el Modelo de Regresión Lineal Simple

En esta validación de regresión lineal simple se considera el estudio de la tensión sistólica, el mismo que consistió en tomar la tensión sistólica y la edad a un grupo de 69 pacientes. Lo que se busca es determinar la influencia de la Edad en la tensión sistólica de los pacientes. La Tabla 3 indica los datos de estas dos variables.

Tabla 3: Tensión Arterial Sistólica y Edad de 69 pacientes.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

Nº	Tensión Sistólica	Edad
1	114	17
2	134	18
3	124	19
4	128	19
5	116	20
6	120	21
7	138	21
8	130	22
9	139	23
10	125	25
11	132	26
12	130	29
13	140	33
14	144	33
15	110	34
16	148	35
17	124	36
18	136	36
19	150	38
20	120	39
21	144	39
22	153	40
23	134	41
24	152	41

Nº	Tensión Sistólica	Edad
25	158	41
26	124	42
27	128	42
28	138	42
29	142	44
30	160	44
31	135	45
32	138	45
33	142	46
34	145	47
35	149	47
36	156	47
37	159	47
38	130	48
39	157	48
40	142	50
41	144	50
42	160	51
43	174	51
44	156	52
45	158	53
46	174	55
47	150	56
48	154	56

Nº	Tensión Sistólica	Edad
49	165	56
50	164	57
51	168	57
52	140	59
53	170	59
54	185	60
55	154	61
56	169	61
57	172	62
58	144	63
59	162	64
60	158	65
61	162	65
62	176	65
63	176	66
64	158	67
65	170	67
66	172	68
67	184	68
68	175	69
69	180	70

Autoría: Macías S. – Pincay C.

Para este ejemplo la variable dependiente o variable respuesta será la *Tensión Sistólica* y la variable explicativa es *Edad* y el número de observaciones es: n = 69. La Tabla 4 contiene las estadísticas básicas de dichas variables, lo cual se realiza para observar el comportamiento básico de las variables.

Tabla 4: Estadísticas básicas de las variables "Tensión Sistólica" y "Edad"
Caso: "Regresión Lineal Simple".
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

Estadísticas	Tensión Sistólica (y)	Edad (x)
Media	148.72±2.22	46.13±1.82
Error Estándar	2.22	1.82
Desviación Estándar	18.48	15.08
Mínimo	110.00	17.00
Cuartil 1	134.50	36.00
Mediana	149.00	47.00
Cuartil 3	162.00	59.00
Máximo	185.00	70.00
Moda	144, 15	47.00
Sesgo	-0.02	-0.31

Autoría: Macías S. – Pincay C.

Aplicando el modelo de Regresión Lineal Simple para el ejemplo de la Tensión Sistólica dicha ecuación es la siguiente:

$$\hat{y} = 103.35 + 0.98\,x \qquad \textbf{(4.1)}$$

De este modelo se determina la tabla ANOVA, como sigue:

Tabla 5: Tabla de Análisis de Varianza (ANOVA) de las variables "Tensión Sistólica" y "Edad" Caso: "Regresión Lineal Simple".
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	1	14965.312	14965.312	121.589
Error	67	8246.456	123.081	
Total	68	23211.768		

