Universidad Nacional Autónoma de México

FACULTAD DE CIENCIAS

ANA-RELI .M VER. 2.0: UN SISTEMA COMPUTACIONAL FACIL DE USAR PARA EL ANALISIS DE REGRESION LINEAL

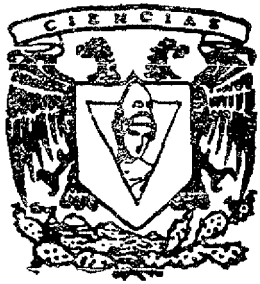
T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

RIGOBERTO REAL MIRANDA



DIRECTOR DE TESIS:

DR. JESUS LOPEZ ESTRADA

1 9 9 9

PAGINACION

DISCONTINUA

UNIVERSIDAD NACIONAL
AVENTA DE
MEXICO

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo de Tesis:

ANA-RELI.M VER. 2.0: UN SISTEMA COMPUTACIONAL FACIL DE USAR PARA EL ANALISIS DE REGRESION LINEAL

realizado por Rigoberto Real Miranda

con número de cuenta 9129563-1 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. Jesús López Estrada
Propietario	M. en A. P. Ma. del Pilar Alonso Reyes
Propietario	Dr. Humberto Madrid de la Vega
Suplente	M. en C. José Antonio Flores Díaz
Suplente	Act. Guadalupe Iztapalapa Cervantes

Consejo Departamental de Matemáticas

M. en A. P. MARIA DEL PILAR ALONSO REYES

Ana_Reli.M Ver. 2.0: Un Sistema Computacional fácil de usar para el Análisis de Regresión Lineal.

Rigoberto Real Miranda.

12 de marzo de 1999

Prefacio

El Análisis de Regresión es una de las más extensas técnicas estadísticas usadas para el análisis de datos múltiples. Las aplicaciones del análisis de regresión son numerosas y se presenta con frecuencia en cualquier campo, tal como: la ingeniería, ciencias físicas, economía, administración, ciencias biológicas y ciencias sociales.

Con frecuencia el número de variables, tomadas para el estudio del fenómeno, es muy grande por lo que, para efectuar las operacioes aritméticas, se recurre al uso de una computadora; sin embargo se debe tener cuidado con los métodos numéricos utilizados para el cálculo de las estadísticas, pues éstas deben ser lo más precisas y confiables posible, ya que a partir de estas se toman decisiones importantes sobre el fenómeno de estudio; y en caso de no considerar los métodos numéricos adecuados la toma de decisiones puede verse afectada al no poderse evitar los errores por redondeo en las operaciones realizadas con ayuda de una computadora digital. Por esta razón, los métodos numéricos utilizados en la realización del presente trabajo fueron obtenidos de textos que gozan de prestigio internacional; además, algunas de las funciones utilizadas fueron obtenidas del paquete Matlab 5.0, el cual goza de reconocimiento internacional y que poco a poco va ganando terreno en su uso, en distintas áreas del conocimiento.

El objetivo del presente trabajo es el desarrollo de un sistema computacional amigable, interactivo y fácil de usar; con el cual se pueda llevar a cabo el análisis de regresión para un modelo lineal de la forma:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \text{con} \quad \underline{\varepsilon} \sim N\left(\underline{0}, \sigma^2 I\right). \tag{0.1}$$

Con tal sistema computacional se busca entonces, el calcular las estimaciones Gauss-Markov para β y σ^2 así como el análisis de las observaciones y de los regresores que intervienen en el modelo, realizando de esta manera análisis estadístico mediante pruebas de hipótesis (validación del modelo y pruebas $t - student$), coeficiente de determinación, análisis de residuales, detección de colinealidad, predicción, etc.; también se realiza un análisis gráfico para llevar a cabo la verificación de los supuestos que acompañan al modelo dado en (0.1), que son el de varianza constante (homocedasticidad), errores independientes y normalidad del vector de errores no-observables $\underline{\varepsilon}$.

En el Capítulo **1** se hace una breve revisión de los conceptos básicos utilizados en el análisis de regresión lineal.

En el Capítulo **2** se presentan estadísticas y coeficientes (o indicadores) que permiten la validación del modelo de regresión lineal ajustado, mediante pruebas de hipótesis, correlaciones, coeficientes de determinación entre otros.

Los capítulos anteriores representan el **Manual de Referencia de aspectos Estadísticos del sistema.**

En el Capítulo **3** se presentan temas que permiten la detección de observaciones aberrantes en el modelo, así como el análisis de residuales por medio de métodos gráficos y, finalmente la verificación de los supuestos del modelo como el de normalidad y varianza constante (homocedasticidad). Este capítulo representa el **Manual de Referencia de procedimientos de Análisis Gráficos.**

En el Capítulo **4** se discuten métodos numéricos, tal como la descomposición QR de la matriz X, que permiten la obtención de las estimaciones Gauss-Markov para β y σ^2 mediante procedimientos que resultan más confiables que los comúnmente conocidos. Este capítulo representa el **Manual de Referencia de procedimientos de Análisis Gráficos.**

En el Capítulo **5** se hace la presentación del sistema ANA_RELI.M VER. 2.0, cuyas características y objetivos son el ser un sistema amigable y de fácil manejo, transparente en sus resultados y confiable en el aspecto numérico; para alcanzar tales logros, se programó en Matlab 5.0 teniéndose así la ventaja de algunas de las funciones numéricos como la descomposición QR de la matriz X, así como del diseño de ventanas, botones y diálogos de ayuda, que permiten al usuario tener pleno conocimiento de las opciones con que cuenta el sistema. Este capítulo representa el **Manual de Ayuda rápida.**

En el Apéndice **A** se revisa la estructura del sistema, las funciones que lo componen y su clasificación, dependiendo de la finalidad que tengan en el sistema. Este apéndice es considerado como el **Manual de Aspectos Técnicos del Sistema.**

En el apéndice **B** se presenta el Documento inicial del Análisis del Sistema ANA_RELI.M, mencionándose la parte estadística y numérica que se ha logrado con la presentación de este trabajo, así como las partes pendientes del mismo.

En el Apéndice **C** son presentados los Demos de algunos ejemplos propuestos obtenidos de textos reconocidos en el ambiente del Análisis de Regresión y el Análisis Númerico, dichos ejemplos son presentados como salidas del sistema ANA_RELI.M VER 2.0. Además, se mencionan algunos comandos de Matlab que permiten la realización (o creación) de ventanas, menús, botones y presentación de texto en la ventana.

Finalmente, quiero externar los siguientes agradecimientos a la gente que siempre estuvo presente en los momentos difíciles.

Un agradecimiento especial al Dr. Jesús López Estrada por la valiosa colaboración prestada en la realización de este trabajo, así como a todas aquellas personas que directa o indirectamente participaron en la realización del mismo.

A los honorables miembros del jurado:

Dr. Jesús López Estrada.

M. en A. P. Ma. del Pilar Alonso Reyes.

Dr. Humberto Madrid de la Vega.

M. en C. José Antonio Flores Díaz.

Act. Guadalupe Tzintzún Cervantes.

por su acertada orientación en la realización del mismo.

Un agradecimiento muy especial a la **M. en A. P. Ma. del Pilar Alonso Reyes** por su gran apoyo incondicional, antes y después del término de este trabajo.

A la **Universidad** y a la **Facultad de Ciencias.**

Un eterno agradecimiento a mi **padre** por haberme dado la gran oportunidad de superarme y llevar a cabo la realización de esta carrera profesional, que para mi es y será la mejor herencia que jamás podré recibir de él.

También un eterno agradecimiento a mi **madre**, quien siempre estuvo presente en esos momentos difíciles y que con su ejemplo y dedicación logró llevarme a la meta tan deseada.

Un agradecimiento también a todos mis hermanos y familiares por estar siempre conmigo.

A todos los grandes **amigos** que siempre estuvieron conmigo, tanto en los momentos difíciles como en los de alegría, y que ahora no me atrevo a mencionar por temor de que en este momento alguno de sus nombres quede fuera.

Un agradecimiento sincero al **Ing. Juan Juárez Martínez** por estar a mi lado en esta última etapa del tan largo y complicado camino.

Rigoberto Real Miranda.

México, D. F.
Marzo de 1999.

Índice General

Capítulo 1

Conceptos básicos utilizados en el Análisis de Regresión Lineal.

Introducción.

En este capítulo se presenta una breve revisión sobre los fundamentos matemáticos del Análisis de Regresión Lineal, tema al cual está enfocado el siguiente trabajo. Para ello, se utilizarán ejemplos por medio de los cuales se ilustrará la aplicación de esta teoría.

Se hará referencia a la bibliografía en que se pueden encontrar las demostraciones de los teoremas mencionados, en vez de incluirlas en este texto, ya que son ellas muy conocidas.

1.1 Modelos lineales.

1.1.1 Conceptos básicos.

El Análisis de Regresión es una de las más extensas técnicas estadísticas usadas para el análisis de datos múltiples. Las aplicaciones de la regresión son numerosas y se presenta casi en cualquier campo, incluyendo ingeniería, ciencias físicas, economía, administración, ciencias biológicas y las ciencias sociales.

Los modelos de regresión son usados para varios propósitos, incluyendo los siguientes:

1. Descripción de datos.

2. Estimación de parámetros.

3. Predicción y estimación de datos.

4 Control de fenómenos.

1.1.2 Construcción de modelos de regresión.

El Análisis de Regresión es una herramienta estadística que se basa en una relación funcional entre dos o más variables cuantitativas tal que una variable esté en función de las otras. Específicamente supóngase que se tiene una variable observable y en función de las variables de control o regresores $x_0,\ x_1, \ldots,\ x_{p-1}$ de un determinado proceso bajo estudio, i. e.

$$y = f\left(x_0, x_1, \ldots, x_{p-1}; \beta_0, \beta_1, \ldots, \beta_q\right) + \varepsilon \tag{1.1}$$

donde ε representa una variable aleatoria debida a diversos factores que intervienen en dicho proceso y que el experimentador no puede controlar, y $\beta_0, \beta_1, \ldots, \beta_q$ son ciertos parámetros, generalmente desconocidos.

La forma más sencilla de (1.1) es la siguiente:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon \tag{1.2}$$

La forma funcional dada por (1.2), se propone con frecuencia como inicial entre estas variables, que es lineal en los parámetros y en las variables o regresores.

Para iniciar el análisis con el modelo de regresión lineal, es necesario que el estudio se haga sobre un conjunto de datos previamente obtenidos, de esta manera se pueden agrupar las observaciones obtenidas junto con las variables de control en forma matricial, de la siguiente forma:

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}, \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

De esta manera el estudio de y se puede enfocar a estudiar la función de asociación del vector $\underline{y} \in \Re^n$ con respecto a la matriz $X \in \Re^{nxp}$ $n \geq p$; así, el modelo lineal (1.2) en forma matricial toma la forma:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{1.3}$$

donde $\underline{y}$ es el vector de observación, X la matriz de variables de control, $\underline{\beta}$ un vector de parámetros desconocidos y $\underline{\varepsilon}$ el vector de errores no observables.

Como se mencionó, $\underline{\varepsilon}$ es el vector de errores no observables, debidos a elementos del proceso que no son controlables, ya que cuando se observa algún fenómeno, existe siempre cierta diferencia o desviación entre lo que se observa y lo que teóricamente debe ocurrir, estas diferencias son los errores de observación. Estos se consideran variables aleatorias, idénticamente distribuidas, independientes, de media cero y varianza constante, esto es, vector de medias $E(\underline{\varepsilon}) = \underline{0}$ y matriz de Varianzas-Covarianzas $Var(\underline{\varepsilon}) = \sigma^2 I$, es decir:

$$\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I) \tag{1 4}$$

Si se desea hacer inferencias estadísticas sobre el modelo de regresión, entonces se le pide un supuesto adicional a los residuales, este es que tengan una distribución Normal con los mismos parámetros, es decir:

$$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I) \tag{1.5}$$

1.2 Mínimos cuadrados.

El problema inicial, una vez propuesto el modelo, es el ajuste del mismo al conjunto de datos, de tal manera que se minimicen las distancias de los valores observados con los ajustados por el modelo. El total de estas distancias está dada por la siguiente expresión:

$$SCE\left(\underline{\tilde{\beta}}\right) = \underline{\varepsilon}^t \underline{\varepsilon} = \left(\underline{y} - X\underline{\tilde{\beta}}\right)^t \left(\underline{y} - X\underline{\tilde{\beta}}\right),$$

donde $\underline{\varepsilon} = \underline{y} - X\underline{\tilde{\beta}}$ es el vector de residuales y $X\underline{\tilde{\beta}}$ es el vector de valores que son ajustados por el modelo.

El método a usar es el conocido como Mínimos Cuadrados, el cual es el más apropiado porque permite la utilización del cálculo diferencial para solucionar el problema.

Diferenciando la función $SCE\left(\underline{\tilde{\beta}}\right)$ con respecto a $\underline{\tilde{\beta}}$ tenemos que:

$$\frac{\partial SCE\left(\underline{\tilde{\beta}}\right)}{\partial \underline{\tilde{\beta}}} = -2X^t\underline{y} + 2X^tX\underline{\tilde{\beta}} = \underline{0}$$

que es igualada a cero para encontrar un punto crítico $\underline{\hat{\beta}}$, llegando a:

$$X^tX\underline{\hat{\beta}} = X^t\underline{y}, \tag{1.6}$$

Ahora solo faltaría verificar que el punto crítico, dado por la relación en (1.6), minimiza la función $SCE\left(\underline{\tilde{\beta}}\right)$. Para ello, basta con comprobar que

$$\frac{\partial^2 SCE\left(\underline{\tilde{\beta}}\right)}{\partial \underline{\tilde{\beta}}^2}$$

es una matriz definida positiva. En efecto, se tiene que:

$$\frac{\partial^2 SCE\left(\underline{\tilde{\beta}}\right)}{\partial \underline{\tilde{\beta}}^2} = 2X^tX > 0 \qquad p.t. \quad \underline{\tilde{\beta}},$$

es una matriz simétrica y definida positiva siempre que X sea una matriz de rango igual p.

Esto comprueba que efectivamente existe un mínimo $\underline{\hat{\beta}}$ para $SCE\left(\underline{\tilde{\beta}}\right)$, si se tiene que $X \in \Re^{n\times p}$, $n \geq p$, tiene rango p.

A las ecuaciones (1.6) se les conoce como las ecuaciones normales.

Si X tiene rango p, entonces X^tX es definida positiva, y las ecuaciones normales tienen una única solución, dada por:

$$\underline{\hat{\beta}} = \left(X^tX\right)^{-1}X^t\underline{y},$$

donde el vector $\underline{\hat{\beta}}$ se conoce como el estímador de Mínimos Cuadrados de $\underline{\beta}$.

1.2.1 Propiedades de los estimadores Mínimos Cuadrados.

Bajo el supuesto que $\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I)$, las principales características de los estimadores Mínimos Cuadrados son:

1. Para $\hat{\underline{\beta}}$.

   (a) $E\left(\hat{\underline{\beta}}\right) = \underline{\beta}$.

   Por lo que el estimador $\hat{\underline{\beta}}$ es un estimador insesgado de $\underline{\beta}$ (Montgomery [8], pag. 128).

   (b) $Var\left(\hat{\underline{\beta}}\right) = \sigma^2 (X^t X)^{-1}$

   En efecto, ya que $Var\left(\underline{y}\right) = \sigma^2 I$, se tiene que

$$\begin{aligned} Var\left(\hat{\underline{\beta}}\right) &= Var\left((X^t X)^{-1} X^t \underline{y}\right) \\ &= (X^t X)^{-1} X^t Var\left(\underline{y}\right) X (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1} (X^t X) (X^t X)^{-1} \\ &= \sigma^2 (X^t X)^{-1} \end{aligned}$$

2. Para σ^2.

   (a) El estimador insesgado de σ^2, el cual es independiente de $\hat{\underline{\beta}}$ es:

$$s^2 = \frac{\left(\underline{y} - X\hat{\underline{\beta}}\right)^t \left(\underline{y} - X\hat{\underline{\beta}}\right)}{n-p} = \frac{SCE}{n-p}$$

   En efecto, debido a que $\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I)$, basta demostrar que la covarianza es cero.

$$
\begin{aligned}
Cov\left[\underset{\sim}{\hat{\beta}}, \underset{\sim}{y} - X\underset{\sim}{\hat{\beta}}\right] &= Cov\left[\left(X^tX\right)^{-1}X^t\underset{\sim}{y}, \underset{\sim}{y} - X\left(X^tX\right)^{-1}X^t\underset{\sim}{y}\right] \\
&= Cov\left[\left(X^tX\right)^{-1}X^t\underset{\sim}{y}, \left(I - X\left(X^tX\right)^{-1}X^t\right)\underset{\sim}{y}\right] \\
&= \left(X^tX\right)^{-1}X^t\,Cov\left(\underset{\sim}{y}, \underset{\sim}{y}\right)\left(I - X\left(X^tX\right)^{-1}X^t\right)^t \\
&= \left(X^tX\right)^{-1}X^t\,Var\left(\underset{\sim}{y}\right)\left(I - X\left(X^tX\right)^{-1}X^t\right)^t \\
&= \sigma^2\left(X^tX\right)^{-1}X^t\left(I - X\left(X^tX\right)^{-1}X^t\right)^t \\
&= 0
\end{aligned}
$$

(b) $E\left(s^2\right) = \sigma^2$.

Por lo que se tiene que el estimador s^2 es insesgado para σ^2. (Montgomery [8], pag. 129).

3. Si además se cumple (1.5) entonces $\underset{\sim}{y}$ tiene distribución $N\left(X\underset{\sim}{\beta}, \sigma^2 I\right)$ donde X es de $(n \times p)$ de rango p, y entonces se tiene que:

(a) $$\frac{\left(\underset{\sim}{\beta} - \underset{\sim}{\hat{\beta}}\right)^t X^tX\left(\underset{\sim}{\beta} - \underset{\sim}{\hat{\beta}}\right)}{\sigma^2} \sim \chi^2_{(p)}.$$

(b) $$\frac{SCE}{\sigma^2} = \frac{(n-p)\,s^2}{\sigma^2} \sim \chi^2_{(n-p)}.$$

(Seber [15], pag. 54-56).

Con base en tales suposiciones se determinarán estadísticas que permitirán evaluar qué tan bien se ajusta el modelo mencionado a los datos, eliminar variables, etc.

Finalmente es conveniente mencionar que $\underset{\sim}{\hat{\beta}}$ es un estadístico además de insesgado, de mínima varianza, consistente y eficiente.

1.2.2 Resultados probabilísticos.

Algunas estadísticas básicas de mucha utilidad en el análisis de regresión se basan en los siguientes hechos probabilísticos:

1. Si k-variables aleatorias $X_1, \ldots, X_k$, se distribuyen normal e independientes con media μ_i y varianzas σ_i^2, entonces

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

tiene una distribución ji-cuadrada con k grados de libertad (Mood [9], pag. 242), lo cual se denota por $U \sim \chi^2_{(k)}$. Como ejemplos de esta distribución se tiene a SCT, SCR y SCE que se definen en la siguiente sección, y que juegan un papel relevante en el análisis de regresión

2. Sean U y V variables aleatorias que se distribuyen como ji-cuadrada con m y n grados de libertad respectivamente, e independientes entre sí, entonces la variable aleatoria

$$Z = \frac{U/m}{V/n}$$

se distribuye como una F con m y n grados de libertad (Mood [9], pag. 247), lo cual se denotará por $Z \sim F_{(m,n)}$.

3. Si $Q_i \sim \chi^2_{(r_i)}$ para $i = 1, 2$, $r_1 > r_2$, y $Q = Q_1 - Q_2$ es independiente de Q_2, entonces $Q \sim \chi^2_{(r)}$ donde $r = r_1 - r_2$. (Seber [15], pag. 20).

4 Si Z es una variable aleatoria normal estándarizada, y U tiene una distribución ji-cuadrada con k grados de libertad, y Z y U son independientes, entonces

$$W = \frac{Z}{\sqrt{U/k}}$$

tiene una distribución t de Student con k grados de libertad (Mood [9], pag. 250), y se denotará $W \sim t_{(k)}$.

En la siguiente sección se definen las sumas de cuadrados utilizadas en el análisis de la varianza de los estimadores.

1.3 Sumas de cuadrados.

Se tiene que "la suma de cuadrados de las desviaciones de las y_i observadas de sus valores esperados" (Searle [14], pag.92) es conocida como la *Suma de Cuadrados de los Errores*, y es denotada por SCE. En símbolos,

$$SCE \underset{def.}{=} \left(\underline{y} - \hat{\underline{y}}\right)^t \left(\underline{y} - \hat{\underline{y}}\right) = \left(\underline{y} - X\hat{\underline{\beta}}\right)^t \left(\underline{y} - X\hat{\underline{\beta}}\right) \tag{1.7}$$

La *Suma de Cuadrados Totales* está definida por:

$$SCT = \underline{y}^t \underline{y} \tag{1.8}$$

Se tiene que SCE es una medida de la variación en y después de estimar a $\hat{\underline{\beta}}$ e indica qué tanto se separan los datos del modelo, si SCE es "pequeño" implica que los datos se ajustan bien al modelo, y la SCT es una medida de la variación en y sin considerar la estimación de $\underline{\beta}$.

La diferencia entre (1.7) y (1.8) es conocida como la Suma de Cuadrados atribuible a la Regresión, y es denotada por:

$$SCR = SCT - SCE,$$

y "representa qué porción de SCT es atribuible al tener la regresión ajustada" (Searle [14], pag. 94) es decir, indica la variación de la estimación $\left(\hat{y}_i\right)$ con respecto al promedio $(\overline{y})$.

1.3.1 Presentación de los datos.

Existen dos formas de pre-procesamiento de los datos, el centrarlos respecto a su media y el de estándarizarlos:

1. Centralización: tiene su justificación en el hecho de que si los intervalos que definen los valores de las variables son muy grandes, al efectuar operaciones en una computadora esto puede provocar redondeos que generen soluciones numéricas inadecuadas.

2 Estandarización: además de tener la ventaja de centralizar los datos, se tiene la posibilidad de comparar dos modelos de regresión que sean similares en cuanto al significado de las variables de estudio, pero diferentes en cuanto a las unidades de medida empleadas en la obtención de los datos.

Centralización de los datos.

Para centrar los datos con respecto a sus medias es necesario:

1. Obtener las medias de las columnas de la matriz X que se denotan como $\overline{x}_i$, $i = 1, \ldots, p-1$.
2. Obtener la media de la variable de respuesta y $(\overline{y})$.
3. Tomar $x_{ij}^* = x_{ij} - \overline{x}_j, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p-1.$
4. Tomar $y_i^* = y_i - \overline{y}, \quad i = 1, 2, \ldots, n.$

Al centrar los datos los parámetros $\hat{\underline{\beta}}$ estimados son los mismos que se obtienen en el caso de que los datos no se centren con respecto a sus medias; si se desea considerar en el modelo una constante β_0 (véase 1.2), ésta queda determinada por la siguiente expresión:

$$\hat{\beta}_0 = \overline{y} - \sum_{j=1}^{p-1} \hat{\beta}_j \, \overline{x}_j$$

Las suma de cuadrados que se tienen cuando se centran los datos son las siguientes:

$$SCR_m = \hat{\underline{\beta}} X^* \underline{y}^*$$

$$SCT_m = \left(\underline{y}^*\right)^t \underline{y}^*$$

donde X^* y $\underline{y}^*$ son los datos centrados.

Estandarización de los datos.

El procedimiento para estándarizar los datos es el siguiente:

1. Obtener las medias de las columnas de la matriz X que se denotan como $\overline{x}_i$, $i = 1, \ldots, p-1$.

2. Obtener las varianzas para cada columna de la matriz X, denotadas por s_j^2, $j = 1, \ldots, p-1$.

$$s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}; \quad j = 1, 2, \ldots, p-1.$$

3. Obtener la media de la variable de respuesta y $(\overline{y})$

4. Obtener la varianza de la variable de repuesta y $\left(s_y^2\right)$.

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

5. Tomar $x_{ij}^* = \dfrac{x_{ij} - \overline{x}_j}{s_j}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p-1.$

6. Tomar $y_i^* = \dfrac{y_i - \overline{y}}{s_y}, \quad i = 1, 2, \ldots, n.$

Con estos cambios, a partir del modelo

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

se construye un nuevo modelo

$$\underline{y}^* = X^*\underline{\beta}^* + \underline{\varepsilon}.$$

Los datos que han sido estándarizados tienen media cero y varianza igual a uno, y los parámetros $\hat{\underline{\beta}}^*$ estimados están relacionados a los parámetros $\hat{\underline{\beta}}$ de la siguiente manera:

$$\hat{\beta}_j = \hat{\beta}_j^* \sqrt{\frac{s_y^2}{s_j^2}}, \qquad j = 1, 2, \ldots, p-1$$

$$\hat{\beta}_0 = \overline{y} - \sum_{j=1}^{p-1} \hat{\beta}_j \, \overline{x}_j$$

1.4 Estadística F.

Siempre que se ajusta un modelo de regresión es necesario saber si las variables de control propuestas aportan explicación alguna a la variable observable del modelo; para ello, se plantea el ensayo de hipótesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} \quad Vs. \quad H_1 : \beta_i \neq 0 \quad \text{p. a. } i = 1, \ldots, p-1$$

el estadístico que nos ayuda a tomar la decisión sobre esta hipótesis es:

$$F = \frac{\frac{1}{(p-1)} U}{\frac{1}{(n-p)} V} = \frac{SCR/(p-1)}{SCE/(n-p)} \tag{1.9}$$

recordando de la sección (1.2.2) que este estadístico se genera a partir del cociente de ji-cuadradas divididas entre sus grados de libertad y que las siguientes sumas de cuadrados tienen distribuciones ji-cuadradas:

$$U = \frac{SCR}{\sigma^2} \sim \chi^2_{(p-1)}, \qquad V = \frac{SCE}{\sigma^2} \sim \chi^2_{(n-p)}$$

En el Análisis de la Varianza se compara a la estadística F determinada por (1.9) con el cuantil $(1-\alpha)$ que hay en las tablas de la distribución F con $p-1$, $n-p$ grados de libertad respectivamente. Éste cuantil se denota como: $F^{1-\alpha}_{(p-1,n-p)}$.

Cuando el valor de la estadística F es mayor que el del cuantil encontrado en las tablas de la distribución F a un nivel $(1-\alpha)$, se rechaza la hipótesis de que $\beta_1 = \cdots = \beta_{p-1} = 0$ al nivel de significancia α.

Ejemplo 1.4.1 (Neter) *En un pequeño estudio de regresión, se obtuvieron los siguientes datos:*

Datos	*Ventas*	*Gastos*	*Marcas*
Mes	y	x_1	x_2
Enero	42	7	33
Febrero	33	4	41
Marzo	75	16	7
Abril	28	3	49
Mayo	91	21	5
Junio	55	8	31

donde (y) representa las ventas realizadas por una Compañía durante el primer semestre de un cierto año, (x_1) representa los gastos realizados para dichas ventas y (x_2) representa el número de marcas registradas en el mercado por la competencia. Mediante un modelo de Regresión Lineal, las ventas esperadas están dadas por:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A partir de estos datos el método de Mínimos Cuadrados proporciona la estimación:

$$\hat{E(y)} = 33.9321 + 2.7847 x_1 - 0.2644 x_2$$

La estadística F correspondiente a éste modelo es $F = 72.73453522$, dada por (1.9) y el valor en tablas del cuantil $F^{1-\alpha}_{(2,3)}$ con un nivel de significancia de $\alpha = .05$ es de $F^{95}_{(2,3)} = 9.55$. Dado que $F = 72.734 > 9.55 = F^{95}_{(2,3)}$ se concluye que el modelo explica la variación que hay en y por las variables x_1 y x_2, por lo tanto se puede decir que al menos β_1 ó β_2, (o ambos) es diferente de cero, es decir, existe una relación entre las variables de control y la variable de respuesta.

1.5 Intervalos de confianza.

Con el método de Mínimos Cuadrados se obtienen estimadores puntuales de los parámetros del modelo lineal, $\underline{\beta}$, σ^2, y $E(y)$. Una vez encontrados estos estimadores se pueden encontrar intervalos que con una confianza dada contengan el valor verdadero del parámetro.

1.5.1 Para $\hat{\beta}_i$.

Directamente no se puede calcular un intervalo de confianza para el vector $\underline{\beta}$, es por ello que se calcula para cada una de sus coordenadas, sabiendo que:

$$\hat{\underline{\beta}} \sim N\left(\underline{\beta}, \sigma^2 \left(X^t X\right)^{-1}\right)$$

o sea que

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 c_{ii}\right), \qquad i = 0, \ldots, p-1 \tag{1.10}$$

donde c_{ii} es el i-ésimo elemento de la diagonal de la matriz $C = \left(X^t X\right)^{-1}$.

Como la distribución en (1.10) contiene 2 parámetros desconocidos, β_i y σ^2, directamente no se puede encontrar un intervalo para β_i, por lo que se utiliza un resultado visto en la sección 1.2.2, el cociente de una normal estándarizada y la raíz de una ji-cuadrada entre sus grados de libertad tiene distribución t-student con los grados de libertad de la ji-cuadrada. Usando este resultado se llega a lo siguiente:

$$P\left(\left|\beta_i - \hat{\beta}_i\right| < a s\sqrt{c_{ii}}\right) = 1 - \alpha$$

equivalente a:

$$P\left(\hat{\beta}_i - a s\sqrt{c_{ii}} < \beta_i < \hat{\beta}_i + a s\sqrt{c_{ii}}\right) = 1 - \alpha$$

donde s es la raíz cuadrada del valor estimado de la varianza y $a = t_{(n-p)}^{1-\frac{\alpha}{2}}$ es el cuantil al $\left(1 - \frac{\alpha}{2}\right)$ de confianza de una distribución t-student con $(n-p)$ grados de libertad.

Por lo tanto se tiene que los intervalos al $(1-\alpha) \times 100\%$ de confianza para cada β_i, están dados por:

$$\beta_i \quad \in \quad \left(\hat{\beta}_i - a s\sqrt{c_{ii}}, \quad \hat{\beta}_i + a s\sqrt{c_{ii}}\right), \qquad i = 0, 1, \ldots, p-1$$

1.5.2 Para σ^2.

En ocasiones es necesario saber, con un nivel de significancia dado, entre qué valores está la varianza original de los errores. Es a partir de su estimación s^2 que podemos saberlo, teniendo como resultados previos, de las secciones 1.2.1 y 1.4 que:

$$\frac{SCE}{\sigma^2} = \frac{(n-p)\,s^2}{\sigma^2} \sim \chi^2_{(n-p)}$$

o sea que, con un nivel de significancia a tenemos:

$$P\left(a < \frac{(n-p)\,s^2}{\sigma^2} < b\right) = 1-\alpha$$

equivalente a:

$$P\left(\frac{(n-p)\,s^2}{b} < \sigma^2 < \frac{(n-p)\,s^2}{a}\right) = 1-\alpha$$

donde s^2 es la estimación de la varianza y $a = \chi^2_{\left(\frac{\alpha}{2},\ \ n-p\right)}$, $\quad b = \chi^2_{\left(1-\frac{\alpha}{2},\ \ n-p\right)}$, son los cuantiles al $\left(\frac{\alpha}{2}\right)$ y $\left(1-\frac{\alpha}{2}\right)$ de confianza, respectivamente, de una distribución ji-cuadrada con $(n-p)$ grados de libertad.

Por lo tanto, el intervalo al $(1-\alpha)\times 100\%$ de confianza para σ^2 es:

$$\sigma^2 \quad \in \quad \left(\frac{(n-p)\,s^2}{b},\quad \frac{(n-p)\,s^2}{a}\right)$$

Ejemplo 1.5.1 *A continuación se calculan los intervalos de confianza para los parámetros y la varianza del Ejemplo 1.4.1:*

1. *Para los parámetros* β_i, $i = 0, 1, 2$.

$$s^2 = 20.6911, \qquad a = t^{.975}_{(3)} = 3\ 1825$$

(a) Para β_0

$$\beta_0 \in \left(33.9321 \pm 3.1825 * \sqrt{20\ 6911} * \sqrt{34.5785}\right),$$

por lo que el intervalo al 95% de confianza para β_0 *es:*

$$(-51.194,\quad 119\ 06).$$

(b) Para β_1

$$\beta_1 \in \left(2\,7847 \pm 3.1825 * \sqrt{20.6911} * \sqrt{.0803}\right),$$

por lo que el intervalo de confianza al 95% para β_1 *es:*

$$(-1.3175, \quad 6.8869).$$

(c) Para β_2

$$\beta_2 \in \left(-0.2644 + 3.1825 * \sqrt{20.6911} * \sqrt{.0126}\right),$$

por lo que el intervalo de confianza al 95% para β_2 *es:*

$$(-1.8894, \quad 1.3606).$$

2. Para σ^2.

$$a = \chi^{2(\,025)}_{(3)} = .2158, \quad b = \chi^{2(\,975)}_{(3)} = 9\,3484$$

por lo que el intervalo de confianza al 95% para la varianza queda:

$$\left(\frac{3 * 20.6911}{9.3484}, \frac{3 * 20.6911}{.2158}\right)$$

entonces, el valor real de la varianza σ^2, *tiene límites de confianza*

$$(6.64, \quad 287.64).$$

Capítulo 2

Revisión de estadísticas que permiten la validación del Modelo de Regresión Lineal.

Introducción.

En este capítulo se presentarán los supuestos básicos planteados para llevar a cabo el Análisis de Regresión Lineal, y se obtendrán las estadísticas que permiten la evaluación de dicho modelo de regresión y de las variables que intervienen en el mismo.

2.1 Supuestos básicos en el Modelo de Regresión Lineal.

Para obtener en forma conveniente el ajuste de un Modelo Lineal Múltiple ó de Regresión Lineal de la forma:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{2.1}$$

es necesario observar que se cumplan los siguientes supuestos:

1. El modelo debe ser lineal en los parámetros β_i, $i = 0, \quad , p-1$.

2. Los errores en las observaciones deben ser estocásticamente independientes, tener una distribución idéntica de media cero y la misma varianza constante, esto es,

$$\underline{\varepsilon} \sim \left(\underline{0}, \sigma^2 I\right)$$

El supuesto de distribución normal en los errores se plantea para que con ello sea posible efectuar pruebas de hipótesis al modelo y verificar sus supuestos, esto es,

$$\underline{\varepsilon} \sim N\left(\underline{0}, \sigma^2 I\right)$$

3 X es una matriz de entradas reales de $(n \times p)$, $n > p$, donde se pide que el $rgo(X) = p$; o sea, que las columnas de X son linealmente independientes.

A continuación se plantea un problema que da lugar a un modelo lineal con tres variables para estudiar la satisfacción de un paciente al haber egresado de un hospital

Ejemplo 2.1.1 (Neter, pag. 266) *:El administrador de un hospital desea estudiar la relación entre la satisfacción de un paciente* (y), *la edad del paciente* $(x_1$, *en años*$)$, *severidad de la enfermedad* $(x_2$, *un índice*$)$ *y el nivel de ansiedad* $(x_3$, *un índice*$)$. *Se seleccionaron 23 pacientes al azar, colectando los siguientes datos, donde valores grandes de* y, x_2 *y* x_3 *están asociados respectivamente con mayor satisfacción, mayor severidad de la enfermedad y más ansiedad.*

$x_1 = AGE$ =*Edad en años.*

$x_2 = SEV$ =*Severidad de la enfermedad, (Índice).*

$x_3 = ANS$ =*Nivel de ansiedad, (Índice).*

$y \quad = SAS$ =*Satisfacción del paciente, (%).*

Los datos a este ejemplo son presentados en la figura 2.1, y los coeficientes del modelo de regresión ajustado se presentan en la figura 2.4, en la columna que tiene el nombre de "Beta", las otras estadísticas se revisarán en secciones posteriores.

Dados los valores de los parámetros el modelo de regresión lineal ajustado ajustado queda de la siguiente forma:

$$y = 162.9358 - 1.2187x_1 - .6731x_2 - 8.3501x_3$$

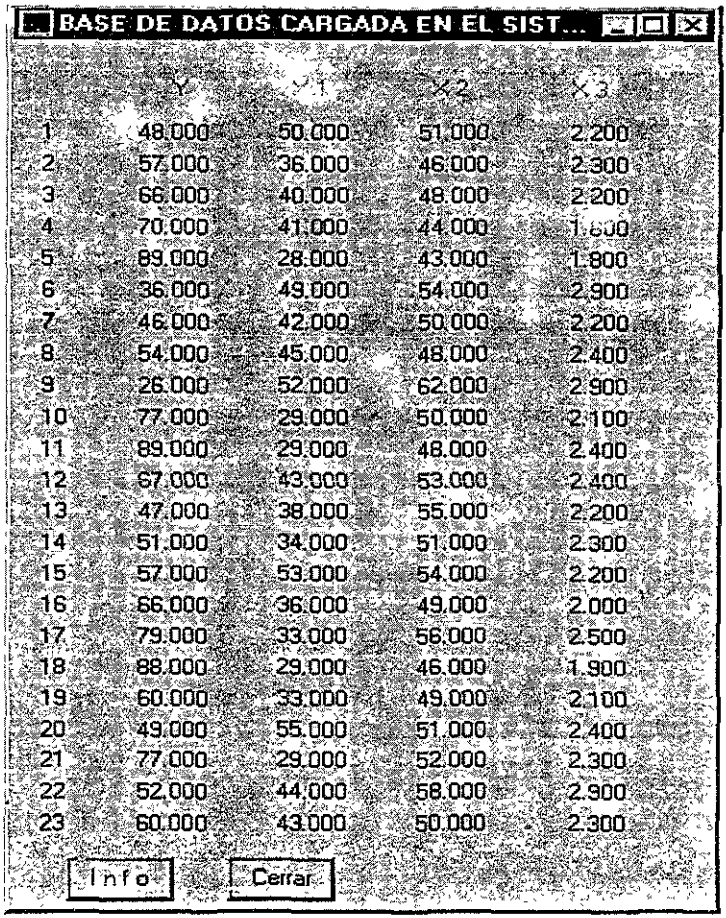


Figura 2.1: Datos del ejemplo 2.1.1.

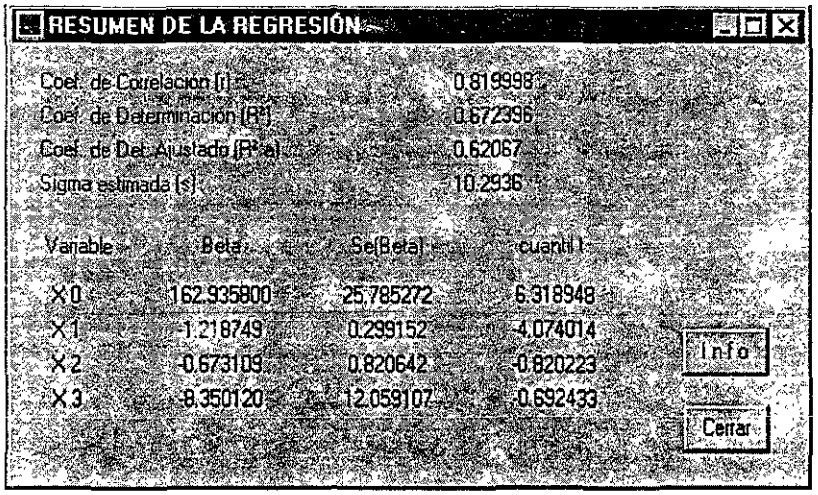


Figura 2.2: Resumen de la Regresión

Las siguientes secciones presentan algunas de las estadísticas que permiten analizar (estudiar o verificar) la validación del modelo.

2.2 Matriz de correlación.

La matriz de correlación de un modelo de Regresión Lineal Múltiple contiene todos los posibles coeficientes de correlación entre todas las parejas de regresores x_i, $i = 1, \ldots, p-1$, incluidas en el modelo, éstos están definidos como:

$$r_{ij} = \frac{cov\,(x_i, x_j)}{\sqrt{Var\,(x_i) * Var\,(x_j)}} = \frac{cov\,(x_i, x_j)}{se\,(x_i) * se\,(x_j)}$$

donde:

$$cov\,(x_i, x_j) = \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)\,(x_{kj} - \overline{x}_j) \quad y \quad Var\,(x_i) = \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2$$

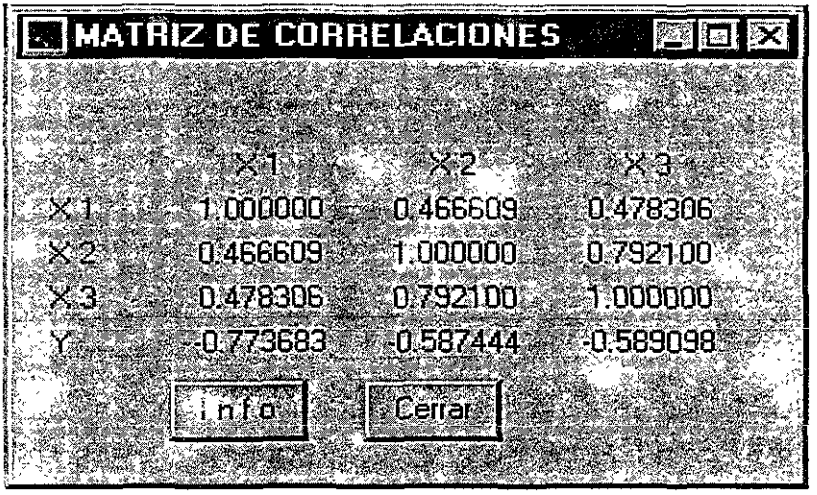


Figura 2.3: Matriz de Correlaciones.

si el valor de $|r_{ij}|$ tiende a 1, indica que los respectivos regresores i y j, presentan una estrecha relación lineal y si el coeficiente tiende a 0 indica una pobre relación entre dichos regresores.