R^2x100 : 64.5%
Valor *p* : 0.00

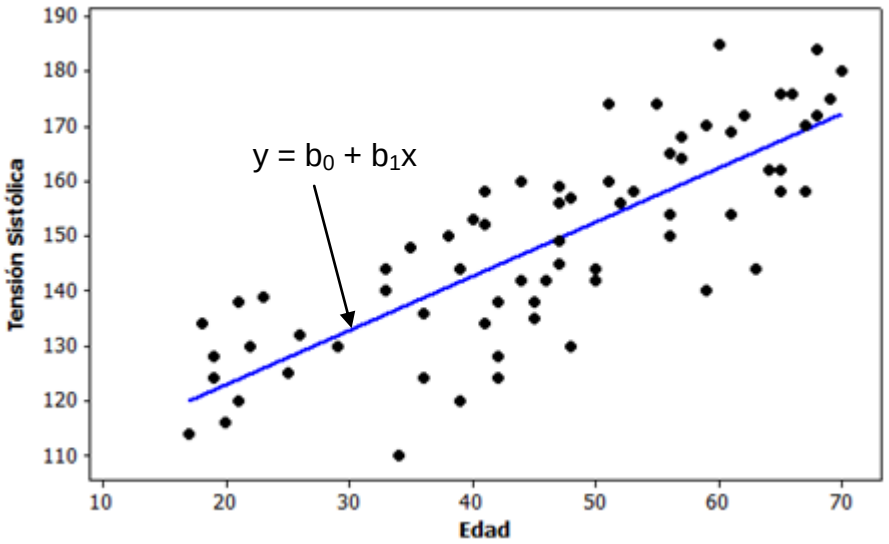
Prueba *t* de β_0, β_1

ESTIMADORES	T	VALOR *p*	INTERVALOS DE CONFIANZA (95%)
$\hat{\beta}_0$	23.891	0.00	$94.718 \leq \hat{\beta}_0 \leq 111.988$
$\hat{\beta}_1$	11.027	0.00	$0.806 \leq \hat{\beta}_1 \leq 1.162$

Autoría: Macías S. – Pincay C.

La Figura 12 representa la Gráfica de dispersión de los datos de la Tensión sistólica versus la Edad de los pacientes y la recta de regresión dada en la ecuación **(4.1)**. Se puede observar que las variables tienen tendencia rectilínea en X, es decir es adecuado formular el modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ de Regresión Lineal Simple.

Figura 12: Gráfica de dispersión de las variables "Tensión Sistólica" vs. "Edad".
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Autoría: Macías S. – Pincay C.

De acuerdo con la ecuación **(4.1)** los estimadores de los betas son $\hat{\beta}_0 = 103.35$ y $\hat{\beta}_1 = 0.98$.

Para iniciar la validación se realizarán simulaciones para lo cual se tomarán 30 muestras de tamaño n = 69 en la cual se supone el error ~N(0,1), por lo tanto en cada simulación con se obtendrán estimadores para los **β**. La Tabla 6 presentan los estimadores de b_0, b_1, s_{b_0} y s_{b_1} de los cuales se busca observar su comportamiento para la validación del modelo de Regresión Lineal Simple en el Software ERLA.

Tabla 6: Estimadores de parámetros Betas. Muestra: 30, n=69 y e ~N(0,1).
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

Muestra	Estimadores	
	$b_0 \pm S_{b_0}$	$b_1 \pm S_{b_1}$
1	102.7937 ± 4.3636	0.9933 ± 0.0900
2	103.4102 ± 4.3472	0.9803 ± 0.0896
3	103.4934 ± 4.3161	0.9830 ± 0.0890
4	103.5848 ± 4.1525	0.9752 ± 0.0856
5	103.9804 ± 4.4466	0.9699 ± 0.0917
6	103.5431 ± 4.2833	0.9832 ± 0.0883
7	103.1418 ± 4.2788	0.9886 ± 0.0882
8	103.4336 ± 4.3411	0.9824 ± 0.0895
9	102.5098 ± 4.3688	0.9999 ± 0.0901
10	103.0473 ± 4.3713	0.9902 ± 0.0901
11	103.4148 ± 4.3913	0.9817 ± 0.0905
12	103.9191 ± 4.3659	0.9737 ± 0.0900
13	102.6607 ± 4.2954	0.9954 ± 0.0886
14	102.7466 ± 4.3905	0.9946 ± 0.0905
15	103.2120 ± 4.2940	0.9897 ± 0.0885
16	102.7792 ± 4.2933	0.9946 ± 0.0885
17	103.0995 ± 4.3213	0.9873 ± 0.0891
18	103.9296 ± 4.3092	0.9731 ± 0.0889
19	103.5879 ± 4.3672	0.9828 ± 0.0900
20	103.6638 ± 4.2941	0.9791 ± 0.0885
21	102.8549 ± 4.3581	0.9933 ± 0.0899
22	103.0017 ± 4.3074	0.9909 ± 0.0888
23	102.5257 ± 4.3514	0.9987 ± 0.0897
24	103.7928 ± 4.3679	0.9742 ± 0.0901
25	103.0982 ± 4.3676	0.9882 ± 0.0901
26	102.8532 ± 4.3346	0.9957 ± 0.0894
27	103.8882 ± 4.3264	0.9734 ± 0.0892
28	102.8559 ± 4.2433	0.9929 ± 0.0875
29	102.5022 ± 4.3131	1.0008 ± 0.0889
30	103.8310 ± 4.2561	0.9698 ± 0.0878

Autoría: Macías S. – Pincay C.

En la Tabla 7 se tienen las estadísticas básicas de los estimadores (b_0 y b_1). El estimador de β_1 presenta sesgo pequeño hacia la derecha.