Ejemplo 2.2.1 *La Matriz de Correlaciones del Ejemplo 2.1.1 se muestra en figura 2.3. De ésta matriz se tiene que los regresores mayormente correlacionados son ANS y SEV, con* $|r_{23}| = .7921$ *y los menos correlacionados son SEV y AGE con* $|r_{12}| = .4666$.

2.3 Errores estándares.

Los errores estándares y las covarianzas estimadas de los β_i, $i = 0, \ldots, p-1$ son encontradas a partir de σ^2 y $(X^tX)^{-1}$, si σ^2 es conocida; y de s^2 y $(X^tX)^{-1}$, si σ^2 es desconocida. Para calcular los errores estándares y las covarianzas se supondrá el segundo caso, σ^2 desconocida.

De la sección 1.2.1 se tienen los siguientes resultados:

$$E\left(\underline{\hat{\beta}}\right) = \underline{\beta} \quad \text{y} \quad Var\left(\underline{\hat{\beta}}\right) = \sigma^2 \left(X^tX\right)^{-1}$$

de las cuales se obtienen las siguientes expresiones:

$$se\left(\hat{\beta}_i\right) = s\sqrt{c_{ii}}$$
$$cov\left(\hat{\beta}_i, \hat{\beta}_j\right) = s^2 c_{ij}$$

donde c_{ii} es el i-ésimo elemento de la diagonal de la matriz $C = (X^tX)^{-1}$ y c_{ij} es el elemento en el i-ésimo renglón y la j-ésima columna de la misma matriz, sin considerar en ambos casos el renglón y la columna correspondiente a la constante.

Ejemplo 2.3.1 *Para el Ejemplo 2.1.1, la matriz* $(X^tX)^{-1}$ *es:*

	const	*AGE*	*SEV*	*ANS*
const	6.27496	.00151	−.15099	60083
AGE	.00151	.00084	−.00037	−.00685
SEV	−.15099	−.00037	.00635	−.06840
ANS	.60083	−.00685	−.06840	1.37245

Entonces se tiene que el error estándar de $\hat{\beta}_1$ *y la covarianza entre* $\hat{\beta}_1$ *y* $\hat{\beta}_2$ *son:*

$$se\left(\hat{\beta}_1\right) = s\sqrt{c_{11}} = 10.29357 * \sqrt{.000845} = .29922$$
$$cov\left(\hat{\beta}_1, \hat{\beta}_2\right) = s^2 c_{12} = (10.29357)^2 * (-.00037) = -.039204$$

El error estándar del estimador $\hat{\beta}_1$ es una medida de la precisión con la que el estimador ha sido calculado i.e., mientras mayor sea su valor la estimación será menos exacta.

2.4 Estadística t-student.

Una pregunta interesante, una vez ajustado el modelo es saber el peso que tienen cada uno de los regresores en la explicación de la variable de respuesta y considerando la existencia de otras variables en el modelo, la estadística t

proporciona una medida de tal influencia, mientras más grande es el valor que toma esta estadística, más grande es el peso de la variable regresora en el modelo, y si es pequeño indica que el regresor tiene poco peso en la explicación de la variable de respuesta y; es por ello que el valor de esta estadística puede sugerir la eliminación de ciertas variables en el modelo de regresión. El cálculo de esta estadística se recomienda que se efectúe después de la validación del modelo.

En el caso de que no exista colinealidad, y un regresor presente una estadística t pequeña, se puede recurrir a la siguiente prueba de hipótesis para determinar si dicho regresor puede ser eliminado del modelo:

Hipótesis:

$$H_0 : \beta_i = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0; \qquad i = 0, \ldots, p-1$$

Estadístico:

$$t_i = \frac{\hat{\beta}_i}{se\left(\hat{\beta}_i\right)}$$

Regla de decisión: La estadística t_i se compara con el cualtil $\left(1 - \frac{\alpha}{2}\right)$ que tiene la distribución t-student con $(n-p)$ grados de libertad, denotado por $t^{1-\frac{\alpha}{2}}_{(n-p)}$, esto al nivel de significancia α, y si el valor absoluto de la estadística es mayor al valor encotrado en las tablas de la distribución, se rechaza la hipótesis de que $\beta_i = 0$, en caso contrario se acepta.

Ejemplo 2.4.1 *Para el Ejemplo 2.1.1, la figura 2.4 muestra el Resumen estadístico de la Regresión:*

Como ilustración se considera la siguiente prueba de hipótesis:

$$H_0 : \beta_i = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0; \quad i = 0, 1, 2, 3$$

el valor de la estadística $t^{1-\frac{\alpha}{2}}_{(19)}$ *de tablas, con un nivel de significancia* $\alpha = .05$ *es* $t^{975}_{(19)} = 2.093$; *luego, para los regresores* x_2 *y* x_3 *se tiene que el valor absoluto de* t_i *es menor que el valor de* $t^{975}_{(19)}$, *por lo que se inclinaría a no rechazar la hipótesis nula* H_0 *para estos dos regresores, lo que indica que son posibles candidatos para descartarse del modelo.*

A menudo se recurre al uso de esta estadística cuando se tiene el problema de querer seleccionar el número mínimo de regresores que expliquen de forma

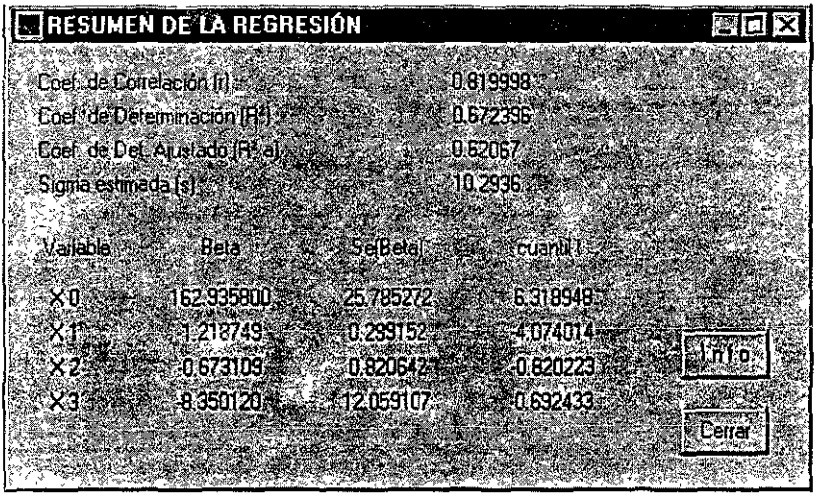


Figura 2.4: Resumen de la Regresión

adecuada la variable de respuesta y, para hacer dicha selección se realiza el siguiente procedimiento:

1. Se ordenan los coeficientes de regresión en forma decreciente con respecto a la magnitud de la estadística $|t_i|$ donde $i = 1, \ldots, p-1$, y se introduce en el modelo un regresor a la vez en este orden, lo cual permite encontrar el mejor o uno de los mejores modelos reducidos.

2. En el caso de que haya varias variables $(x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ que tienen valores de t muy pequeños y se sospecha que se pueden eliminar se recurre a una prueba de hipótesis en términos de la estadística F, en la que se discute la siguiente prueba de hipótesis:

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_m} = 0 \qquad Vs. \qquad H_1 : \beta_{i_j} \neq 0, \quad \text{para algún } j = 1, \ldots, m$$

Nota: Como se mencionó previamente éste procedimiento funciona cuando no se presenta colinealidad.

2.5 Coeficiente de determinación.

Una pregunta de interés es cómo medir la proporción de variabilidad en y explicada por la regresión sobre las $x_i's$, ésto se suele hacer en términos del Coeficiente de Determinación el cual se define como sigue:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \tag{2.2}$$

y es por ello que se utiliza conjuntamente con otro tipo de estadísticas y pruebas para establecer un diagnóstico completo del modelo de regresión.

Un R^2 grande no necesariamente implica que el modelo ajustado es muy útil. Adicionar más regresores al modelo puede sólo incrementar R^2 y nunca reducirlo, porque SCE no puede nunca llegar a ser más grande con más regresores y SCT es siempre el mismo para un conjunto dado de respuestas. La interpretación que se da a éste coeficiente debe ser cuidadosa pues en ocasiones se puede tener R^2 significativo pero el ajuste no ser muy bueno, para ello se puede hacer uso del siguiente indicador.

Coeficiente de determinación múltiple ajustado.

El Coeficiente de Determinación Múltiple Ajustado, denotado por R_a^2, ajusta R^2 dividiendo cada Suma de Cuadrados por sus grados de libertad.

$$R_a^2 = 1 - \frac{\frac{1}{n-p}SCE}{\frac{1}{n-1}SCT} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SCE}{SCT} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Este valor puede llegar a ser más pequeño cuando otro regresor es introducido al modelo, porque la disminución en SCE puede ser más que compensado por la pérdida de un grado de libertad en el denominador $n-p$ Es conveniente que los valores R_a^2 y R^2 no difieran mucho entre sí, en caso de que llegara a ocurrir esto, entonces se puede pensar que el modelo ha sido sobrespecificado; esto es, términos que no contribuyen significativamente al ajuste han sido incluidos.

En el ejemplo 2.1.1 se tienen los siguientes valores para estas estadísticas: $R^2 = 67239$ y $R_a^2 = .62066$, (obtenidos de la figura 2.4), por lo que se puede decir que aproximadamente el 68% de la variabilidad observada en la respuesta es modelada por las $x_i's$, ya que $R^2 = .67239$; ahora bien, como ambas estadísticas difieren en 5 puntos porcentuales aproximadamente, se

podría considerar que el modelo tiene al menos un regresor que no contribuye en forma significativa al ajuste del modelo.

No se puede llevar a cabo ninguna prueba de hipótesis con esta estadística debido a que las expresiones dadas en el cociente en (2.2) no son independientes, lo cual impide encontrar alguna distribución probabilística que permita llevar a cabo tales propósitos.

A pesar de esto, su uso es común. Aunque debe interpretarse con cuidado, ya que se puede obtener un Coeficiente de Determinación cercano a 1 sin saber con exactitud, y en base a cierta probabilidad, qué tan representativo sea este valor en cuanto a la explicación del ajuste del modelo.

En general, si el coeficiente de determinación R^2 tiende a 1 ésto indica que el modelo lineal se ajusta de manera aceptable a los datos.

2.6 Prueba de hipótesis general.

En la sección 2.4, se vio que el valor de la t puede sugerir la eliminación de ciertos regresores en el Modelo Lineal General (MLG), otra alternativa es que el investigador decida descartar algunos regresores en dicho modelo.

En esta sección se revisará la prueba de hipótesis que permite tomar la decisión de eliminar varios regresores a la vez que se suponen no son relevantes para el modelo; para el caso de una sola variable, en la sección 2.4 se vio como resolver este problema.

Si se considera el modelo lineal general:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \qquad \underline{\varepsilon} \sim N\left(\underline{0}, \sigma^2 I\right).$$

se obtiene un Modelo Lineal Reducido (MLR), éste modelo tiene sólo un subcojunto de todas las variables que se incluyen en el modelo inicial por lo que el número de parámetros considerados en el MLR es menor que el número de parámetros del MLG.

En la hipótesis nula se prueban algunos parámetros específicos para los respectivos regresores, quedando un modelo reducido (MLR), (un criterio para la elección de estos coeficientes es el que se basa en los valores de la estadística t) y se prueban las hipótesis siguientes:

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_k} = 0 \qquad Vs. \qquad H_1 : \beta_{i_j} \neq 0, \quad \text{para alguna } j.$$

El procedimiento que lleva a tomar la decisión de rechazar o aceptar la hipótesis nula es el siguiente:

1. Obtener los valores de $\hat{y}$ y $\hat{y'}$ que son los valores estimados que se fijan con el MLG y el MLR respectivamente.

2. La ausencia de ajuste relacionada a los datos en el MLG es la suma de cuadrados del error o de los residuales, denotada por $SCE\,(MLG)$. Esto es,

$$SCE\,(MLG) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

en el caso del MLR tal cantidad es:

$$SCE\,(MLR) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i'\right)^2$$

Para el MLG se supone que se tienen p parámetros, si se considera la presencia de una constante (β_0) en el modelo y $p-1$ en otro caso, sin pérdida de generalidad, supondremos que en el MLR se considera la estimación de la constante (β_0), entonces, para el MLR se consideran k parámetros.

3 Para saber qué tan bien se adecua el MLR, se compara

$$SCE\,(MLR) - SCE\,(MLG) \text{ con } SCE\,(MLG)$$

utilizando para ello el siguiente cociente:

$$F = \frac{[SCE\,(MLR) - SCE\,(MLG)]\,/\,(p-k)}{SCE\,(MLG)\,/\,(n-p)} \tag{2.3}$$

donde las constantes $(p-k)$ y $(n-p)$, en el numerador y denominador, sirven para compensar el número de parámetros que se involucran en los dos modelos y para obtener las distribuciones ji-cuadrada (véase sección 1 4) que permiten construir la estadística de prueba con distribución F, con $(p-k)$ y $(n-p)$ grados de libertad.

4. El valor resultante del cociente en (2.3), se compara con el cuantil $(1-\alpha)$ de la distribución F con $(p-k)$ y $(n-p)$ grados de libertad, y a un nivel de significancia α, si el cociente es mayor que el cuantil $(1-\alpha)$ de la distribución F, se rechaza la hipótesis nula, y se considera que el MLR no explica de manera satisfactoria la relación que existe entre los regresores y la variable de respuesta.

Un caso particular de las pruebas de hipótesis son las que se utilizan para la validación del modelo

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon},$$

y que se verán a continuación.

2.7 Validación del modelo de regresión lineal ajustado.

Al plantearse un modelo de regresión siempre se incluyen ciertas suposiciones (o hipótesis de trabajo), tales como:

- Los errores tienen media cero y matriz de varianza-covarianza $\sigma^2 I$ (equivalente a la independencia de los errores y la homocedasticidad de las varianzas de los errores).
- Los errores tienen distribución normal.

Con base en estas suposiciones es que se pueden efectuar todas las pruebas de hipótesis que permiten el análisis estadístico del modelo.

Resulta natural la pregunta de verificar si estadísticamente existe una relación lineal entre los regresores (variables independientes) y la variable de respuesta (dependiente)

A continuación se presenta el procedimiento para efectuar tal verificación, usualmente conocido como validación del modelo.

Esto se lleva a cabo a partir del siguiente ensayo de hipótesis:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0, \quad \text{para alguna } i \qquad (2.4)$$

Cuando la hipótesis nula (H_0) se acepta, estadísticamente se tiene que no existe evidencia de que los regresores $x_1, \ldots, x_{p-1}$ se asocien linealmente

a la variable de respuesta y, en caso contrario, cuando se rechaza la hipótesis nula estadísticamente se tiene que sí existe evidencia de que los regresores x_i se asocian linealmente a la variable de respuesta.

Este procedimiento se lleva a cabo con el Análisis de Varianza, con base a la estadística F. Para esto se supone que todos los coeficientes de regresión son cero (excepto la constante), equivalente a suponer que no hay una asociación lineal entre los regresores y la variable de respuesta. Esto se logra con el planteamiento de hipótesis dado por (2.4).

Esta prueba estadística se basa en la varianza obtenida por el modelo completo

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

la cual es comparada con el modelo sin regresores (dado por H_0) mediante las sumas de cuadrados, donde $SCR = SCT - SCE$ es la suma de cuadrados de y explicada por el modelo lineal general que no es explicada por el modelo lineal reducido. Los grados de libertad asociados a la SCR es el número de restricciones dadas en la hipótesis nula, que en este caso son $(p-1)$, el de la SCE es $(n-p)$ y el de SCT en la suma de los grados de libertad de los anteriores. Estos datos son presentados en la siguiente tabla:

Tabla de Análisis de Varianza (ANOVA) $\left(H_0 : \beta_1 = \cdots = \beta_{p-1} = 0\right)$.

Fuente de Variación	*Sumas de Cuadrados*	*Grados de Libertad*	*Cuadrados Medios*	F
Re*gresión*	SCR	$p-1$	$SCR/(p-1)$	$\frac{SCR/(p-1)}{SCE/(n-p)}$
Re*sidual*	SCE	$n-p$	$SCE/(n-p)$	
Total	SCT	$n-1$		

Si la estadística F dada en esta tabla es mayor que el cuantil de la distribución F, denotado por $F^{1-\alpha}_{(p-1,n-p)}$, con $(p-1)$ y $(n-p)$ grados de libertad, y un nivel de significancia α, se concluirá que la información dada por los regresores al modelo completo es significativamente mejor que aquel modelo que no las toma en cuenta (modelo reducido), por lo que se rechaza la hipótesis nula (H_0).

Esta prueba se utiliza cuando se desea verificar qué tan bien se ajustan los datos al modelo encontrado al estimar el vector de parámetros $\hat{\underline{\beta}}$ del modelo lineal general.

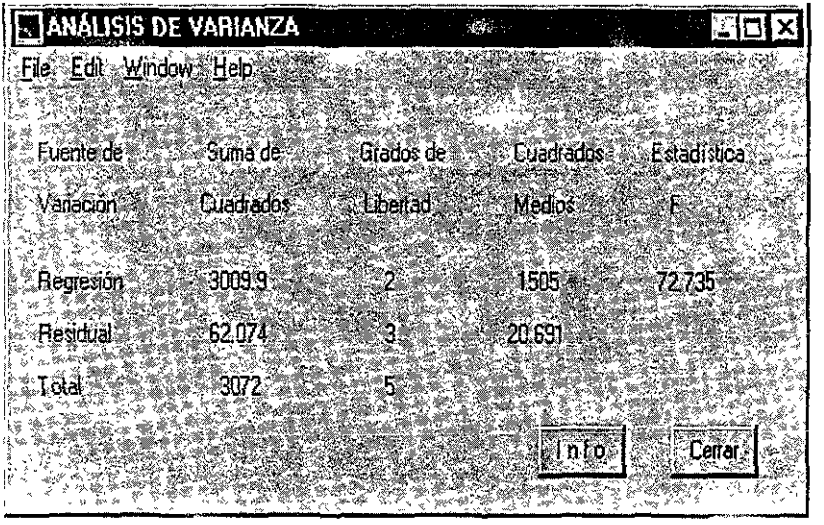


Figura 2.5: Análisis de Varianza.

Ejemplo 2.7.1 *La Tabla de Análisis de Varianza para el Ejemplo 1.4.1 se muestra en la figura 2.5:*

Ahora se obtiene el cuantil de la distribución F con un nivel de significancia $\alpha = 05$, es decir, $F^* = F^{95}_{(2,3)} = 9.55$ y este valor es comparado con el estadístico de la tabla $F = 72.734$, como este valor cae en la región de rechazo dada por F^*, se puede concluir que al menos uno de los regresores (x_1 ó x_2 ó ambos) explica la variación en el modelo dado por la variable de respuesta y.

El rechazo de la hipótesis nula $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ indica que el modelo lineal es adecuado o que estadísticamente existe una relación lineal entre los regresores $(x_1, \ldots, x_{p-1})$ y la variable de respuesta (y)

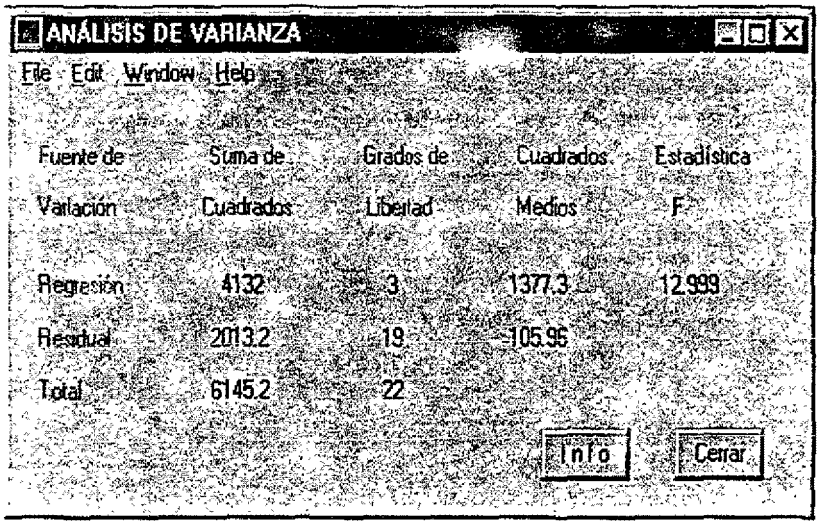


Figura 2.6: Análisis de Varianza.