Tabla 7: Estadísticas Básicas de los Estimadores de los parámetros Betas.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

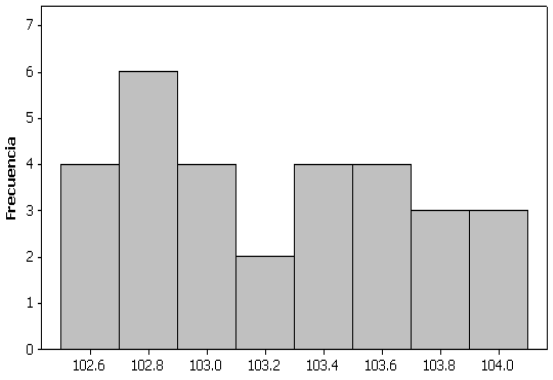
Parámetro	**β_0**	**β_1**
Estadísticas	**(b_0)**	**(b_1)**
Media	103.24 ± 0.47	0.99 ± 0.00
Error Estándar	0.09	0.00
Desviación Estándar	0.47	0.01
Mínimo	102.50	0.97
Cuartil 1	102.84	0.98
Mediana	103.18	0.99
Cuartil 3	103.61	0.99
Máximo	103.98	1.00
Sesgo	0.03	-0.21

Autoría: Macías S. – Pincay C.

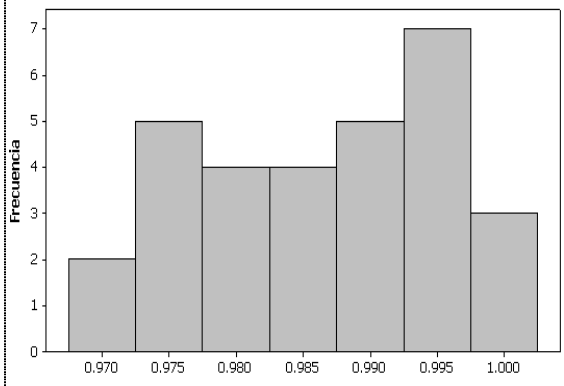
De acuerdo con la Tabla 12, el estimador b_0 tiene sesgo hacia la izquierda en tanto que b_1 tiene el sesgo hacia derecha. En la Figura 13 se observa el histograma de Frecuencias y Diagrama de Cajas de b_0, b_1, s_{b_0} y s_{b_1}.

Figura 13: Histogramas de Frecuencias y Diagramas de Cajas de b_0, b_1, s_{b_0} y s_{b_1}

"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



Histograma de Frecuencias b_0



Histograma de Frecuencias b_1

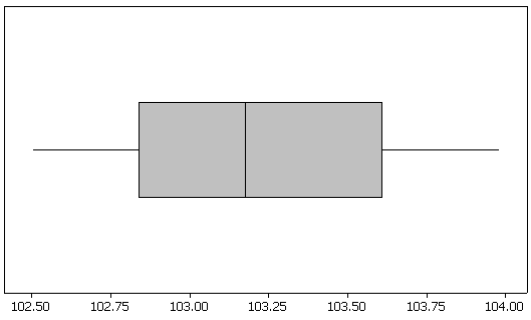


Diagrama de Cajas b_0

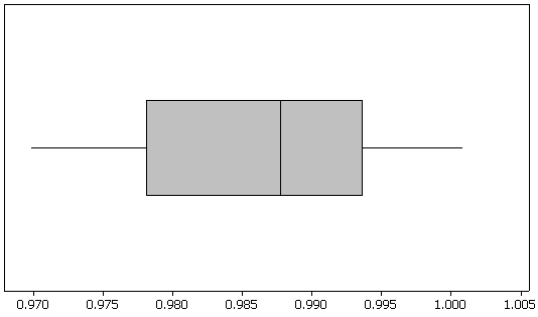
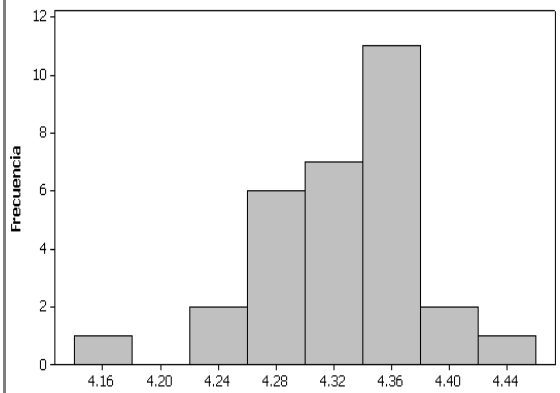
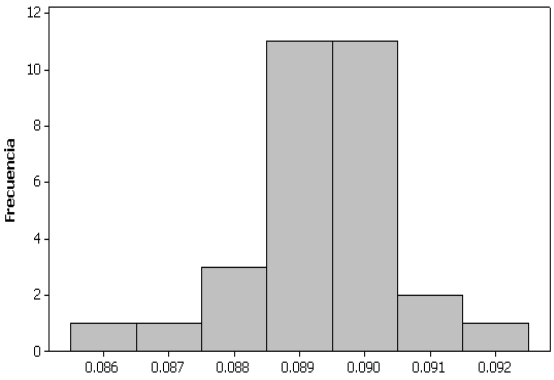