Ejemplo 2.7.2 *La Tabla ANOVA del Ejemplo 2.1.1 es mostrada en la figura 2.6, para realizar la prueba de hipótesis:*

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0, \quad \text{para algún } i = 1, 2, 3$$

Ya que $F = 12.998$ *excede el valor del cuantil* $F^{95}_{(3,19)} = 3.13$, *se puede concluir que la satisfacción del paciente es explicado por al menos alguno de los regresores que son considerados en el modelo.*

2.8 Colinealidad.

Una de las suposiciones más importantes en el análisis de regresión, es que los regresores x_i son independientes entre sí o que al menos no están fuertemente correlacionados, si tal suposición no se cumple, el cambio que ocurra en algún regresor puede implicar cambios en otro regresor, y esto provoca una interpretación equivocada de los coeficientes de regresión, ya que se considera

que el i-ésimo coeficiente de regresión es "la medida del cambio en la variable de respuesta cuando el regresor x_i es incrementado en una unidad, y los regresores restantes permanecen constantes" (Montgomery [8], pag. 110); y esto como se mencionó no ocurre en la mayoría de los casos.

Si se observa que no existe alguna relación entre los regresores, se dice que son independientes, sin embargo, en la práctica la mayoría de los casos estudiados presentan regresores no independientes por lo que se considera que existe colinealidad o datos colineales.

La presencia del fenómeno de *colinealidad* en la matriz de datos del modelo de regresión lineal, ha provocado una gran preocupación a estadistas y analistas numéricos. Es un tema controvertido el cual es abordado con técnicas estadísticas, aunque para otros es un problema numérico y lo abordan con técnicas finas del Análisis Numérico.

La *colinealidad* para un estadista significa redundancia en las variables de control en el modelo de regresión lineal, y le interesa su diagnóstico pues esto significa la posible aparición de estimadores mínimos cuadrados inaceptables y/o inflación de la varianza de tales estimaciones, obteniendo pruebas de hipótesis poco confiables que llevan a la toma de decisiones equivocadas. La detección de *colinealidad* también conlleva a una selección de aquellas variables de control que no son redundantes para el modelo y que "mejor expliquen" a los datos observados (véase sección 3.4).

El diagnóstico de *colinealidad;* para una analista numérico significa la detección de rango deficiente de una matriz, lo cual puede dar lugar a una fuerte acumulación de errores por redondeo en su tratamiento numérico. A la hora de detectar colinealidad, el analista numérico está interesado en determinar el rango numérico de la matriz del modelo y un subconjunto de sus columnas que "mejor genere" al espacio imagen de tal matriz (Ver López & Madrid [6]).

2.8.1 Perspectiva histórica.

Se han empleado muchos procedimientos para la detección de la colinealidad. Los más comúnmente usados son los siguientes (Belsley [1]), se indicarán sus problemas y debilidades.

1. Hipotéticamente, los signos de los parámetros estimados $\hat{\beta}_i$ son incorrectos. Regresores que se consideran "importantes" en la explicación

de la variable de respuesta, tienen estadísticos t muy chicos. Hay grandes cambios en el vector de estimaciones $\hat{\underline{\beta}}$, al borrar un renglón o una columna de X.

Desafortunadamente, ninguna de estas condiciones es necesaria o suficiente para la existencia de colinealidad, y técnicas más avanzadas son requeridas para detectar la presencia de colinealidad.

2. La examinación de la matriz de correlaciones C de los regresores, para observar el nivel de asociación entre los distintos regresores x_i y x_j, $i \neq j$ (véase sección 2.2); o la inversa de esta matriz, C^{-1}.

   La matriz de correlaciones, por si sola, no siempre es buena para la detección de colinealidad, pues sólo presenta correlaciones dos a dos y se puede tener una fuerte dependencia de algún x_i con los restantes x_j.

   Al observar la inversa de la matriz de correlaciones, C^{-1} se tienen indicadores importantes, conocidos como Coeficientes de Inflación de la Varianza (CIV) y denotados por:

$$CIV_j = c_{jj}^{(-1)} = \frac{1}{1 - R_j^2}; \qquad \text{con } C^{-1} = \left\{c^{(-1)}\right\}_{ij}$$

   donde $c_{jj}^{(-1)}$ es el $j-$ésimo elemento en la diagonal de la matriz C^{-1}, y R_j^2 es el coeficiente de determinación múltiple (véase sección 2.5) obtenido al haber ajustado el $j-$ésimo regresor en términos de los $j-1$ regresores restantes, i. e.,

$$x_j = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_{p-1} x_{p-1}$$

   Cuando R_j^2 tiende hacia 1 indica la presencia de una relación lineal entre los regresores, y el CIV para el coeficiente estimado de x_j tiende a infinito. Esto sugiere que cuando un CIV excede el 10, la colinealidad puede causar problemas en la estimación de los parámetros.

   La desventaja de éste método es que es débil en su cálculo pues es inestable -numéricamente-, debido a la colinealidad presente, por lo que no son del todo confiables los resultados.

   Nota: Si cada uno de los regresores de la matriz de datos X es estándarizado, $\tilde{X}$ (véase sección 1.3.1), entonces la matriz $\tilde{X}^t \tilde{X}$ será la matriz de correlaciones C, entre los regresores y $C^{-1} = \left(\tilde{X}^t \tilde{X}\right)^{-1}$.

3. Otra técnica para el diagnóstico, empleando información de las matrices C y C^{-1}, es el mencionado por Farrar y Glauber (1967).

   Bajo la suposición de que X tiene columnas ortogonales, ellos discuten que una transformación del $\det(C)$ es aproximadamente distribuida como una χ^2 y por lo tanto provee una medida de la desviación de la ortogonalidad o la presencia de colinealidad. Además proponen el uso de la medida

$$r_{ij} = \frac{-c_{ij}^{(-1)}}{\sqrt{c_{ii}^{(-1)}}\sqrt{c_{jj}^{(-1)}}},$$

   que es, la correlación parcial entre x_i y x_j, ajustadas por todos los otros regresores, para investigar las interdependencias en mayor detalle. Estas técnicas han sido presas de varias críticas.

   Primera, el uso del $\det(C)$ no puede diagnosticar la presencia de varias dependencias cercanas coexistentes, la existencia de una dependencia cercana hará un $\det(C)$ cercano a cero.

   Segunda, el uso de los elementos r_{ij} carece de discriminación. Estas correlaciones, todas se aproximan a la unidad (± 1) cuando la colinealidad llega a ser más molesta. De este modo r_{ij} puede ser cercano a la unidad aunque las variables x_i y x_j no estén envueltas en cualquier relación colineal (Belsley [1], apéndice 3C).

4. El análisis de los eigenvalores y eigenvectores de la matriz de correlaciones: C Kendall (1957) y Silvey (1969) han sugerido el uso de los eigenvalores de $X'X$ como una llave a la presencia de colinealidad, la cual dicen que es indicada por la presencia de un eigenvalor "pequeño". Desafortunadamente no se está informado de lo que es "pequeño", y hay una tendencia natural por comparar pequeño con el cero. En algunos casos es indicado, pero sin justificación, que la colinealidad puede existir si "un eigenvalor es pequeño en relación a los otros". Aquí, pequeño es interpretado en relación a los más grandes eigenvalores que en relación al cero.

Una vez que se ha detectado el problema de colinealidad en los regresores, es necesario, eliminarlo o disminuirlo al máximo, de tal manera que no afecte en forma considerable en la estimación de parámetros del modelo de regresión lineal.

2.8.2 Remedios para eliminar la colinealidad.

Técnicas avanzadas se han desarrollado para la eliminación del problema de colinealidad. En éste trabajo sólo se hará mención a ellas, indicando la bibliografía donde el lector puede profundizar sobre cada una de estas técnicas.

Componentes Principales.

En este método se construye el modelo de regresión lineal dado por:

$$\underline{y} = Z\underline{\alpha} + \underline{\varepsilon}$$

donde

$$Z = XQ, \quad \underline{\alpha} = Q^t\underline{\beta}, \quad Q^tX^tXQ = Z^tZ = \Lambda$$

recordando que $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ es una matriz diagonal con los eigenvalores de X^tX y Q es una matriz ortogonal donde sus columnas son los eigenvectores asociados a $\lambda_1, \lambda_2, \ldots, \lambda_p$. Las columnas de Z, definen un nuevo conjunto de regresores ortogonales.

El estimador de mínimos cuadrados de $\underline{\alpha}$ es

$$\hat{\underline{\alpha}} = \left(Z^tZ\right)^{-1} Z^t\underline{y} = \Lambda^{-1}Z^t\underline{y}$$

y la matriz de covarianzas de $\hat{\underline{\alpha}}$ es:

$$Var\left(\hat{\underline{\alpha}}\right) = \sigma^2\left(Z^tZ\right)^{-1} = \sigma^2\Lambda^{-1}$$

El método de regresión con Componentes Principales combate la colinealidad usando menos del conjunto completo de Componentes Principales en el modelo. Para obtener el estimador de Componentes Principales, se asume que los regresores son arreglados en orden decreciente, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Suponga que los últimos s de estos eigenvalores son aproximadamente igual a cero. En la Regresión con Componentes Principales se eliminan los estimadores asociados a x_j que tienen un valor λ_j cercano a cero.

Esto es,

$$\hat{\underline{\alpha}}_{CP} = B\,\hat{\underline{\alpha}}$$

donde

$$B = \begin{bmatrix} I_{p-s} & 0 \\ 0 & 0 \end{bmatrix}$$

Entonces el estimador de Componentes Principales es

$$\underset{\wedge}{\underline{\alpha}}_{CP} = \begin{bmatrix} \underset{\wedge}{\alpha}_1 \\ \vdots \\ \underset{\wedge}{\alpha}_{p-s} \\ \underline{0}_s \end{bmatrix}$$

o en términos de los regresores originales

$$\underset{\wedge}{\underline{\beta}}_{CP} = Q \underset{\wedge}{\underline{\alpha}}_{CP} = \sum_{j=1}^{p-s} \lambda_j^{-1} \underline{q}_j \underline{q}_j^t X^t \underline{y}$$

Un estudio de Simulación por Gunst y Mason (1977) muestra que la Regresión con Componentes Principales ofrece considerables mejoramientos sobre los mínimos cuadrados cuando los datos son mal condicionados.

Referencias: Montgomery & Peck [8].

Selección de Variables.

Para la Selección de Variables existen varios métodos a aplicar al modelo de regresión lineal, entre ellos están:

1. Método Forward.
2. Método Backward.
3. Estadística C_p de Mallows.

Estas técnicas de Selección de Variables serán revisadas más a fondo en la sección 3.4, por lo que son omitidas en esta sección.

Referencias: Montgomery & Peck [8], Neter [11], Ryan [13].

Regresión Ridge.

La Regresión Ridge es uno de varios métodos que han sido propuestos como remedio al problema de la estimación de los parámetros del modelo de regresión lineal bajo colinealidad, modificando el método de mínimos cuadrados para obtener estimadores sesgados. El vector de estimadores $\underset{\wedge}{\underline{\beta}}$ usual, cumple con la propiedad de ser insesgado pero impreciso, mientras que el estimador

ridge $\hat{\underline{\beta}}_R$ es mucho más preciso pero presenta un sesgo pequeño. La probabilidad de que $\hat{\underline{\beta}}_R$ caiga cerca del valor verdadero $\underline{\beta}$ es mucho más grande que para el estimador insesgado $\hat{\underline{\beta}}$.

Para los mínimos cuadrados ordinarios, las ecuaciones normales están dadas por:

$$X'X\,\hat{\underline{\beta}} = X'y \tag{2.5a}$$

Los estimadores para la regresión ridge son obtenidos después de introducir en las ecuaciones normales (2.5a) una constante $k \geq 0$, en la siguiente forma:

$$\left(X'X + kI\right)\hat{\underline{\beta}}_R = X'y \tag{2.6}$$

donde $\hat{\underline{\beta}}_R$ es el vector de parámetros estimados para la regresión ridge. La solución de las ecuaciones normales (2.6) estará dado por:

$$\hat{\underline{\beta}}_R = \left(X'X + kI\right)^{-1} X'y \tag{2.7}$$

La constante k refleja la cantidad del sesgo en los estimadores. Cuando $k = 0$, (2.7) se reduce a la estimación de los coeficientes de regresión para los mínimos cuadrados ordinarios.

Puede mostrarse que la componente del sesgo del error cuadrático medio total del estimador de la regresión ridge $\hat{\underline{\beta}}_R$ decrece cuando k tiende a crecer, mientras que la varianza llega a ser más pequeña. Puede mostrarse además que existe siempre algún valor k para el cual el estimador ridge $\hat{\underline{\beta}}_R$ tiene un error cuadrático medio total más pequeño que el estimador de mínimos cuadrados ordinarios $\hat{\underline{\beta}}$. La dificultad es que el valor óptimo de k varía de una aplicación a otra y es desconocido.

Referencias: Montgomery & Peck [8], Neter [11].

2.9 Resumen estadístico.

A continuación se presenta un resumen estadístico básico en el que se pueden observar el promedio, varianza, desviación estándar, y los valores mínimos y máximos de cada regresor. Estos resultados nos ayudan a tener una idea a

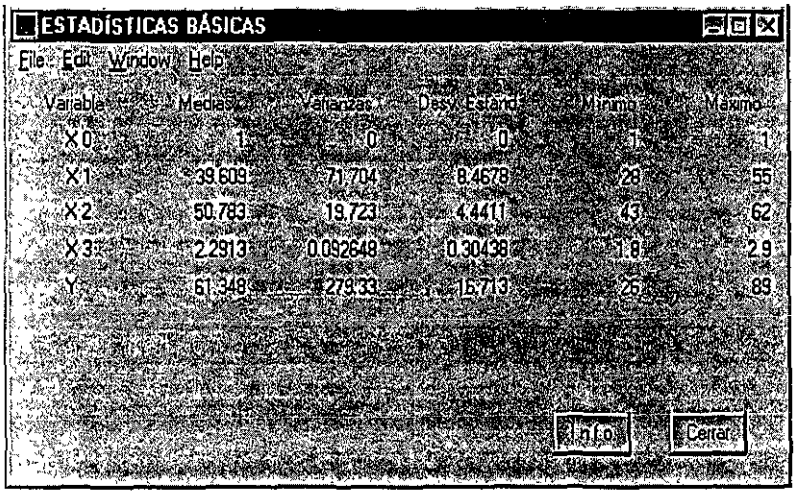


Figura 2.7: Estadísticas Básicas.

priori del posible comportamiento que tendrán los datos una vez ajsutados al modelo de regresión.

Para los datos del Ejemplo 2.1.1, la figura 2.7 muestra éste primer resumen estadístico.

Un segundo resumen que permite verificar en forma breve las características del modelo de regresión

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

incluye las estadísticas:

- La estimación de $\hat{\underline{\beta}}$.
- La estimación de la varianza, s^2.
- La estadística F.
- El Coeficiente de Determinación R^2 y el ajustado R_a^2.
- Matriz de $(X^tX)^{-1}$.

- Matriz de Correlación.

2.9.1 Aspectos numéricos en el cálculo de la media y la varianza.

Cuando se efectúan operaciones aritméticas en una computadora, se incurre en errores por redondeo, es particularmente importante evitar cancelación numérica, y para conseguirlo se recurre a algoritmos que efectúan los cálculos numéricos de la manera más exacta posible. En los casos particulares del cálculo de la media y la varianza, uno de los procedimientos que se puede utilizar es el siguiente:

Para la media.

1. Ordenar los datos en forma descendente.
2. Sumarlos en el orden en que fueron arreglados
3. Dividir la suma obtenida en el paso anterior entre el número de datos.

Para la varianza.

1. Restar a cada dato su media y elevar el resultado al cuadrado.
2. Ordenar los resultados obtenidos en el paso 1 en forma descendente.
3. Sumarlos.
4. Dividir la suma anterior entre el número de datos menos 1.

Este procedimiento que se sugiere en Thisted [18], pag. 10, es costoso en tiempo de máquina por el número de operaciones que se realizan, ya que es necesario ordenar en dos ocasiones los datos, es por ello que para el cálculo de la media y la varianza se prefirió utilizar el algoritmo desarrollado por R. J. Hanson [5].

En el siguiente ejemplo se verá la aplicación de las estadísticas mencionadas en ésta sección.

Ejemplo 2.9.1 *Como parte de un estudio para investigar la relación que existe entre el estrés y otras variables, se recopilaron los siguientes datos de una muestra aleatoria de 15 ejecutivos asociados. Las medidas fueron:*

x_1 =*Medida de la importancia de la empresa.*

x_2 =*Número de años en la presente posición.*

x_3 =*Salario anual/1000.*

x_4 =*Edad.*

y = *Medida del estrés*

- *Matriz de Correlación:*

  En general, las correlaciones entre los regresores no son muy grandes, excepto $r_{12} = 50108$, *este valor no se puede considerar muy grande para concluir que existen problemas de colinealidad, pero se deja a consideración del analista el decidir sobre el valor de esta estadística.*

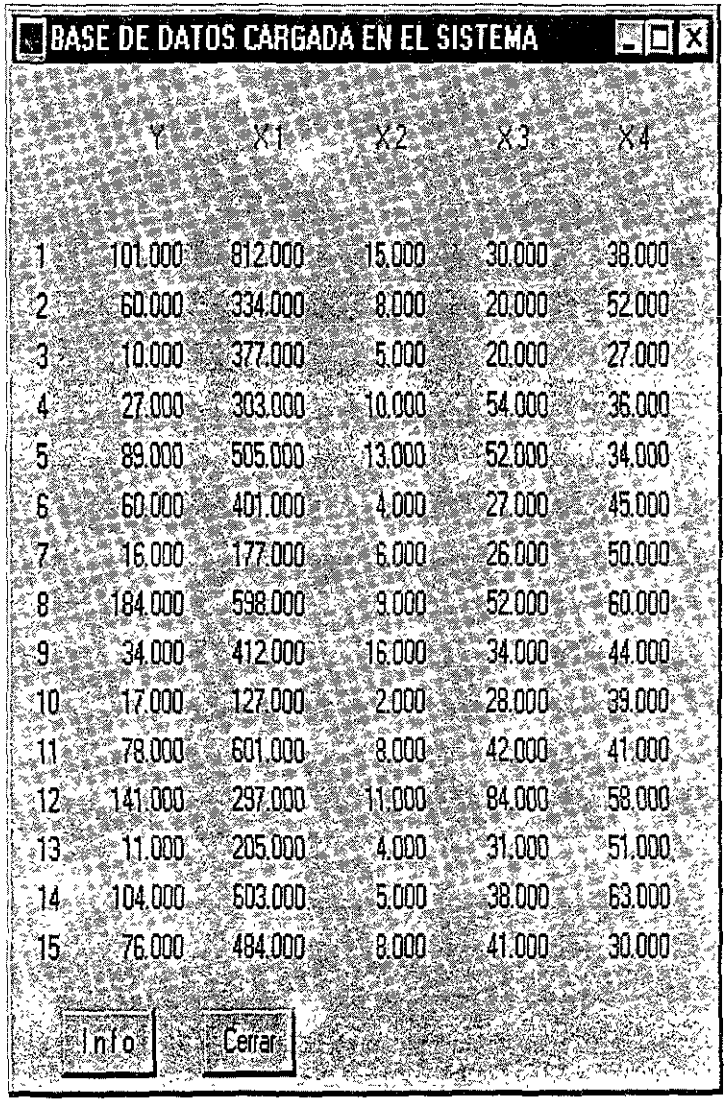
- *Validación del modelo y estadísticas t.*

  Tomando un nivel de significancia $\alpha = .05$, *tenemos que el cuantil* $F^{95}_{(4,10)} = 3.48$, *es menor que el valor* $F = 13.365$ *dado en la figura 2.11, lo cual lleva a la conclusión de que al menos uno de los regresores tomados explica la variación de la Medida del estrés* (y).

  El cuantil t-student es $t^{975}_{(10)} = 2.228$, *que comparado con* $|t_i|$, $i = 1, \ldots, 4$, *estadística t para cada regresor, dado en la figura 2.10, se tiene que el Número de años en la presente posición* (x_2), *es menor que el cuantil dado por lo que se concluye que la variable número de años tiene poco peso en la explicación del estrés.*

- *Coeficientes de Determinación.*

  En la figura 2.10 se observa que $R^2 = 84242$, *lo cual indica que los regresores explican aproximadamente el 84% de la variación total sobre la variable de respuesta que es el estrés. Para este mismo caso* $R^2_a = .77939$, *mostrado en la misma figura, por lo que se verifica que existe una diferencia significativa entre ambas estadísticas (7% aprox.), con lo cual se concluye que puede existir al menos un regresor en el modelo que no tiene gran relación con la variable de respuesta, y que hace que estos dos indicadores difieran en forma "significativa". Como se mencionó en la sección 2.5, la interpretación del* R^2 *debe ser cuidadosa*



	Y	X1	X2	X3	X4
1	101.000	812.000	15.000	30.000	38.000
2	60.000	334.000	8.000	20.000	52.000
3	10.000	377.000	5.000	20.000	27.000
4	27.000	303.000	10.000	54.000	35.000
5	89.000	505.000	13.000	52.000	34.000
6	60.000	401.000	4.000	27.000	45.000
7	16.000	177.000	6.000	26.000	50.000
8	184.000	598.000	9.000	52.000	60.000
9	34.000	412.000	16.000	34.000	44.000
10	17.000	127.000	2.000	28.000	39.000
11	78.000	601.000	8.000	42.000	41.000
12	141.000	297.000	11.000	84.000	58.000
13	11.000	205.000	4.000	31.000	51.000
14	104.000	603.000	5.000	38.000	63.000
15	76.000	484.000	8.000	41.000	30.000

Figura 2.8: Datos para el ejemplo del Estrés.