Diagrama de Cajas b_1



Histograma de Frecuencias s_{b_0}



Histograma de Frecuencias s_{b_1}

Figura 13: Histogramas de Frecuencias y Diagramas de Cajas de b_0, b_1, s_{b_0} y s_{b_1}

"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"

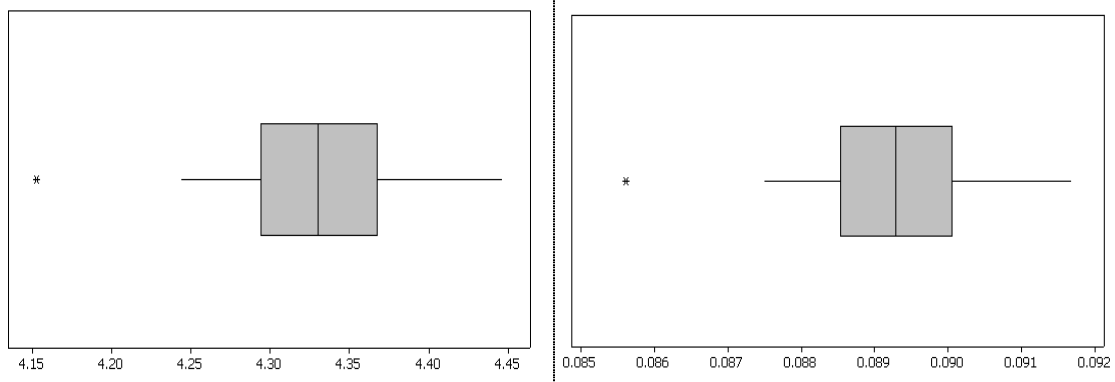


Diagrama de Cajas s_{b_0}

Diagrama de Cajas s_{b_1}

Autoría: Macías S. – Pincay C.

4.3. Validación para el Modelo de Regresión Lineal Múltiple

Para el caso de la validación de Regresión Lineal Múltiple el ejemplo que se considerará es el de Importaciones de cierto producto en el lapso de 41 años. Las variables que se analizan son Importaciones, Precio Relativo y PIB Real. El modelo de Regresión utilizado es: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Tabla 8
"Selección de Modelos y Pruebas de Homocedasticidad"
Estadísticas básicas de las variables "Importaciones", "Precio Relativo" y "PIB real"
Caso: "Regresión Lineal Múltiple".

Estadísticas	Importaciones Reales	Precio Relativo	PIB Real
Media	391.70	1.54	2771.00
Error Estándar	28.10	0.06	175.00
Desviación Estándar	179.80	0.41	1120.00
Mínimo	152.90	0.92	1049.00
Cuartil 1	268.10	1.08	1744.00
Mediana	334.30	1.58	2940.00
Cuartil 3	502.10	1.78	3452.00
Máximo	882.20	2.35	5073.00
Sesgo	1.16	0.12	0.22

Autoría: Macías S. – Pincay C.

Para la variable "Importaciones" el modelo de Regresión Lineal Múltiple es:

$$\hat{y} = 207.01 - 238.66\,x_1 + 0.19\,x_2 \qquad \textbf{(4.2)}$$

Con estos datos y con el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ se concluye la Tabla ANOVA que se muestra en la Tabla 9.

Tabla 9
"Selección de Modelos y Pruebas de Homocedasticidad"
Tabla de Análisis de Varianza (ANOVA) de las variables "Importaciones", "Precio Relativo" y "PIB real"
Caso: "Regresión Lineal Múltiple".

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMAS CUADRÁTICAS	MEDIAS CUADRÁTICAS	F
Regresión	2	1153267.916	576633.958	156.872
Error	38	139681.774	3675.836	
Total	40	1292949.690		

R^2x100 : 89.2%
Valor *p* : 0.00

Prueba *t* de β_0, β_1, β_2

ESTIMADORES	t	VALOR *p*	INTERVALOS DE CONFIANZA (95%)
$\hat{\beta}_0$	5.551	0.00	131.522 ≤ $\hat{\beta}_0$ ≤ 282.504
$\hat{\beta}_1$	-7.291	0.00	-304.920≤ $\hat{\beta}_1$ ≤ -172.394
$\hat{\beta}_2$	16.611	0.00	0.175 ≤ $\hat{\beta}_2$ ≤ 0.224

Autoría: Macías S. – Pincay C.

Para esta prueba se tomaron 30 muestras de tamaño n=41 al igual que en caso de regresión Lineal simple, con (error~N(0,1)).

Tabla 10
"Selección de Modelos y Pruebas de Homocedasticidad"
Estimadores de parámetros Betas. Muestra: 30, n=41 y e ~N(0,1)
Caso: "Regresión Lineal Múltiple".

Muestras	Estimadores		
	$b_0 \pm S_{b_0}$	$b_1 \pm S_{b_1}$	$b_2 \pm S_{b_2}$
1	206.2396 ± 37.1291	-237.7555 ± 32.5905	0.1995 ± 0.0120
2	205.8952 ± 37.3231	-237.4717 ± 32.7608	0.1994 ± 0.0120
3	206.6026 ± 37.3081	-238.4039 ± 32.7476	0.1997 ± 0.0120
4	207.8778 ± 37.2602	-239.2150 ± 32.7055	0.1996 ± 0.0120
5	206.6426 ± 37.2191	-238.4980 ± 32.6694	0.1997 ± 0.0120
6	207.3172 ± 37.1855	-238.9681 ± 32.6399	0.1999 ± 0.0120
7	207.3322 ± 37.2727	-239.0078 ± 32.7165	0.1998 ± 0.0120
8	205.8918 ± 37.2992	-237.9643 ± 32.7398	0.1997 ± 0.0120
9	206.6168 ± 37.1970	-238.3380 ± 32.6500	0.1996 ± 0.0120
10	207.1935 ± 37.2971	-238.5735 ± 32.7379	0.1996 ± 0.0120
11	208.4345 ± 37.2685	-240.1310 ± 32.7128	0.2000 ± 0.0120
12	206.8149 ± 37.1011	-238.0958 ± 32.5659	0.1995 ± 0.0120
13	206.8391 ± 37.2692	-238.4686 ± 32.7134	0.1997 ± 0.0120
14	207.2050 ± 37.3970	-238.7416 ± 32.8256	0.1997 ± 0.0121
15	207.4255 ± 37.3151	-238.6495 ± 32.7537	0.1995 ± 0.0120
16	206.6882 ± 37.3861	-238.8142 ± 32.8161	0.1999 ± 0.0121
17	206.9769 ± 37.2090	-238.2632 ± 32.6606	0.1995 ± 0.0120
18	206.2779 ± 37.4375	-237.0636 ± 32.8611	0.1992 ± 0.0121
19	206.5265 ± 37.1468	-238.5819 ± 32.6059	0.1998 ± 0.0120
20	207.4963 ± 37.3654	-239.7261 ± 32.7979	0.2001 ± 0.0120
21	207.4525 ± 37.2111	-238.9007 ± 32.6624	0.1997 ± 0.0120
22	207.2845 ± 37.4197	-238.4083 ± 32.8455	0.1995 ± 0.0121
23	206.5542 ± 37.1987	-238.7233 ± 32.6516	0.1998 ± 0.0120
24	207.3626 ± 37.2578	-239.5884 ± 32.7034	0.2000 ± 0.0120
25	206.3897 ± 37.3211	-238.6605 ± 32.7590	0.1999 ± 0.0120
26	207.7043 ± 37.3936	-239.0963 ± 32.8226	0.1996 ± 0.0121
27	207.1466 ± 37.2034	-239.1195 ± 32.6557	0.1999 ± 0.0120
28	206.6802 ± 37.3382	-238.1802 ± 32.7740	0.1996 ± 0.0120
29	206.6108 ± 37.3503	-237.7961 ± 32.7846	0.1995 ± 0.0120
30	207.3635 ± 37.3267	-239.0962 ± 32.7639	0.1997 ± 0.0120

Autoría: Macías S. – Pincay C.