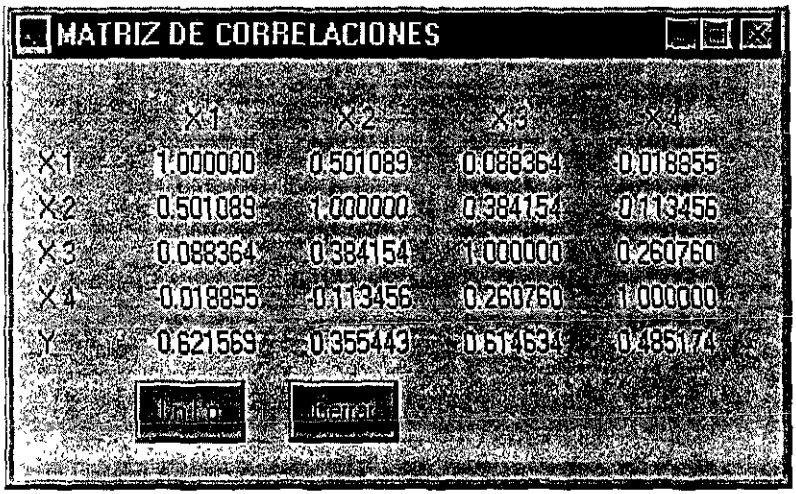


Figura 2.9 Matriz de Correlaciones

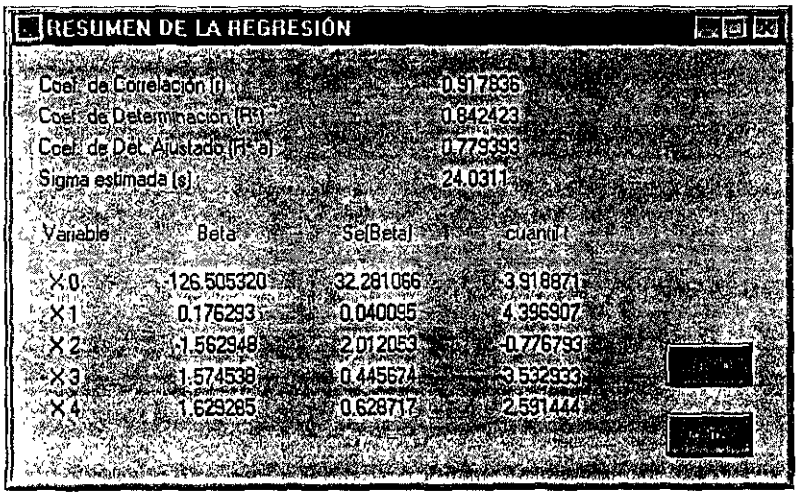


Figura 2.10 Resumen de la Regresión para el ejemplo del Estrés.

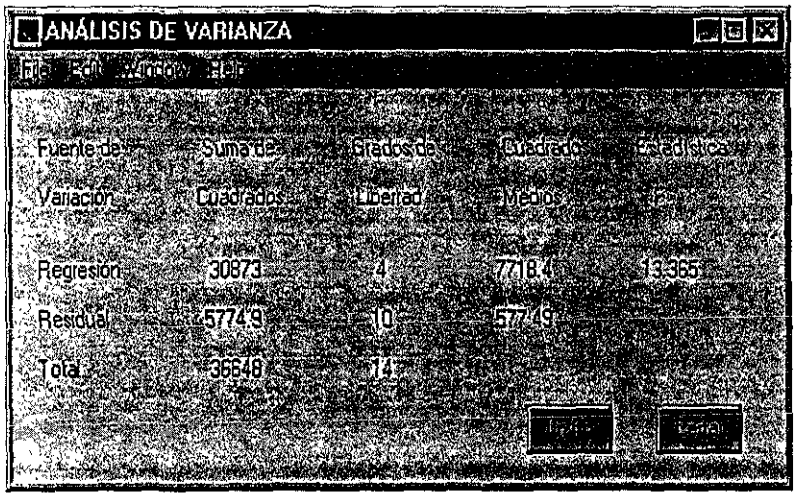


Figura 2.11: Análisis de Varianza para el ejemplo del estrés.

pues éste indicador sólo dice la explicación de los regresores sobre la variable de respuesta, sin importarle el número de regresores considerados en el modelo y que si son demasiados pueden llegar a traer problemas en los resultados, como se menciona en el siguiente punto.

- *Colinealidad.*

  La colinealidad es un problema que con frecuencia se presenta en el análisis de datos. La detección de este problema no tiene una solución práctica todavía. Pero aquí se hace referencia a este problema, únicamente observando las estadísticas básicas. Obsérvese la figura 2.12:

  El signo de cada uno de los estimadores encontrados debe coincidir con su respectivo coeficiente de correlación entre la variable de respuesta y el regresor, pues si se recuerda, en la sección 2.8 se mencionó que cada coeficiente de regresión representa la medida de cambio en la variable de respuesta, es por eso que los valores dados en la última figura deben coincidir en su signo. Lo cual lleva a la conclusión de que hay presencia de colinealidad

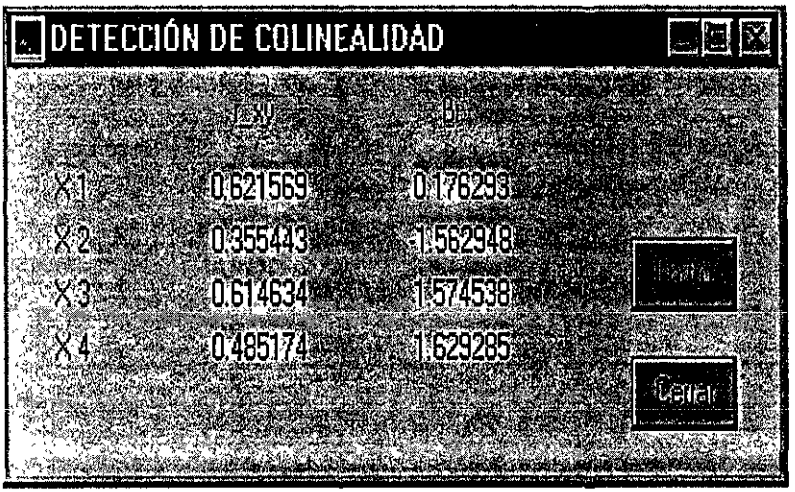


Figura 2.12: Detección de Colinealidad.

Nota: Si lo anterior no ocurre, no necesariamente indica que no hay colinealidad entre los regresores.

Capítulo 3

Análisis de los datos y las variables que intervienen en el Modelo de Regresión Lineal.

Introducción.

Los siguientes temas serán vistos en el presente capítulo:

1. El análisis de los residuales, obtenidos al estimar los parámetros $\hat{\beta}'s$ del modelo lineal

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

   mediante la graficación de los mismos.

2. La influencia de las observaciones (outliers o datos aberrantes), mediante la estadística conocida como Distancias de Cook.

3. Los criterios para seleccionar los mejores regresores que intervienen en el modelo lineal, para llegar a un modelo reducido.

3.1 Análisis de residuales.

El Análisis de los Residuales permite determinar observaciones aberrantes, violaciones a los supuestos del modelo como son que el vector de errores

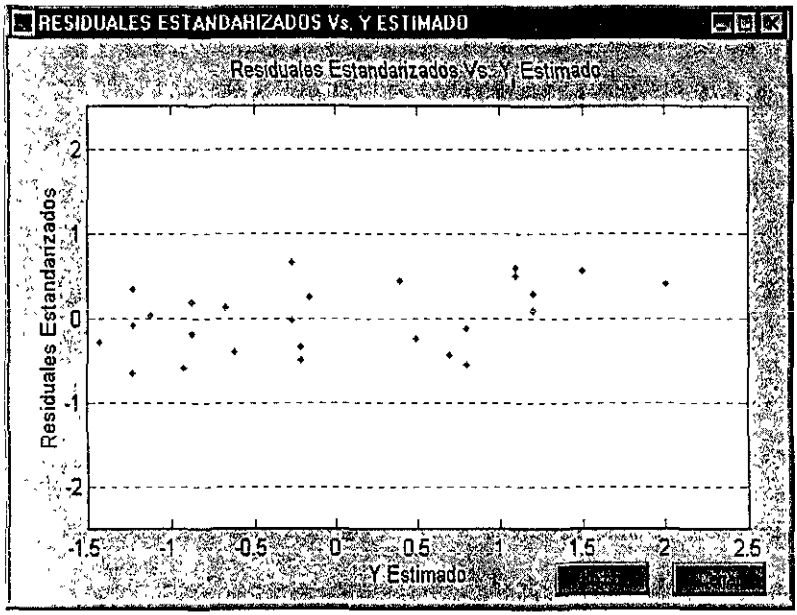


Figura 3.1: Residuales Estándarizados Vs. Y Estimado.

no observables son independientes, que tienen una distribución normal con media cero y varianza constante, también mediante este análisis es posible verificar si la relación que existe entre la variable de respuesta y y los regresores es lineal

Este análisis se basa en los residuales y los residuales estándarizados que son respectivamente

$$e_i = y_i - \hat{y}_i, \qquad rs_i = \frac{e_i}{s}, \quad i = 1, \ldots, n \quad y \quad s^2 = \frac{SCE}{n-p}$$

Al estándarizar los residuales estos tienen media cero y varianza aproximada a la unidad.

A continuación se indicarán las gráficas que comúnmente se acostumbran revisar para dicho análisis.

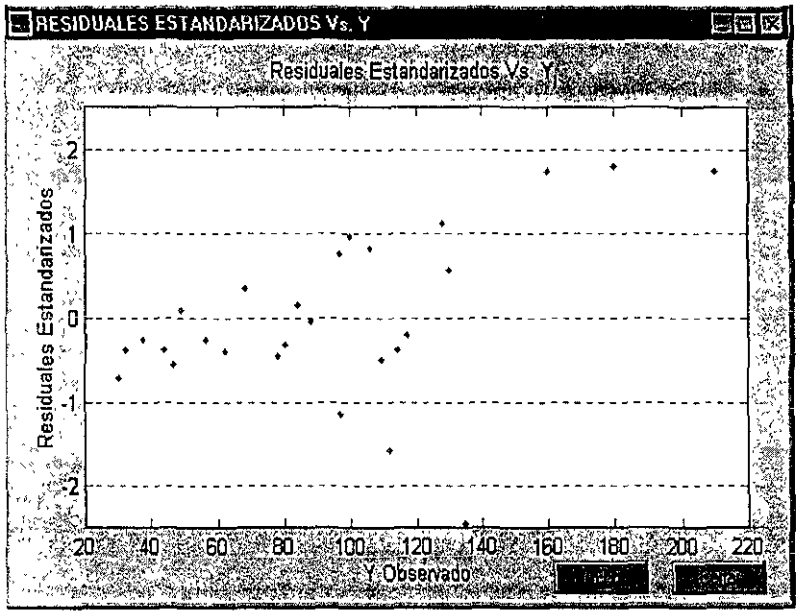


Figura 3.2: Residuales Estándarizados Vs. Y Observado.

3.1.1 Graficación de rs_i Vs. $\hat{y}_i$.

Al graficar los residuales estándarizados (rs_i) y los valores ajustados por el modelo $\left(\hat{y}_i\right)$, se verifica el supuesto de homocedasticidad (i. e. $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I)$), y se detecta la posible ocurrencia de datos aberrantes.

En general, cuando el modelo es correcto los residuales estándarizados se encuentran dispersos de manera aleatoria alrededor del cero, y toman valores dentro del intervalo $(-2, 2)$; esto tiene su justificación en el hecho de que la probabilidad que se encuentra dentro de este intervalo, para una variable aleatoria con distribución normal estándarizada es del 95%, por lo que al graficar los residuales estándarizados deben estar en este intervalo, si un residual estándarizado queda muy alejado de este intervalo uno estaría dispuesto a

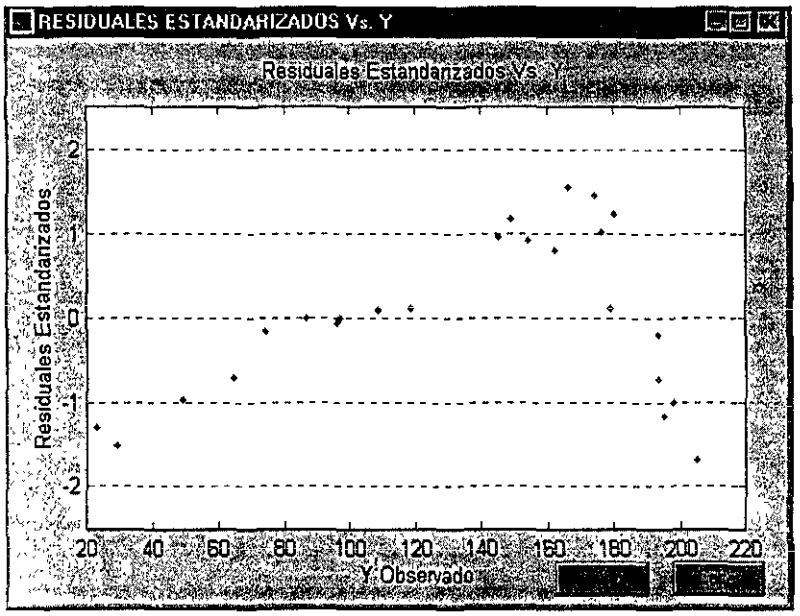


Figura 3.3. Residuales Estándarizados Vs. Y Observado

pensar que la respectiva observación no corresponde a la población y por ello resulta razonable considerarlo como un dato aberrante ó "outlier" en Inglés.

La existencia de un patrón sistemático de variación en los residuales puede considerarse como evidencia de una violación en uno de los supuestos del modelo, que es el de la varianza constante de los residuales u homocedasticidad (i. e., $\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I)$).

Cuando se cumple la suposición de varianza constante en los errores de un modelo lineal, se dice que los errores son homocedásticos (véase gráfica 3 1); cuando no se cumple tal suposición se dice que hay heterocedasticidad o que los errores son heterocedásticos

En la gráfica 3.2, se observa que la varianza de los errores no es constante y además que es una función creciente de la variable de respuesta; existen casos en los que la varianza de los errores se incrementa al decrecer los valores de la variable de respuesta y.

Cuando se tiene el caso de que la varianza de los errores no es constante, los resultados en las pruebas no serán de utilidad, pues aunque los estimadores obtenidos son insesgados, no son generalmente los mejores en cuanto a su precisión o varianza.

Algunas de las consecuencias que provoca el que la varianza en los errores no sea constante, son las siguientes:

1. Los estimadores podrán tener grandes desviaciones estándar.
2. Como consecuencia del punto anterior, los intervalos de confianza para los parámetros serán grandes
3. Las pruebas tendrán baja sensibilidad

Cuando una transformación es aplicada sobre los regresores (véase Chattierjee, [3]) para obtener, dentro del modelo, una varianza constante en los errores, se logra también en forma casual, el obtener buenas normalizaciones.

En la gráfica 3.3, se ilustra un caso donde el problema no es de varianza constante sino que la tendencia curva de los puntos indica más bien que el comportamiento de la variable y con respecto a sus regresores x_i, no es lineal.

En un modelo con un sólo regresor es recomendable graficar dicho regresor (x) vs. la variable de respuesta (y) para corroborar si la relación entre ambas es lineal.

Si se determina que un modelo lineal no es el adecuado para describir la relación que existe entre la variable de respuesta y y los regresores x_i, se sugiere el efectuar una transformación de variables que dependerá del tipo de gráfica resultante.

Otra gráfica que permite llegar a conclusiones análogas a las mencionadas en esta sección es la de los residuales estándarizados y los regresores.

En la siguiente sección se discute la gráfica que permite verificar el supuesto de normalidad.

3.1.2 Gráfica de probabilidad Normal de los residuales.

Uno de los supuestos básicos, aunados al planteamiento del modelo de regresión lineal, es que los errores tengan distribución normal con media $E(\underline{\varepsilon}) = \underline{0}$ y matriz de varianza -covarianza $Var(\underline{\varepsilon}) = \sigma^2 I$, esto es,

$$\underline{\varepsilon} \sim N\left(\underline{0}, \sigma^2 I\right)$$

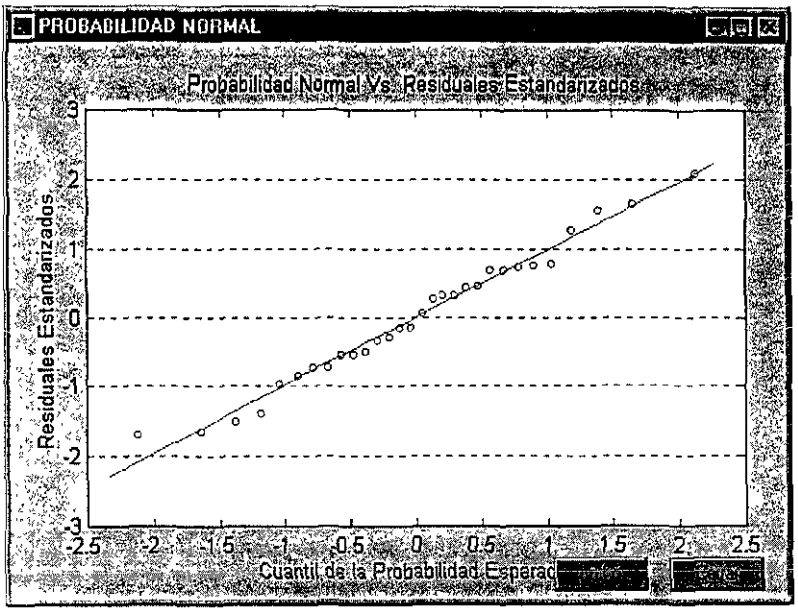


Figura 3.4: Verificación del supuesto de Normalidad en los Residuales.

ya que a partir de tal supuesto se generan las inferencias estadísticas que permiten la validación del modelo de regresión lineal ajustado, así como el cálculo de otras estadísticas importantes (vistas en el capítulo anterior). Para la verificación de dicho supuesto, a continuación se plantea una forma sencilla y rápida, basada en la graficación de los residuales estándarizados con su valor esperado. El procedimiento es el siguiente:

1. Se ordenan los residuales estándarizados en orden ascendente.

2. El valor esperado del $i-$ésimo residual para una muestra aleatoria de tamaño n (Neter [11], pag. 125) es:

$$s * z\left(\frac{i - .375}{n + .25}\right)$$

donde $s^2 = \dfrac{SCE}{n-p}$, es la estimación insesgada de la varianza σ^2 de los errores, y $z(A)$ es el cuantil de la distribución normal estándarizada que ha acumulado la probabilidad de A.

3 A continuación se grafican los residuales estándarizados con su valor esperado respectivo.

El supuesto de normalidad se cumple si los puntos resultantes en esta gráfica se encuentran dispersos sobre una linea recta, (como se muestra en la gráfica 3.4), en caso de no cumplirse ésto, se considerará que se está violando el supuesto de normalidad en los errores.

Nota:La decisión sobre tal supuesto queda siempre a consideración del analista.

La verificación de este supuesto es importante para el caso en el que se deseen realizar inferencias estadísticas; sin dicho supuesto no se podría concluir con cierta confiabilidad acerca de los resultados obtenidos

3.2 Prueba Durbin-Watson para autocorrelación.

Para los modelos de regresión lineal considerados se asume que los errores aleatorios ε_i son variables aleatorias no correlacionadas o variables aleatorias independientes, idénticamente distribuidas con distribución normal. A menudo, la suposición de errores no correlacionados o independientes no es apropiado, los errores están frecuentemente correlacionados positivamente sobre el tiempo. A los errores correlacionados sobre el tiempo se les llama autocorrelacionados

3.2.1 Problemas de autocorrelación.

La presencia de autocorrelación en un modelo de regresión, tiene un número de consecuencias importantes con el uso de las estimaciones mínimo cuadradas·

1. Los coeficientes de regresión estimados son insesgados, pero quizá no tengan la propiedad de varianza mínima y quizá sean ineficientes.

2 Los Cuadrados Medios de los Errores (CME) pueden bajo-estimar la varianza de los errores.

3. $se\left(\hat{\beta}_i\right)$ pueden bajoestimar la desviación estándar verdadera de los coeficientes de regresión estimada.

3.2.2 Durbin-Watson.

La prueba Durbin-Watson para autocorrelación asume el modelo de error autoregresivo.

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t1} + \cdots + \beta_{p-1} x_{t,p-1} + \varepsilon_t \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \end{aligned} \tag{3.1}$$

donde $|\rho| < 1$, y u_t es una variable aletoria $N(0, \sigma^2)$ independiente de los valores de los regresores ajustados.

La prueba consiste en determinar si el parámetro de autocorrelación ρ en (3.1) es cero. Note que si $\rho = 0$, $\varepsilon_t = u_t$. De aquí, los errores ε_t son entonces independientes porque u_t son independientes.

Las hipótesis usualmente consideradas son:

$$H_0 : \rho = 0 \qquad Vs. \qquad H_1 : \rho > 0 \tag{3.2}$$

El estadístico D para esta prueba es obtenido por el método de mínimos cuadrados al ajuste de la función de regresión, y su expresión es la siguiente

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

donde n es el número de casos.

Un procedimiento exacto para esta prueba no está disponible, pero Durbin y Watson han obtenido límites inferior y superior d_L y d_U tal que un valor de D fuera de estos límites conduce a una decisión definitiva. La regla de decisión para esta prueba (entre las hipótesis (3.2)) es:

Si $D > d_U$, inclinarse a aceptar H_0.

Si $D < d_L$, inclinarse a aceptar H_1.

Si $d_L \leq D \leq d_U$, la prueba es inconclusa

Valores pequeños de D conducen a la conclusión de que $\rho > 0$ porque el error adyacente de ε_t y ε_{t-1} tiende a ser de la misma magnitud cuando ellos son positivamente autocorrelacionados por lo tanto, la diferencia en los residuales, $e_t - e_{t-1}$, tendería a ser pequeña cuando $\rho > 0$, conduciendo a un numerador pequeño en D y por lo tanto a un estadístico de prueba D pequeño.

3.3 Influencia de las observaciones.

En ocasiones, hay observaciones que influyen de manera importante en el ajuste del modelo, por lo que es necesario identificarlas y determinar si deben o no permanecer en el modelo, para conocer la magnitud de tal influencia se utilizan las siguientes estadísticas: el leverage del $i-$ésimo caso, los residuales estudentizados y las distancias de Cook.

El leverage o potencia del $i-$ésimo caso u observación está dado por:

$$h_{ii} = \underline{x}_i \left(X^t X\right)^{-1} \underline{x}_i^t, \qquad i = 1, \ldots, n \tag{3.3}$$

donde $\underline{x}_i$ es el $i-$ésimo renglón de la matriz de datos X, entonces 3.3 indica el efecto de la $i-$ésima observación en la regresión, i. e., indica si los valores de $\underline{x}_i$ para la $i-$ésima observación son muy alejados porque puede mostrarse que h_{ii} es una medida de la distancia entre los valores de $\underline{x}_i$ para la i-ésima observación y la media de los valores de $\underline{x}$ (Neter [11], pag. 395), para todos los n casos. Esto es, un valor grande de h_{ii} indica que el $i-$ésimo caso es distante del centro de todas las observaciones

Un valor de h_{ii} es usualmente considerado grande si es más de dos veces el valor de la media de los leverages, denotado por $\overline{h} = \frac{p}{n}$.

Los valores del leverage mayores que $\frac{2p}{n}$ son considerados por esta regla como casos extremos.

Los residuales estudentizados se encuentran definidos por:

$$rt_i = \frac{e_i}{s\sqrt{1 - h_{ii}}} \qquad i = 1, \ldots n$$

si el modelo es correcto, los valores de esta estadística deben presentar una distribución t con $n - p$ grados de libertad, si esta cantidad es grande (mayor en valor absoluto que 2) para la $i-$ésima observación, entonces se investiga.

Nota: Los residuales estándarizados y los estudentizados son diferentes entre sí, pues tienen distinta distribución e interpretación.

Una estadística que mide el impacto combinado de la $i-$ésima observación sobre todos los coeficientes de regresión estimados son las Distancias de Cook D_i (Neter, [11], pag 403). Esta medida es derivada del concepto de una región de confianza para todos los p coeficientes de regresión β_k $(k = 0, 1, \ldots, p-1)$ simultáneamente Puede mostrarse (Montgomery, [8]) que el límite de esta región de confianza para el modelo lineal y errores con distribución normal, está dado por:

$$\frac{\left(\underline{\hat{\beta}} - \underline{\beta}\right)^t (X^t X) \left(\underline{\hat{\beta}} - \underline{\beta}\right)}{ps^2} = F^{1-\alpha}_{(p,n-p)}$$

Como se menciona en Montgomery [8] (pag. 163), "es conveniente considerar la ubicación del punto y la variable de respuesta al medir la influencia" de las observaciones, así, de la expresión anterior se originan las Distancias de Cook, otra estadística de suma importancia, además de las dos mencionadas previamente, es definida por:

$$D_i = \frac{\left(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}}\right)^t (X^t X) \left(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}}\right)}{ps^2} \qquad i = 1, \ldots, n$$

en donde se sugiere una medida del cuadrado de la distancia entre los estimadores Mínimo Cuadrados (Montgomery [8], pag. 163) tomando en cuenta todas las observaciones $\underline{\hat{\beta}}$, y los estimadores que se obtienen al no considerar la i-ésima observación $\underline{\hat{\beta}}_{(i)}$, e indica que para los casos en los que su valor es grande al ser eliminados, habrá cambios substanciales en el análisis.

Si ocurre que $F^5_{(p,n-p)} \simeq 1$, entonces para valores $D_i > 1$ se considera dicha observación como influyente en el ajuste del modelo.

Para calcular D_i es posible utilizar la siguiente expresión en forma alternativa (Montgomery, [8]):

$$D_i = \frac{r t_i^2 * h_{ii}}{p(1 - h_{ii})}, \qquad i = 1, \ldots, n$$

donde D_i es el producto del cuadrado del i-ésimo residual estudentizado y una función monótona de h_{ii}. Por lo tanto un valor grande de D_i, puede deberse

a un valor grande de e_i, de h_{ii}, o a ambos; es por ello que para efectuar un análisis completo es necesario considerar las estadísticas D_i, e_i y h_{ii} en forma conjunta, para cada caso.

Al decidir eliminar una observación por tener ésta una influencia importante con respecto a las demás, se modificarán los datos del modelo, y esto determinará una nueva estimación de los parámetros $\hat{\beta}$, para obtener su cálculo, se recurre a la actualización de la descomposición QR de la matriz X del modelo original, actualización que es discutida en la sección A.2.

Ejemplo 3.3.1 *En el Ejemplo 2.1.1, se tiene que para la observación 14:*

$$e_{14} = -16.9645, \qquad h_{14,14} = .06875, \qquad s = 10.29357$$

$$rt_{14} = \frac{-16.9645}{10.29357 * \sqrt{1 - .06875}} = -1\ 70781$$

$$D_{14} = \frac{(-1.70781)^2 * .06875}{4 * (1 - .06875)} = .05383$$

En este caso, la observación no influye mucho en el ajuste del modelo puesto que rt_{14} es relativamente un valor aceptable, así como $h_{14,14}$ que está muy por debajo del valor de $\frac{2p}{n} = .66667$ y D_{14} resulta ser un valor relativamente pequeño, por lo cual se concluye que ésta observación no es influyente en el cálculo de los parámetros del modelo.

En la siguiente sección se presentan diferentes criterios para llegar al mejor subconjunto de regresores que intervienen en el modelo.

3.4 Selección de variables.

Existen principalmente dos motivos por los que se efectúa una selección de variables sobre un modelo de regresión lineal. El primero es la sospecha de que existe colinealidad en el modelo y el segundo es el deseo del investigador de determinar el mínimo número de regresores que expliquen a la variable de respuesta y

Además de las dos razones mencionadas de una manera informal, Montgomery [8] sugiere que para tomar la decisión de efectuar una selección de las variables del modelo se consideren las siguientes preguntas.

1. ¿Es razonable la ecuación que representa el modelo? (i. e. ¿los regresores considerados en el modelo tienen sentido con el problema?).

2. ¿El modelo cumple los objetivos para los que se diseñó?.

3. ¿Las β_i's estimadas presentan valores razonables? (i. e. ¿sus signos y magnitudes son aceptables y sus errores estándar son relativamente pequeños?).

4. ¿Son satisfactorios los diagnósticos usuales para verificar lo adecuado del modelo? (Una prueba es la validación del modelo mediante una prueba de hipótesis).

Al eliminar o agregar una variable al modelo original realmente se está modificando el modelo. Existen varios métodos planteados para la reducción del modelo, que en su mayoría no conducen al mismo modelo reducido, por lo que no se puede concluir cuál de ellos es "el mejor", queda a consideración del analista el decidir cuál de los modelos reducidos es el que más se adecua a la realidad. A continuación se describen algunos de estos métodos:

3.4.1 Método Forward.

Como se mencionó anteriormente existen distintos métodos que permiten la obtención de un modelo reducido. Los pasos de este método son los siguientes.

1. Comienza suponiendo que no hay regresores en el modelo más que la intercepción (β_0)

2. Se calculan los coeficientes de correlación entre los regresores (x_i) y la variable de respuesta (y), dados por la siguiente expresión

$$r_{x_i y} = \sqrt{\frac{SCT - SCE(x_i)}{SCT}}, \qquad i = 1, \ldots, p-1 \qquad (3.4)$$

donde $SCE(x_i)$ es la Suma de Cuadrados de los Errores para el modelo reducido

$$y = \beta_0 + \beta_i x_i$$

El que tenga la mayor correlación en valor absoluto -supongamos x_{i_1}- es candidato a formar parte del modelo reducido. Se puede tomar la

misma decisión usando $|r_{x,y}|$ o $r^2_{x_i y}$. Se plantea el juego de hipótesis siguiente.

$$H_0 : \beta_{i_1} = 0 \qquad Vs. \qquad H_1 : \beta_{i_1} \neq 0$$

Se calcula el estadístico de prueba F_1 dado por:

$$F_1 = \frac{(n-2)\, r^2_{x_{i_1} y}}{1 - r^2_{x_{i_1} y}} \tag{3.5}$$

y se compara con el cuantil $F^{1-\alpha}_{(1,n-2)}$ al nivel de significancia α. Si $F_1 > F^{1-\alpha}_{(1,n-2)}$ se rechaza la hipótesis H_0 y el regresor elegido se queda en el modelo (i. e. x_{i_1} ahora es parte del modelo).

Nota: Como $F_1 = F\left(r_{x_{i_1} y}\right)$, *donde* $F\left(r_{x_{i_1} y}\right) = \dfrac{(n-2)\, r^2_{x_{i_1} y}}{1 - r^2_{x_{i_1} y}}$ *es creciente para* $r > 0$, se sigue que F_1 es máxima para x_{i_1} si $\left|r_{x_{i_1} y}\right|$ es máxima

3. Se calculan los coeficientes de correlación parcial entre los regresores restantes y la variable de respuesta, dado que el i_1-ésimo regresor está en el modelo, la expresión de estos se da a continuación:

$$r_{x_{i_2} y\, x_{i_1}} = \sqrt{\frac{SCE\left(x_{i_1}\right) - SCE\left(x_{i_1}, x_{i_2}\right)}{SCE\left(x_{i_1}\right)}} \tag{3.6}$$

El regresor con la mayor correlación en valor absoluto es candidato a entrar. Se plantea el siguiente juego de hipótesis:

$$H_0 : \beta_{i_2} = 0 \qquad Vs \qquad H_1 : \beta_{i_2} \neq 0$$

Se calcula el estadístico de prueba F_2, dado por:

$$F_2 = \frac{(n-3)\, r^2_{x_{i_2} y\, x_{i_1}}}{1 - r^2_{x_{i_2} y\, x_{i_1}}} \tag{3.7}$$

y se compara con el cuantil $F^{1-\alpha}_{(1,n-3)}$. Si $F_2 > F^{1-\alpha}_{(1,n-3)}$ se rechaza H_0 y el siguiente regresor se queda en el modelo.

4. Se realiza nuevamente el paso 2, ajustando las estadísticas dadas por (3.6) y (3.7) para este paso.

Este procedimiento termina cuando la estadística F_k no exceda al cuantil $F^{1-\alpha}_{(1,n-(k+1))}$, quedando en el modelo $k-1$ regresores.

En general, las correlaciones parciales al cuadrado están dadas por:

$$r^2_{x_{k+1}y\,x_1x_2\,\ldots\,x_k} = \frac{SCE\,(x_1,x_2,\ldots,x_k) - SCE\,(x_1,x_2,x_3,\ldots,x_{k+1})}{SCE\,(x_1,x_2,\ldots,x_k)} \tag{3.8}$$

donde $SCE\,(x_1,x_2,\ldots,x_k)$ es la Suma de Cuadrados de los Errores del modelo ajustado

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

y en forma análoga $SCE\,(x_1,x_2,x_3,\ldots,x_{k+1})$ es la Suma de Cuadrados de los Errores del modelo ajustado

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k+1} x_{k+1},$$

y la expresión dada por 3.8 se leería, la correlación parcial entre x_{k+1} y y, dado que los regresores $x_1, x_2, \ldots, x_k$ están en el modelo.

El estadístico F_k se calcularía de la siguiente manera:

$$F_{k+1} = \frac{(n-(k+2))\, r^2_{x_{k+1}y.x_1x_2\,\ldots\,x_k}}{1 - r^2_{x_{k+1}y\,x_1x_2\,\ldots\,x_k}}$$

y el cuantil con el que se tendría que comparar es: $F^{1-\alpha}_{(1,n-(k+2))}$, donde $k+1$ sólo nos indica el paso en el que estamos, que sería el $(k+2)$-ésimo paso.

3.4.2 Método Backward.

Este método funciona en forma inversa que el Método Forward, y es igual de efectivo para la difícil tarea de escoger "el mejor" modelo reducido. Los pasos de este método se describen a continuación

1. Se inicia el análisis considerando todos los regresores en el modelo
2. Se calculan las estadísticas F parciales para cada uno de los regresores considerados en el modelo, dadas por:

$$F = \frac{SCR(x_i/x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_{p-1})}{\frac{1}{n-p}SCE(x_1,\cdots x_{p-1})}$$
$$= \frac{SCE(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_{p-1}) - SCE(x_1,\cdots,x_{p-1})}{\frac{1}{n-p}SCE(x_1,\cdots x_{p-1})} \quad (3.9)$$

El valor más pequeño de estas estadísticas indica cual regresor es candidato a salir del modelo, comparándose con el cuantil $F^{1-\alpha}_{(1,n-p)}$. Si $F < F^{1-\alpha}_{(1,n-p)}$ el respectivo regresor sale del modelo.

3. Se calculan nuevamente las estadísticas F parciales dadas por (3.9), para cada uno de los regresores que quedaron en el modelo. El valor más pequeño es comparado nuevamente ahora con el cuantil $F^{1-\alpha}_{(1,n-(p-1))}$. Si $F < F^{1-\alpha}_{(1,n-(p-1))}$ el regresor sale del modelo.

Este procedimiento termina cuando la estadística F excede al cuantil $F^{1-\alpha}_{(1,n-(p-k+1))}$ quedando en el modelo $p-k$ regresores.

Nota: El criterio para adicionar o borrar un regresor es equivalente al usar el coeficiente de correlación parcial o la estadística F parcial, sólo es necesario conocer las equivalencias.

3.4.3 Estadística C_p de Mallows.

Otro método no menos importante que los anteriores es conocido como la estadística C_p de Mallows, la cual tiene su origen en el siguiente criterio:

Deducción de la Estadística C_p de Mallows.

Este criterio de selección de variables fue propuesto por Colin Mallows (Montgomery, [8]) y está relacionado directamente con el Error Cuadrático Medio (ECM) de un valor fijo. Esta estadística se utiliza para determinar un subconjunto de regresores a usar en un modelo de regresión lineal, que puedan explicar de manera confiable a la variable de respuesta y.

El planteamiento inicial consiste en lo siguiente:

Se tienen dos modelos de regresión; uno con $k-$regresores de estudio que se define como Modelo Lineal General (MLG) y otro con un subconjunto de estos $k-$regresores, digamos $p-$regresores, al que se definirá Modelo Lineal

Reducido (MLR), y la estadística C_p de Mallows se vale de la comparación de la Suma de Cuadrados de los Errores (SCE) de ambos modelos.

Para encontrar esta estadística se asume que el modelo completo (con $k-$regresores) es el modelo correcto, esto implica que será insesgado en sus estimaciones , esto es $E\left(\underline{y}\right) = \underline{y}^*$, mientras que el modelo reducido con el subconjunto de los regresores ($p-$regresores) será sesgado en su estimación, esto es, $E\left(\underline{y}_s\right) \neq \underline{y}^*$ donde el subíndice s indica que se está manejando el modelo con $p-$regresores.

Como este estadístico se deriva del ECM de un valor fijo, se empezará a trabajar directamente con éste; hay que ver entonces el ECM para las observaciones, dado por:

$$\begin{aligned} ECM\left(\hat{y}_{is}\right) &= E\left[\left(\hat{y}_{is} - E\left(y_i\right)\right)^2\right] \\ &= \left[E\left(y_i\right) - E\left(\hat{y}_{is}\right)\right]^2 + Var\left(\hat{y}_{is}\right) \\ &= \left[Sesgo\left(\hat{y}_{is}\right)\right]^2 + Var\left(\hat{y}_{is}\right) \end{aligned} \tag{3.10}$$

donde $\underset{\sim}{\hat{y}}_s = X_s \underset{\sim}{\hat{\beta}}_s$ es el modelo con $p-$regresores, (se toma $E\left(y_i\right)$ pues el valor verdadero es desconocido).

El Sesgo al Cuadrado Total para estos $p-$regresores se obtiene con la siguiente expresión:

$$SSB\left(p\right) = \sum_{i=1}^{n}\left[E\left(y_i\right) - E\left(\hat{y}_{is}\right)\right]^2 = \sum_{i=1}^{n}\left[Sesgo\left(\hat{y}_{is}\right)\right]^2 \tag{3.11}$$

A partir de las expresiones dadas en (3.10) y (3.11), el Error Cuadrático Medio Total Estandarizado se define de la siguiente manera:

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} ECM\left(\hat{y}_{is}\right) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\left(E\left(y_i\right) - E\left(\hat{y}_{is}\right)\right)^2 + Var\left(\hat{y}_{is}\right)\right] \\ &= \frac{1}{\sigma^2}\left(SSB\left(p\right) + \sum_{i=1}^{n} Var\left(\hat{y}_{is}\right)\right) \end{aligned} \tag{3.12}$$

La expresión dada en (3.12) es difícil de interpretar por lo que, si se realizan unos cuantos cálculos (véase Ryan, [13]), se pueden verificar las siguientes igualdades:

$$\sum_{i=1}^{n} Var\left(\hat{y}_{is}\right) = p\sigma^2 \quad \text{y} \quad E\left[SCE\left(p\right)\right] = SSB\left(p\right) + \left(n-p\right)\sigma^2 \qquad (3.13)$$

donde $SCE\left(p\right)$ es la Suma de Cuadrados de los Errores para el modelo con $p-$regresores, por lo que tenemos que (3.12), utilizando las igualdades de (3.13), se tranforma en:

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\left\{E\left[SCE\left(p\right)\right] - \left(n-p\right)\sigma^2 + p\sigma^2\right\} \\ &= \frac{1}{\sigma^2}\left\{E\left[SCE\left(p\right)\right] - n\sigma^2 + 2p\sigma^2\right\} \end{aligned}$$

Por lo tanto se tiene que el Error Cuadrático Medio Total Estandarizado está determinado por:

$$\Gamma_p = \frac{E\left[SCE\left(p\right)\right]}{\sigma^2} - n + 2p \qquad (3.14)$$

Como se desconocen los valores de $E\left[SCE\left(p\right)\right]$ y σ^2, se toman los estimadores de cada uno de ellos que serían $SCE\left(p\right)$ y s^2, donde s^2 es la estimación insesgada de σ^2 para el Modelo Lineal General (con los $k-$regresores). Un estimador entonces, para la expresión (3.14) es:

$$C_p = \hat{\Gamma}_p = \frac{SCE\left(p\right)}{s^2} - n + 2p \qquad (3.15)$$

Una vez obtenido el cálculo de esta estadística para los modelos reducidos, se presenta la incógnita de cómo saber cuál subconjunto de regresores es el mejor. Para esto Mallows propone el siguiente criterio.

Criterio de Selección.

El criterio propuesto por Mallows consiste en la construcción de la gráfica $C_p = p$ para cada ecuación de la regresión

Aquellos modelos con sesgo pequeño tendrán valores C_p cercanos a la linea $C_p = p$, mientras que los modelos con grandes sesgos tendrán distancias lejanas a la gráfica.