En la Tabla 11 se muestran las Estadísticas Básicas de los estimadores de los betas (b_0, b_1 y b_2) se observa que la desviación estándar del estimador b_2 es prácticamente cero.

Tabla 11
"Selección de Modelos y Pruebas de Homocedasticidad"
Estadísticas Básicas de los Estimadores de los parámetros Betas
Caso: "Regresión Lineal Múltiple".

Parámetro	**β_0**	**β_1**	**β_2**
Estadísticas	**(b_0)**	**(b_1)**	**(b_2)**
Media	206.96	-238.61	0.20
Error Estándar	0.10	0.11	0.00
Desviación Estándar	0.58	0.65	0.00
Mínimo	205.89	-240.13	0.19
Cuartil 1	206.59	-239.03	0.20
Mediana	206.91	-238.62	0.20
Cuartil 3	207.36	-238.24	0.20
Máximo	208.43	-237.06	0.20
Sesgo	0.26	0.04	-0.04

Autoría: Macías S. – Pincay C.

4.4. Validación para los Indicadores de Selección de Modelos: R^2 Ajustado, C_p Mallows, Akaike Y PRESS.

En esta subsección como datos para la validación de los indicadores de selección, se considera el caso de una "Central Eléctrica".

Las variables que se consideran son:

C: Costo en dólares

D: Fecha de expedición permiso de construcción

T1: Tiempo entre la solicitud de permiso y la expedición o permiso

T2: Tiempo entre la emisión de la licencia de funcionamiento y permiso de construcción

S: Capacidad de Energía neta de la planta

PR: Existencia previa de un reactor en el mismo sitio.

NE: Planta construida en la región noreste

CT: Uso de la torre de enfriamiento

BW: Sistema de suministro de vapor nuclear

N: Número acumulado de plantas de energía

PT: Llave de plantas

El número de observaciones son n=32 y la variable dependiente para el modelo de Regresión es el Costo en dólares ***(C).***

De acuerdo con la ejecución de ERLA, basados en el ejemplo antes mencionado se determinó el valor del R^2 Ajustado, C_p Mallows, Akaike y PRESS de las 1024 combinaciones de las 10 variables de explicación (11 parámetros). Ver Tabla 12.

Tabla 12
"Selección de Modelos y Pruebas de Homocedasticidad"
Valores de los Indicadores R^2 Ajustado, C_p Mallows, Akaike y PRESS – De las 1024 Combinaciones de las diez Variables de Explicación (Once Parámetros).

# Parámetros	R^2 Ajustado	C_p Mallows	AIC	PRESS	# Variables Explicativas
2	0.4364	55.91	-78.68	4.38	**1**
3	0.6314	27.04	-91.36	2.76	**2**
4	0.7326	13.16	-100.75	1.81	**3**
5	0.7814	7.29	-106.36	1.60	**4**
6	0.7980	6.05	-108.10	1.60	**5**
7	0.8068	5.97	-108.77	1.67	**6**
8	0.8065	7.04	-108.03	1.75	**7**
9	0.8149	8.49	-108.81	1.91	**8**
10	0.8072	9.05	-106.93	2.05	**9**
11	0.7985	11.00	-105.014	2.32	**10**

R^2 Ajustado: 8 V.E. (0.8149)
C_p Mallows: 5 V.E. (6.0500**)**
AIC: 8 V.E. (-108.81**)**
PRESS: 4 V.E. (1.6000)

Autoría: Macías S. – Pincay C.

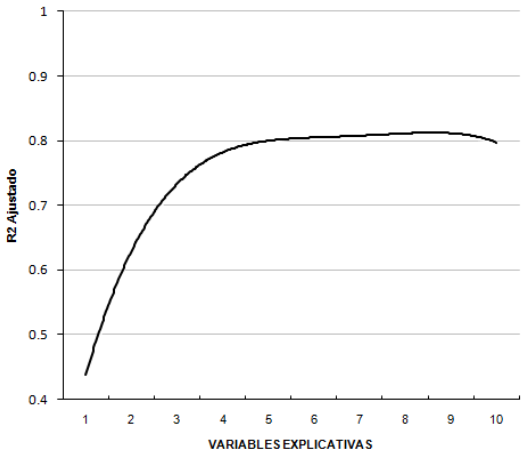
En la Tabla 12 se observa la cantidad de variables de explicación que en mejor grado explican a la variable respuesta "*y*" y por ende se tendría el mejor Modelo de Regresión Lineal. El R^2 Ajustado propone que sean 8 las variables explicativas: **(D, T2, S, PR, NE, CT, N, PT)** donde el modelo seria:

$$C = -11.68 + 0.24\,D + 0.006\,T_2 + 0.001\,S$$
$$- 0.11\,PR + 0.26\,NE + 0.11\,CT - 0.01\,N - 0.21\,PT$$

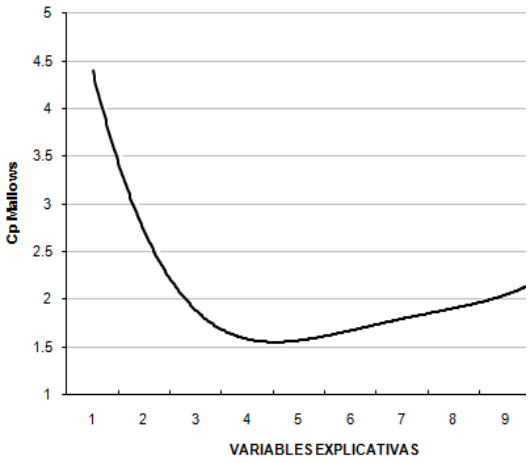
Y con este se obtiene un R^2 Ajustado de 0.8149. En el caso del Akaike se tiene igual cantidad de variables que el R^2 Ajustado y la misma combinación las variables de explicación. **(D, T2, S, PR, NE, CT, N, PT).**

Para determinar cuál es el comportamiento o tendencia de dichos indicadores, en la Figura 13 se presentan las gráficas de tendencias.

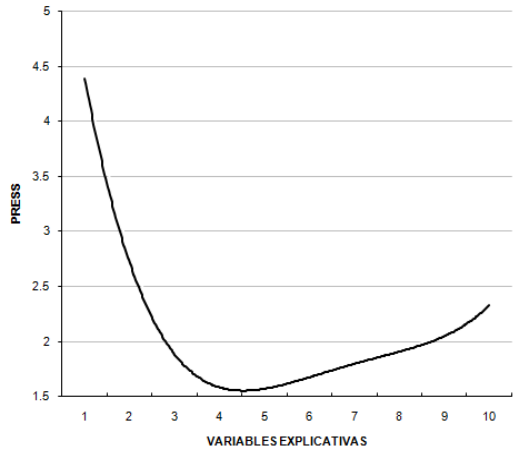
Figura 14: Graficas de Tendencia de los indicadores de Selección de Modelos: R^2 Ajustado, C_p Mallows, Akaike y PRESS.
"Construcción de Software para Regresión El Caso de Selección del Modelo y Pruebas de Homocedasticidad"



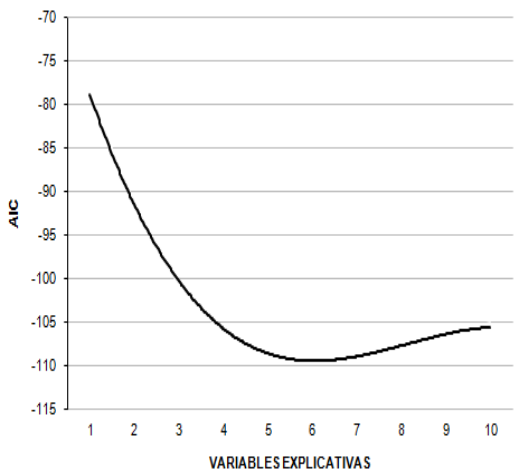
a) V.E. vs. R^2 Ajustado



b) V.E. vs. C_p Mallows



c) V.E. vs. PRESS



d) V.E. vs. AIC

Autoría: Macías S. – Pincay C.

CONCLUSIONES

Las tecnologías de la información (TI) ofrecen grandes posibilidades al mundo de la educación. Pueden facilitar el aprendizaje de conceptos y materias, ayudar a resolver problemas y contribuir a desarrollar las habilidades cognitivas.

Se enuncian las principales conclusiones derivadas del Trabajo Especial de Grado expuesto.

- Existen numerosas técnicas para la construcción de un software estadístico, por lo que es importante escoger y determinar las que mejor se adapten al contexto y a las necesidades que se deseen satisfacer, así como a las características de la población objetivo.

- Asimismo el lenguaje de programación Microsoft Visual Basic 8.0 de la familia de Microsoft Visual Studio 8.0 permitió el desarrollo de un software con una interface amigable con el usuario la cual satisface el requerimiento de ser apto para fines educativos; además de que el usuario final fue un programa computacional con características profesionales y que permiten su fácil entendimiento, entre las cuales se pueden mencionar cuadros de dialogo, consejos como ayuda. Menú emergente para el manejo de resultados, etc.