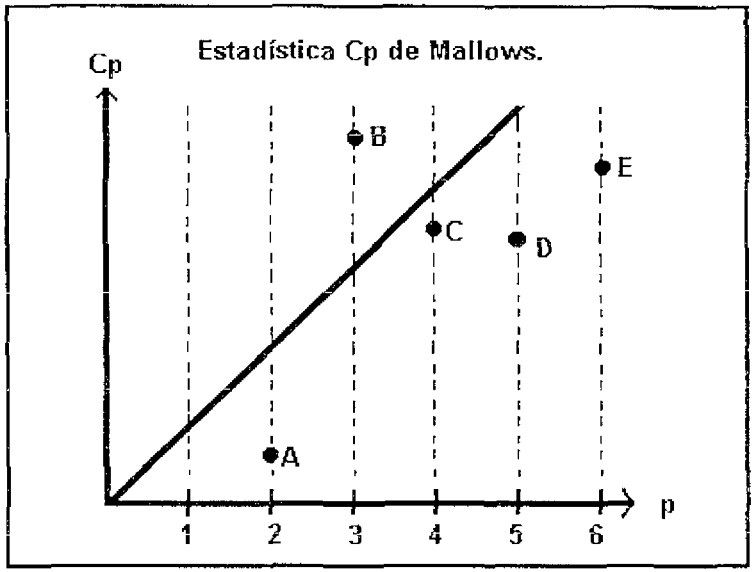


Figura 3.5: Análisis Gráfico de la Estadística C_p de Mallows

Si se considera la gráfica 3.5, el modelo dado en A es el que tiene menor número de regresores, pero debe considerarse que este modelo tiene un sesgo mucho mayor que el dado por C, ya que tiene el valor de C_p más cercano a la linea; este caso tiene más regresores en el modelo lineal reducido pero es el que presenta el menor Error Cuadrático Medio Total, esto es, tiene un menor sesgo, por lo que es preferible inclinarse a escoger el modelo dado en C antes que cualquir otro.

En el siguiente ejemplo se presenta una selección de variables aplicando los tres métodos previamente presentados.

Ejemplo 3.4.1 *Para el Ejemplo 2.1.1, se tienen los siguientes resultados, que se utilizarán para hacer una selección de variables:*

Sumas de Cuadrados.		
$SCE(x_1)$	=	2466.781
$SCE(x_2)$	=	4024.559
$SCE(x_3)$	=	4012.604
$SCE(x_1, x_2)$	=	2063.998
$SCE(x_1, x_3)$	=	2084.48
$SCE(x_2, x_3)$	=	3771.836
$SCE(x_1, x_2, x_3)$	=	2013.195
SCT	=	6145.217
$s^2 = 105.957624$		

- *Método Forward.*

  - *Paso 0: Se supone que la intercepción* (β_0), *es la única en el modelo, i. e., el modelo reducido es:*

$$y = \beta_0$$

  - *Paso 1: Se calculan las correlaciones entre regresores y variable dependiente, con la erpresión dada por (3.4) en la página 58.*

Correlaciones entre x_i *Vs. y*				
$r^2_{x_1y}$	=	$1 - \frac{2466.781}{6145.217}$	=	.59858
$r^2_{x_2y}$	=	$1 - \frac{4024.559}{6145.217}$	=	.34509
$r^2_{x_3y}$	=	$1 - \frac{4012.604}{6145.217}$	=	34703

La variable con mayor correlación es x_1, *la cual es candidata a formar parte del modelo reducido. Calculando el estadístico de*

prueba dado por (3.5) en la página 59, se tiene:

$$F_1^* = \frac{(23-2) * .59858}{1 - .59858} = 31.314$$

Comparando este valor con el estadístico de la distribución F con $\alpha = .05$, *se tiene* $F^{95}_{(1,21)} = 4.32 < F_1^*$, *por lo que se puede concluir que la variable* x_1 *se integra al modelo de regresión. Ahora el modelo reducido es:*

$$y = \beta_0 + \beta_1 x_1$$

- *Paso 2. Se calculan ahora las correlaciones parciales entre* x_2, x_3 *y la variable de respuesta, dado que* x_1 *está en el modelo. Se utiliza la expresión dada por (3.6) en la página 59:*

Correlaciones entre x_2 *y* x_3 *Vs. y dado* x_1.
$r^2_{x_2 y, x_1} = 1 - \frac{2063\ 998}{2466\ 781} = .16328$
$r^2_{x_3 y, x_1} = 1 - \frac{2084\ 48}{2466.781} = .15498$

Ahora x_2 *es candidato a formar parte del modelo reducido pues es quien tiene mayor correlación parcial, se calcula la estadística F dada por (3.7), página 59.*

$$F_2^* = \frac{(23-3) * (\ 16328)}{1 - \ 16328} = 3.9029$$

El cuantil para esta prueba es $F^{95}_{(1,20)} = 4.35 > F_2^*$, *lo cual indica que el regresor* x_2 *ya no es significativo en la explicación de la variable de respuesta, por lo que el "mejor" modelo reducido obtenido con este método sería:*

$$y = \beta_0 + \beta_1 x_1$$

- *Método Backward.*
  - *Paso 0: Inicia suponiendo que todos los regresores son parte del modelo, i.e., se tiene el modelo lineal general, que en este caso sería:*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- *Paso 1. Se calculan las estadísticas F parciales para cada uno de los regresores incluidos en el modelo con la expresión dada por (3.9) en la página 61:*

Estadísticas F-parciales para x_1, x_2 y x_3.
$F_{x_1} = \frac{3771.836-2013.195}{\frac{2013.195}{23-4}} = 16.598$
$F_{x_2} = \frac{2081.48-2013.195}{\frac{2013.195}{23-4}} = .67277$
$F_{x_3} = \frac{2063.998-2013.195}{\frac{2013.195}{23-4}} = .47947$

El regresor x_3 es el que tiene la estadística F-parcial más pequeña entonces, este regresor es candidato a salir del modelo. El cuantil para esta prueba es $F^{95}_{(1,19)} = 4.38 > F_{x_3}$, por lo que el regresor x_3 es sacado del modelo. Ahora el modelo de regresión lineal reducido es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- *Paso 2: Se recalculan las F-parciales para las variables que quedaron en el modelo, dado que x_3 está fuera del modelo:*

Estadísticas F-parciales para x_1 y x_2.
$F_{x_1} = \frac{4024.559-2063.998}{\frac{2063.998}{23-4}} = 18.998$
$F_{x_2} = \frac{2466.781-2063.998}{\frac{2063.998}{23-4}} = 3.9029$

El regresor x_2 es el que tiene la estadística F-parcial más pequeña, por lo que es posible candidato a salir del modelo. El cuantil para esta prueba es $F^{95}_{(1,20)} = 4.35 > F_{x_2}$, por lo que se toma la decisión de sacar a x_2 del modelo. Ahora el modelo de regresión lineal reducido es:

$$y = \beta_0 + \beta_1 x_1$$

– *Paso 3: Se calculan nuevamente las estadísticas F-parciales con únicamente x_1 en el modelo:*

Estadísticas F-parciales para x_1.
$F_{x_1} = \frac{6145.217 - 2466.781}{\frac{2466.781}{23-2}} = 31.315$

El estadístico para esta prueba es $F^{95}_{(1,21)} = 4.32 < F_{x_1}$, por lo que el regresor x_1 permanece en el modelo. Quedando como el "mejor" modelo reducido con este método:

$$y = \beta_0 + \beta_1 x_1$$

En este caso los dos métodos anteriores dejaron el mismo modelo lineal reducido, pero como se mencionó anteriormente, esto no siempre ocurre además, se puede llegar a la conclusión de que ninguna variable queda dentro del modelo de regresión, i. e., ninguno de los regresores es significativo en la explicación de la variable de respuesta.

- *Estadística C_p de Mallows.*

  Esta estadística se muestra en la expresión (3.15) en la página 63.

Variables	*Expresión*	*Cp*	*p*
x_1	$\frac{2466.781}{105.9576} + 2 * 2 - 23$	4.2808	2
x_2	$\frac{4024.559}{105.9576} + 2 * 2 - 23$	18.983	2
x_3	$\frac{4012.604}{105.9576} + 2 * 2 - 23$	18.87	2
x_1, x_2	$\frac{2063.998}{105.9576} + 2 * 3 - 23$	2.4795	3
x_1, x_3	$\frac{2084.48}{105.9576} + 2 * 3 - 23$	2.6728	3
x_2, x_3	$\frac{3771.836}{105.9576} + 2 * 3 - 23$	18.598	3
x_1, x_2, x_3	$\frac{2013.195}{105.9576} + 2 * 4 - 23$	4.0	4

Como dice Mallows en su análisis de ésta estadística, aquellos regresores en los cuales se cumpla que $C_p \approx p$, estarán proporcionando el "mejor" modelo lineal reducido De la tabla anterior se tiene que para el caso en que los tres regresores están incluidos $C_4 = 4.0$, pero este caso no interesa, cuando se consideran los regresores x_1, x_3 y la ordenada al origen (β_0) se tiene que $C_3 = 2.6728 \approx 3$, por lo que se puede concluir que el modelo que proporciona un menor Error Cuadrático Medio Total en las observaciones y_i es.

$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

En ocasiones la selección de variables es complicada dependiendo de las relaciones lineales que guarden los regresores (x_i) entre sí, esto es, la presencia de colinealidad en los regresores (véase sección 2.8), la cual produce sensibilidad en las estimaciones y como consecuencia la toma de decisiones erróneas, es por ello que se recomienda el hacer uso de varios métodos de selección de variables de tal manera que se obtenga uno de los "mejores" modelos lineales reducidos.

Capítulo 4

Análisis de Regresión Lineal vía la descomposición QR.

Introducción.

En este capítulo se presenta, desde el punto de vista del Análisis Numérico, cómo se obtiene la estimación de los parámetros del modelo lineal

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{4.1}$$

con los supuestos básicos de que $X \in \Re^{n\times p}$, $n > p$, X de rango máximo y el vector de errores $\underline{\varepsilon}$ con media $\underline{0}$ y matriz de varianza-covarianza $\sigma^2 I$.

La herramienta básica para el cálculo de las estimaciones Gauss-Markov, $\hat{\underline{\beta}}$ para $\underline{\beta}$ y s^2 para σ^2, es la descomposición QR de la matriz de datos X, ésta descomposición también permite calcular el vector de residuales $\underline{e}\left(=\underline{y} - X\hat{\underline{\beta}}\right)$ y su correspondiente matriz de varianzas y covarianzas $\sigma^2 (X^t X)^{-1}$ $\left[\acute{o}\ s^2 (X^t X)^{-1}\right]$, entre otras estadísticas. En base a estos cálculos se llevan a cabo otros más que permiten el análisis estadístico del modelo y de los datos, mediante pruebas de hipótesis y gráficas.

4.1 Cálculo de los estimadores de los parámetros $\beta's$ mediante las ecuaciones normales.

Para llegar a la obtención de la estimación $\hat{\underline{\beta}}$ de mínimos cuadrados para $\underline{\beta}$ en el modelo lineal dado por (4.1), es necesario resolver el siguiente problema:

$$\underset{\tilde{\underline{\beta}}}{Mín} \left\| \underline{y} - X\,\tilde{\underline{\beta}} \right\|_2^2$$

es decir, minimizar la suma de cuadrados de los errores. Ahora como:

$$\begin{aligned} \left\| \underline{y} - X\,\tilde{\underline{\beta}} \right\|_2^2 &= \left(\underline{y} - X\,\tilde{\underline{\beta}} \right)^t \left(\underline{y} - X\,\tilde{\underline{\beta}} \right) \\ &= \underline{y}^t\underline{y} - \underline{y}^t X\,\tilde{\underline{\beta}} - \tilde{\underline{\beta}}^t\, X^t\underline{y} + \tilde{\underline{\beta}}^t\, X^t X\,\tilde{\underline{\beta}} \\ &= \underline{y}^t\underline{y} - 2\underline{y}^t X\,\tilde{\underline{\beta}} + \tilde{\underline{\beta}}^t\, X^t X\,\tilde{\underline{\beta}} \end{aligned}$$

En la página 7 se calculó el punto crítico para esta expresión, llegando a que el mínimo se encuentra con la resolución del sistema lineal algebráico

$$X^t X\,\tilde{\underline{\beta}} = X^t\underline{y} \tag{4.2}$$

conocido como Ecuaciones Normales.

El cálculo de $\hat{\underline{\beta}}$ mediante la solución del sistema de las ecuaciones normales (4.2) con ayuda de una computadora no es recomendable, ya que implica el cálculo previo de la matriz X^tX, lo cual a parte de ser costoso, puede resultar inconveniente, pues bien X^tX (siendo $X \in \Re^{n\times p}$, $n > p$, de rango máximo) puede resultar numéricamente mal-condicionada o singular.

Ejemplo 4.1.1 *Sea $u < \varepsilon < \sqrt{u}$, donde u es la unidad de redondeo de la computadora en que se efectúan los cálculos.*

Se puede verificar que:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix}$$

tiene rango máximo (ya que $fl(1+\varepsilon) \neq 1$ *) Y que*

$$X^tX = \begin{bmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{bmatrix} \simeq \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

es singular en la computadora, pues $fl(1.+\varepsilon^2) = 1.$

Dado que la matriz X^tX es simétrica y definida positiva, se emplea un método alternativo para la resolución del sistema de ecuaciones (algebráicas) dado por (4.2), que debido a sus propiedades de estabilidad numérica y de economía en cuanto a requerimientos de memoria y de cantidad de cómputo resulta una buena herramienta, este es el método de Cholesky. Así, bajo el supuesto de que $X \in \Re^{n\times p}$, $n \geq p$, tiene rango máximo, se tiene que X^tX es simétrica y definida positiva. Por ello, el método ideal para resolver numéricamente las ecuaciones normales (4.2) es el método de Cholesky, el cual consiste en hallar la descomposición de Cholesky de X^tX, i. e.

$$X^tX = R^tR$$

siendo R una matriz triangular superior, de rango p, con diagonal positiva.

La resolución numérica de las ecuaciones normales (4.2), con base en esta descomposición, se reduce a resolver dos sistemas de ecuaciones triangulares:

$$R^t\underset{\sim}{z} = X^t\underset{\sim}{y} \qquad \text{y} \qquad R\underset{\sim}{\tilde{\beta}} = \underset{\sim}{z}.$$

El Análisis de Error Retrospectivo para el método de Cholesky, indica que:

$$\frac{\left\| \underset{\sim}{\overset{\wedge}{\beta}}^{*} - \underset{\sim}{\overset{\wedge}{\beta}} \right\|_2}{\left\| \underset{\sim}{\overset{\wedge}{\beta}} \right\|_2} \leq n^{\frac{3}{2}} \kappa_2\left(X^tX\right) u + O\left(u^2\right) \tag{4.3}$$

en donde $\underset{\sim}{\overset{\wedge}{\beta}}^{*}$ es la solución numérica y $\underset{\sim}{\overset{\wedge}{\beta}}$ es la solución exacta de las ecuaciones normales (4.2), u es la unidad de redondeo, y

$$\kappa_2\left(X^tX\right) = \left\|X^tX\right\|_2 \left\|\left(X^tX\right)^{-1}\right\|_2 \tag{4.4}$$

es el número de condición de X^tX.

Por lo que, la exactitud de la solución numérica de $\underset{\sim}{\overset{\wedge}{\beta}}^{*}$ de las ecuaciones normales (4.2) está fundamentalmente determinada por el número de condición $\kappa_2\left(X^tX\right)$.

4.1.1 Inversa generalizada de Moore-Penrose.

La Inversa Generalizada de Moore-Penrose X^+ de X, es definida como la solución respecto de $Y \in \Re^{p\times n}$ del sistema de ecuaciones matriciales:

1. $XYX = X$,
2. $YXY = Y$,
3. $(XY)' = XY$,
4. $(YX)' = YX$.

En efecto, se demuestra que este sistema matricial, llamado de Moore-Penrose, tiene una única solución.

Para el caso de matrices rectangulares, el concepto de número de condición se extiende como sigue:

Dada $X \in \Re^{n\times p}$ por el número de condición $\kappa_2(X)$ se entiende

$$\kappa_2(X) = \|X\|_2 \left\|X^+\right\|_2$$

Para la situación bajo discusión ($X \in \Re^{n\times p}$, $n > p$, $rgo(X) = p$), es directo verificar que

$$X^+ = \left(X^t X\right)^{-1} X^t.$$

4.1.2 Descomposición en valores singulares.

La Descomposición en Valores Singulares (DVS) de X, consiste en factorizarla como el siguiente producto de matrices:

$$X = U\Sigma V^t$$

en donde $U \in \Re^{n\times n}$ y $V \in \Re^{p\times p}$ son ortogonales, y

$$\Sigma = \begin{bmatrix} \Delta \\ 0 \end{bmatrix} \in \Re^{n\times p}$$

con

$$\begin{aligned} \Delta &= diag(s_1, s_2, \ldots, s_p), \\ s_1 &\geq s_2 \geq \cdot \cdot \geq s_p \geq 0. \end{aligned}$$

En Golub-Van Loan [4] se da una demostración de que esta DVS de $X \in \Re^{n\times p}$ (dada), siempre existe.

A partir de esta Descomposición es directo verificar que

$$\kappa_2\left(X^tX\right) = \left[\kappa_2\left(X\right)\right]^2. \tag{4.5}$$

De esta igualdad y la relación dada por (4.3) se sigue que en una microcomputadora AT de 32 bits con doble precisión (i. e. $\approx 10^{-16}$) si $\kappa_2\left(X\right) \simeq 10^8$ entonces carece de sentido resolver las ecuaciones normales (4.2), pues de (4.3) se tiene que

$$\frac{\left\|\underset{\sim}{\overset{\wedge}{\beta}}^{*} - \underset{\sim}{\overset{\wedge}{\beta}}\right\|_2}{\left\|\underset{\sim}{\overset{\wedge}{\beta}}\right\|_2} \simeq 1$$

Aún a pesar de haber aplicado el método ideal para matrices simétricas y definidas positivas, que es el método de Cholesky.

Luego, la resolución numérica de las ecuaciones normales (4.2) para calcular la estimación Gauss-Markov $\underset{\sim}{\overset{\wedge}{\beta}}$ para $\underset{\sim}{\beta}$ (estimación bajo la aplicación del criterio de mínimos cuadrados) tiene dos serios inconvenientes:

1. El cálculo previo de la matriz X^tX, el cual requiere $O\left(np^2\right)$ 'flops' (véase Golub-Van Loan, [4], para la definición de flop); además de que X^tX puede resultar numéricamente singular, aún cuando $X \in \Re^{n\times p}$ sea numéricamente de rango p.

2. Si $\kappa_2\left(X\right) \approx \frac{1}{\sqrt{u}}$, siendo u la unidad de redondeo de la Aritmética de Punto Flotante usada, entonces puede carecer de sentido la resolución numérica de las ecuaciones normales (4.2), aún cuando éstas sean resueltas por el método de Cholesky.

Una alternativa a la resolución del problema de Mínimos Cuadrados vía la resolución numérica de sus ecuaciones normales (4.2), está basada en la factorización de la matriz X mediante el empleo de matrices elementales de eliminación ortogonales. Esta alternativa, basada en la llamada descomposición QR de X, aún cuando es un poco más cara (en cuanto a requerimientos de memoria y cantidad de cómputo numérico), no requiere del cálculo de X^tX, y bajo ciertas condiciones adicionales (muy razonables desde el punto de vista práctico), su sensibilidad numérica queda determinada por $\kappa_2\left(X\right)$ en vez de $\kappa_2\left(X^tX\right)$.

4.2 Descomposición QR de la matriz X.

Para iniciar este tema es necesario tratar previamente el tema de matrices ortogonales, ya que a partir de cierto tipo de estas matrices -conocidas como de Householder-, la matriz X se reduce a una matriz triangular, lo cual permite determinar los parámetros β de una manera relativamente fácil.

Una matriz ortogonal(ortonormal) es una matriz Q de $n \times n$ que satisface que $Q'Q = QQ' = I$, donde I es la identidad de $n \times n$, este tipo de matrices es importante porque preserva la norma euclidiana (o norma 2), ya que:

$$\|Q\underline{z}\|_2 = \sqrt{(Q\underline{z})'(Q\underline{z})} = \sqrt{\underline{z}'Q'Q\underline{z}} = \sqrt{\underline{z}'\underline{z}} = \|\underline{z}\|_2$$

A continuación se revisan brevemente las reflexiones de Householder, cuya importancia radica en que facilitan la reducción de una matriz a una forma triangular superior.