- Si bien hay en el mercado diversas opciones de software estadísticos, su utilización se limita en gran parte a la parte básica de la técnica de regresión, por lo que es importante fomentar a “ERLA” en su desarrollo e implementación para que se incremente su uso en las aulas de clase, así como en los diferentes niveles de investigación.

- El sistema de software presentado está asentado en los principios de las teorías constructivistas, ya que se basa la construcción del conocimiento en la capacidad de cada individuo, apoyando así la construcción inicial de modelos predictivos. Sin embargo es importante señalar que un software estadístico basado en un sólo enfoque estaría incompleto, por lo que es necesario involucrar aspectos de las demás teorías existentes, como se lo ha realizado con “ERLA”.

- El desarrollo de un software estadístico incluye profesionales y/o expertos, por lo que a una primera instancia fue necesario considerar un número de graduandos, en el proceso para determinar, de manera más completa, los aspectos que influyen

en el proceso de construcción y aprendizaje, para así lograr un mejor desarrollo y uso de “ERLA”.

- La Cátedra de Regresión Lineal Avanzada tiene como uno de sus objetivos “Relacionar los conocimientos adquiridos de Ingeniería Clásica con aplicaciones avanzadas y recientemente descubiertas por especialistas en el tema, mediante la elaboración de simulaciones de problemas con la ayuda del computador”. Sin embargo esto poco se lleva a la práctica, ya que las actividades o tareas orientadas a cumplir con este objetivo no se han mantenido ni aprovechado de la manera más eficiente con el paso del tiempo, por lo que es vital desarrollar aplicaciones que permitan lograr el objetivo citado.
- El presente Reporte Especial de Grado puede servir de base para su expansión y adaptación a otros tópicos o temas y/o para futuros proyectos en ésta y otras áreas de conocimiento.
- Todo sistema de software depende del apoyo que reciba, de Entidades ya sean Públicas o Privadas; y de la utilización del mismo, por lo que el éxito de este proyecto depende del uso, impulso y aplicación de la Escuela Superior Politécnica del Litoral “ESPOL” y profesionales.

RECOMENDACIONES

Desde la concepción del desarrollo de un sistema de software surgen ideas que deben ser descartadas para poder determinar el alcance del proyecto, sin embargo, dichas ideas pueden servir de base para la expansión y mejoramiento del proyecto.

Algunas de las recomendaciones se exponen en las líneas siguientes:

- Disminuir la incertidumbre en la administración del software en los distintos módulos, usando el manual de usuario.

- Elaborar módulos de estadísticas, donde los usuarios pueden consultar el rendimiento del Software (individual o por sección) y los usuarios puedan consultar su rendimiento de forma personal o global con respecto al Software.

REFERENCIAS BIBLIOGRAFICAS

[1] **Bovas A. y Johannes L.** (2006) *Introduction to Regression Modeling*, Primera Edición, Thomson Brooks/Cole, USA.

[2] **Zurita G.** (2010) *Probabilidad y Estadística*, Segunda Edición, Centro de Difusión y Publicaciones - ESPOL, Guayaquil, Ecuador.

[3] **Rencher A.** *Methods of Multivariate Analysis*, Segunda Edición, Wiley Interscience.

[4] **Freund J., Miller I., Miller M.** (2000) *Estadística Matemática con Aplicaciones*, Sexta Edición, Prentice Hall, México.

[5] **Timm N.** (2002) *Applied Multivariate Analysis*, Springer, New York, USA.

[6] **Mallows, C.** (1973) *Some comments on* C_p, Techmetrics, **15:** 661 – 664.

[7] **Contreras Juana, Del Pino Claudio (2011)** Matemática interactiva, http://matesup.utalca.cl

[8] **Universidad de Málaga. (2011)** Bioestadística: Métodos y Aplicaciones, http://www.bioestadistica.uma.es/libro/node97.htm

[9] **Universidad Nacional de Colombia. (2011)** Métodos de Regresión, http://www.virtual.unal.edu.co/cursos/ciencias

[10] **Galton F. (1889)** *Natural Inheritance*, Primera Edición, Macmillan, Londres.

[11] **ReliaSoft Corporation. (2011)** Hypothesis Tests in Multiple Linear Regression**,** http://www.weibull.com

[12] **Lopez, E.** (1998) *Tratamiento De La Colinealidad en Regresión Múltiple*, **10:** 491 – 507.