4.2.1 Reflexiones de Householder.

Una reflexión de Householder es una matriz H de $n \times n$ que se define de la siguiente manera:

$$H = I - 2\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} = I - 2\frac{\underline{u}\underline{u}'}{\|\underline{u}\|_2^2} \tag{4.6}$$

donde $\underline{u}$ es un vector de $n \times 1$ distinto del vector cero Estas matrices son simétricas y ortogonales. Esta última característica se verifica a continuación:

$$\begin{aligned}
H'H &= HH \\
&= \left(I - 2\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}}\right)\left(I - 2\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}}\right) \\
&= I - 4\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} + 4\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}}\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} \\
&= I - 4\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} + 4\frac{\underline{u}(\underline{u}'\underline{u})\underline{u}'}{(\underline{u}'\underline{u})^2} \\
&= I - 4\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} + 4\frac{\underline{u}\underline{u}'}{\underline{u}'\underline{u}} \\
&= I
\end{aligned}$$

En particular se tiene que:

$$H = H^t = H^{-1}$$

La construcción de estas matrices permite que todas las componentes de un vector -no nulo- al que se le aplique esta reflexión, sean cero a excepción de la primera. Esto se explica claramente en el siguiente:

Lema 4.2.1 *Dado* $\underline{x} \in \Re^n$, $\underline{x} \neq \underline{0}$ *existe* H *de Householder definida por* $\underline{u} = \underline{x} + \alpha \hat{e}_1$, $\alpha = sign(x_1) \|\underline{x}\|_2$ *tal que* $H\underline{x} = -\alpha \hat{e}_1$.

Esto es:

$$H\underline{x} = \begin{bmatrix} -\alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

y debido a que las transformaciones ortogonales preservan la norma 2, se tiene que $|\alpha| = \|\underline{x}\|_2$.

Nótese que el signo de α depende directamente del signo de x_1 en el vector $\underline{x}$, esto se hace por razones de estabilidad numérica (i. e. se trata de evitar cancelación numérica).

Las condiciones para la existencia de una matriz ortogonal Q y una matriz R triangular superior, tal que $X = QR$, y la forma de cómo obtener esta descomposición se presentan en el siguiente:

Teorema 4.2.1 *Dada* $X \in \Re^{n \times p}$, $n \geq p$, *de rango* p, *existen* $Q \in \Re^{n \times n}$ *ortogonal y* $R \in \Re^{p \times p}$ *triangular superior tal que*

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

Demostración : *Escríbase la matriz* X *como sigue:*

$$X = \left[\underline{x}_1 | \underline{x}_2 | \ \cdots | \underline{x}_p\right].$$

Como X *tiene rango* p, *se tiene que* $\underline{x}_1 \neq \underline{0}$. *Luego, por el lema 4.2.1, existe* H_1 *de Householder tal que*

$$
\begin{aligned}
H_1X &= \left[H_1\underline{x}_1|H_1\underline{x}_2|\cdots|H_1\underline{x}_p\right] \\
H_1X &= \begin{bmatrix} -\alpha_1 & x_{12}^{(1)} & x_{13}^{(1)} & \cdots & x_{1p}^{(1)} \\ 0 & x_{22}^{(1)} & x_{23}^{(1)} & \cdots & x_{2p}^{(1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & x_{n2}^{(1)} & x_{n3}^{(1)} & \cdots & x_{np}^{(1)} \end{bmatrix} \\
&= \begin{bmatrix} -\alpha_1 & x_{12}^{(1)} & x_{13}^{(1)} & \cdots & x_{1p}^{(1)} \\ & & & & \\ \underline{0} & \underline{x}_2^{(1)} & \underline{x}_3^{(1)} & \cdots & \underline{x}_p^{(1)} \end{bmatrix} \\
&= \begin{bmatrix} -\alpha_1 & \underline{x}_{1*}^{(1)} \\ & \\ \underline{0} & X_{22}^{(1)} \end{bmatrix}.
\end{aligned}
$$

Donde el superíndice k en $x_{ij}^{(k)}$ indica el número de operaciones que se han realizado sobre dicho factor.

Como X tiene rango p, se tiene que $\underline{x}_2^{(1)} \neq \underline{0}$. Luego, por el lema existe $\widetilde{H}_2$ tal que

$$
\begin{aligned}
\widetilde{H}_2\, X_{22}^{(1)} &= \left[\widetilde{H}_2\,\underline{x}_2^{(1)}|\,\widetilde{H}_2\,\underline{x}_3^{(1)}|\cdots|\,\widetilde{H}_2\,\underline{x}_p^{(1)}\right] \\
&= \begin{bmatrix} -\alpha_2 & x_{23}^{(2)} & x_{24}^{(2)} & \cdots & x_{2p}^{(2)} \\ 0 & & & & \\ \vdots & & X_{22}^{(2)} & & \\ 0 & & & & \end{bmatrix}
\end{aligned}
$$

Tomando la matriz H_2 de la siguiente manera:

$$
H_2 = \begin{bmatrix} 1 & \underline{0}^t \\ & \\ \underline{0}^t & \widetilde{H}_2 \end{bmatrix}
$$

se sigue que

$$
H_2H_1X = \begin{bmatrix} -\alpha_1 & x_{12}^{(1)} & x_{13}^{(1)} & \cdots & x_{1p}^{(1)} \\ 0 & -\alpha_2 & x_{23}^{(2)} & \cdots & x_{2p}^{(2)} \\ 0 & 0 & x_{33}^{(2)} & \cdots & x_{3p}^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & x_{n3}^{(2)} & \cdots & x_{3p}^{(2)} \end{bmatrix}
$$

Así, después de aplicar p pasos de eliminación de Householder a la matriz X se tiene que

$$H_p H_{p-1} \cdots H_2 H_1 X = \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

Sea $Q^{-1} = H_p H_{p-1} \cdots H_2 H_1$, Q^{-1} *es ortogonal por ser producto de matrices* H_k *ortogonales.*

Con esto, se sigue que

$$Q^{-1}X = \begin{bmatrix} R \\ 0 \end{bmatrix} \qquad \textit{por lo que} \qquad X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

y que

$$\begin{aligned} Q &= (H_p H_{p-1} \cdots H_2 H_1)^{-1} \\ &= H_1^{-1} H_2^{-1} \cdots H_p^{-1} \\ &= H_1 H_2 \cdots H_p \end{aligned} \qquad \blacksquare$$

Esta es la llamada Descomposición QR de la matriz X.

Si se considera la siguiente partición de la matriz $Q = [Q_x, Q_\perp]$ donde Q_x tiene p columnas, entonces X se puede simplificar de la siguiente forma:

$$X = [Q_x, Q_\perp] \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_x R$$

esto es, se tiene que

$$X = Q_x R. \tag{4.7}$$

Ahora, sustituyendo esta descomposición en las ecuaciones normales:

$$X^t X \, \underset{\sim}{\tilde{\beta}} = X^t \underline{y}$$

y simplificando la expresión, se llega a que:

$$R \, \underset{\sim}{\tilde{\beta}} = Q_x^t \underline{y} \tag{4.8}$$

En resumen, la resolución del problema de mínimos cuadrados

$$\underset{\underset{\sim}{\tilde{\beta}}}{Mín} \frac{1}{2} \left\| \underline{y} - X \, \underset{\sim}{\tilde{\beta}} \right\|_2^2$$

en términos de la descomposición QR de la matriz X, se reduce a efectuar los siguientes dos pasos:

1. Calcular el producto de

$$Q_x^t \underline{y} \tag{4.9}$$

2. Resolver el sistema triangular superior

$$R\,\underset{\sim}{\beta} = Q_x^t \underline{y}$$

por el conocido método de sustitución sucesiva hacia atrás

Si X es de rango completo, R es no singular. Por ello es posible hablar de R^{-1} que a partir de (4.8) se tiene que:

$$\underset{\wedge}{\beta} = R^{-1} Q_x^t \underline{y}$$

Esto implica que:

$$X^+ = R^{-1} Q_x^t = \left[R^{-1}, 0\right] \begin{bmatrix} Q_x^t \\ Q_\perp^t \end{bmatrix} = \left[R^{-1}, 0\right] Q^t$$

donde X^+ denota a la matriz inversa generalizada de Moore-Penrose.

Por lo que es directo verificar que

$$\kappa_2(R) = \kappa_2(X)$$

En consecuencia, bajo el supuesto de la hipótesis del ángulo agudo (véase sección 4.5.2), la cota (a priori) del error relativo está esencialmente dada por:

$$\frac{\left\| \underset{\wedge}{\beta}^* - \underset{\wedge}{\beta} \right\|_2}{\left\| \underset{\wedge}{\beta} \right\|_2} \leq \kappa_2(R) \frac{\|\delta R\|_2}{\|R\|_2} \simeq \kappa_2(X) \frac{\|\delta X\|_2}{\|X\|_2}.$$

Por lo tanto, si X es de rango máximo y no es muy mal comportada entonces, bajo la hipótesis del ángulo agudo, la descomposición QR es la mejor opción para la resolución numérica del problema de mínimos cuadrados lineal clásico.

A manera de resumen, se puede decir que al efectuar la descomposición QR de la matriz X se obtienen las siguientes características numéricas importantes:

1. No es necesario construir las ecuaciones normales (4.2) relacionadas al sistema.

2. Al ser Q ortogonal la descomposición QR se efectúa de manera completamente estable

3. La sensibilidad numérica del sistema (4.8) es la de X, y no la de $X'X$ como ocurre con las ecuaciones normales.

Es por ello que el uso de esta descomposición se considera la forma más viable de evitar problemas de sensibilidad numérica.

4.3 Otros usos de la descomposición QR de la matriz X.

El paquete de programas MATLAB, cuenta con funciones que calculan la descomposición QR de la matriz X, y en base a un sencillo procedimiento se estiman los parámetros $\hat{\underline{\beta}}$ del modelo lineal, los valores de predicción $\hat{\underline{y}}$ de las observaciones y el vector de residuales que se denota con el vector $\underline{e}$.

A partir de los residuales es posible obtener:

1. la estimación de s^2 para la varianza, $s^2 = \dfrac{\underline{e}'\underline{e}}{n-p} = \dfrac{SCE}{n-p}$

2. la estadística F, $F = \dfrac{\frac{1}{p-1}SCR}{\frac{1}{n-p}SCE}$

3. el coeficiente de determinación R^2, $R^2 = \dfrac{SCR}{SCT}$

4. el coeficiente de determinación ajustado R_a^2, $R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\dfrac{SCE}{SCT}$

5. los residuales estándarizados, $rs_i = \dfrac{e_i}{s}$

6. los residuales estudentizados, $rt_i = \dfrac{e_i}{s\sqrt{1-h_{ii}}}$, donde $h_{ii} = \left(X\left(R'R\right)^{-1}X'\right)_{ii}$

7. la estadística C_p de Mallows, $C_p = \dfrac{SCE\,(p)}{s^2} + 2p - n$

8. graficar los residuales estándarizados Vs. las $\hat{y}_i$ para determinar si hay observaciones aberrantes

9. graficar los residuales estándarizados Vs. las $\hat{y}_i$ para determinar si el modelo cumple con ciertas suposiciones como es la homocedasticidad en los errores

10. graficar los residuales estándarizados Vs. cuantil esperado de la distribución Normal, para verificar si se cumple con el supuesto de normalidad en los errores.

Las funciones de MATLAB calculan también la matriz $(R^tR)^{-1}$ que es exactamente (salvo por errores de redondeo) la matriz $(X^tX)^{-1}$ debido a que de la descomposición QR de X

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

se tiene que

$$\begin{aligned} X^tX &= \begin{bmatrix} R^t & 0^t \end{bmatrix} Q^tQ \begin{bmatrix} R \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} R^t & 0^t \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} \\ &= R^tR, \end{aligned}$$

por lo tanto

$$(X^tX)^{-1} = (R^tR)^{-1}$$

A partir de la matriz $(R^tR)^{-1}$ (sus elementos son r_{ij} $i, j = 1, \ldots, p$) es posible obtener las siguientes estadísticas:

1. la desviación estándar de cada $\hat{\beta}_i$, $se\left(\hat{\beta}_i\right) = s\sqrt{r_{ii}}$

2. el coeficiente de inflación de la varianza para cada $\hat{\beta}_i$, determinado por r_{ii}, (siempre y cuando la matriz X haya sido previamente escalada)

3. la potencia o leverage de cada observación, $h_{ii} = \left(X\left(R^tR\right)^{-1}X^t\right)_{ii}$

4. las distancias de Cook, $D_i = \dfrac{rt_i^2 * h_{ii}}{p(1-h_{ii})}$.

Es por estas razones que el Sistema ANA_RELI.M VER. 2.0 detallado en el siguiente capítulo, se basa en la descomposición QR de la matriz X.

4.4 Proceso de obtención de la media y la varianza muestrales.

Los procedimientos que generalmente se plantean en los libros de texto de estadística para la obtención de la media y sobre todo la varianza, al ser implantados en una computadora generalmente producen resultados erróneos. Esto se debe a que si todos los datos se encuentran cercanos a la media, entonces es muy probable que ocurra una cancelación numérica. El algoritmo que proponen algunos libros de texto es el siguiente:

$$\overline{x} = \frac{\sum_{i=1}^{m} x_i}{m} \qquad \text{y} \qquad s^2 = \frac{\sum_{i=1}^{m} x_i^2 - m\overline{x}^2}{m-1}$$

Se han desarrollado varios algoritmos para evitar la cancelación numérica, uno de ellos es conocido como el de dos pasos, el cual es numéricamente estable, pero muy costoso; debido a que para el cálculo de la varianza es necesario leer el archivo de datos en dos ocasiones, a continuación se presenta este algoritmo.

$$s^2 = \frac{\sum_{i=1}^{m} (x_i - \overline{x})^2}{m-1}$$

Una alternativa es utilizar algoritmos que sean menos costosos y en los que se pueda conocer una cota del error en el resultado; uno de ellos, es el desarrollado por West-Hanson [5] y que a continuación se describe.

Algoritmo de West-Hanson para actualizar la media y la varianza muestrales.

$$\begin{aligned} M_1 &\leftarrow x_1, \\ V_1 &\leftarrow 0.e0; \\ \text{Para } i &= 2,\ldots,m, \text{ haz} \end{aligned}$$

$$
\begin{aligned}
M_i &\leftarrow M_{i-1} + \frac{(x_i - M_{i-1})}{i} \\
V_i &\leftarrow V_{i-1} + (i-1) * (x_i - M_{i-1}) * \left(\frac{x_i - M_{i-1}}{i} \right) \\
\overline{x} &\leftarrow M_n \\
s^2 &\leftarrow \frac{1}{m-1} V_m
\end{aligned}
$$

A continuación se presenta un ejemplo en el que se aplican los tres algoritmos mencionados. Si se desea verificar el análisis de error correspondiente a cada uno de ellos, consulte el Apéndice C de la referencia [12].

Ejemplo 4.4.1 *En una Aritmética de Punto Flotante de 3 decimales, considere las siguientes observaciones* $x_1 = 19$, $x_2 = 20$ *y* $x_3 = 21$; *al aplicar los algoritmos anteriores se obtienen los resultados siguientes:*

Si se calcula la media como $\overline{x} = \dfrac{\sum_{i=1}^{m} x_i}{m}$, *para el ejemplo, se tiene que* $fl\left(\frac{19+20+21}{3}\right) = fl\,(20) = 20$.

1. *Con el algoritmo de dos pasos se tiene lo siguiente:*

$$
\begin{aligned}
s_0^2 &= (19-20)^2 = 1 \\
s_1^2 &= 1 + (20-20)^2 = 1 + 0 = 1 \\
s_2^2 &= 1 + (21-20)^2 = 1 + 1 = 2 \\
s^2 &= \tfrac{2}{2} = 1
\end{aligned}
$$

2. *Con el algoritmo de los libros de texto de estadística se tendría lo siguiente:*

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{m} x_i^2 - m\overline{x}^2}{m-1} \\
fl\left(\sum_{i=1}^{3} x_i^2\right) &= fl\,(361 + 400 + 441) = fl\,(1202) = .120 \times 10^4 \\
y\ \overline{x}^2 &= 400
\end{aligned}
$$

por lo que

$$\tilde{s}^2 = fl\left(\frac{1200 - 3 * 400}{2}\right) = fl\left(\frac{1200 - 1200}{2}\right) = 0$$

3. *Con el algoritmo de West-Hanson*

$$\begin{aligned} M_1 &= 19, V_1 = 0 \\ M_2 &= 19 + \frac{20-19}{2} = 19.5, \quad V_2 = 0 + (2-1) * (20-19) * \left(\frac{20-19}{2}\right) = .5 \\ M_3 &= 19.5 + \frac{21-19.5}{3} = 20.0, \quad V_3 = .5 + (3-1) * (21-19.5) * \left(\frac{21-19.5}{3}\right) = 2.0 \end{aligned}$$

Por lo tanto

$$\overline{x} = 20.0 \qquad s^2 = \frac{2}{2} = 1.0$$

En la siguiente tabla se comparan las cotas de error entre el algoritmo de West (que utiliza el algoritmo que desarrolló Hanson para calcular la media), el de dos pasos y el que se presenta en los libros de texto de estadística.

Algoritmo	Cota de Error con Dígito de Guardia
Dos Pasos	$(m+4)\,u$
Textos de Estadística	$3m\kappa^2(s)\,u + 2u$
Hanson	$\left(2 + \sqrt{m} + \frac{8}{3}\sqrt{2m}\right)\kappa(s)\,u + (m+4)\,u$

Donde:

$\underline{x} = (x_1, x_2, \ldots, x_m)^t$

m = número de datos.

u = es la unidad de redondeo de la computadora en que se efectúan los cálculos.

$$\kappa(s) = \frac{1}{\sqrt{m-1}} \frac{\|\underline{x}\|}{s}, \text{ es el número de condición de la desviación estándar.}$$

con $\|\underline{x}\|$ = norma del vector $\underline{x}$, y

$$s = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \bar{x})^2}{m-1}}$$

4.5 Un poco de Análisis Numérico en la Regresión Lineal.

En las secciones anteriores se trató la importancia de la descomposición QR de la matriz X (método netamente numérico) en la solución al problema de minimizar la Suma de Cuadrados de los Errores, esto es

$$\underset{\tilde{\underline{\beta}}}{Min} \left\| \underline{y} - X \tilde{\underline{\beta}} \right\|_2^2 \tag{4.10}$$

que alcanza su punto crítico con la resolución del sistema lineal algebráico

$$X^t X \hat{\underline{\beta}} = X^t \underline{y} \tag{4.11}$$

conocido como sistema de Ecuaciones Normales, y que utilizando la descomposición QR, lleva a la solución

$$\hat{\underline{\beta}} = R^{-1} Q_x^t \underline{y}.$$

El cálculo de la estimación del vector $\hat{\underline{\beta}}$, afecta en forma directa en los resultados de todas las demás estadísticas a calcular, necesarias para llevar a cabo el análisis de regresión, pues éstas dependen de la veracidad en la estimación dada por $\hat{\underline{\beta}}$ (véase sección 4.3), por esta razón siempre se buscan aquellos métodos numéricos que brindan la confiabilidad y estabilidad numérica de los resultados obtenidos, tales como la descomposición QR de la

matriz X (véase sección 4.2) y los algoritmos para el cálculo de la media y la varianza muestrales (véase sección 4.4).

La solución al sistema de Ecuaciones Normales, dado por (4.11), tiene dos posibilidades en su solución:

1. Si $rgo(X) = p$ entonces X^tX es no-singular y (4.10) tiene una única solución $\underset{\sim}{\hat{\beta}}$ dada por

$$\underset{\sim}{\hat{\beta}} = \left(X^tX\right)^{-1} X^t\underset{\sim}{y} = R^{-1}Q_x^t\underset{\sim}{y}.$$

2. Si $rgo(X) < p$ entonces (4.10) tiene una infinidad de soluciones

$$\underset{\sim}{\hat{\beta}} = \underset{\sim}{\hat{\beta}}_0 + N(X).$$

Una solución generalizada al problema anterior para su segunda posibilidad, está planteada en la siguiente:

Definición 4.5.1 *Se dirá que* $\underset{\sim}{\hat{\beta}}^+$ *es una solución generalizada de (4.10) si*

1. $\left\|\underset{\sim}{y} - X\,\underset{\sim}{\hat{\beta}}^+\right\|_2 \leq \left\|\underset{\sim}{y} - X\,\underset{\sim}{\tilde{\beta}}\right\|_2$, *para todo* $\underset{\sim}{\tilde{\beta}} \in \Re^p$.

2. *si* $\underset{\sim}{\hat{\beta}}^*$ *es tal que*

$$\left\|\underset{\sim}{y} - X\,\underset{\sim}{\hat{\beta}}^*\right\|_2 \leq \left\|\underset{\sim}{y} - X\,\underset{\sim}{\tilde{\beta}}\right\|_2. \text{ para todo } \underset{\sim}{\tilde{\beta}} \in \Re^p$$

entonces

$$\left\|\underset{\sim}{\hat{\beta}}^*\right\|_2 \geq \left\|\underset{\sim}{\hat{\beta}}^+\right\|_2.$$

Se puede demostrar que la solución generalizada $\underset{\sim}{\hat{\beta}}^+$ para el problema (4.10), está determinada por:

$$\underset{\sim}{\hat{\beta}}^+ = X^+\underset{\sim}{y}$$

donde

$$X^+ = V\Sigma^+ U^t, \quad \Sigma^+ = diag\left(s_1^+, s_2^+, \ldots, s_p^+\right), \quad s_i^+ = \begin{cases} s_i^{-1}, & \text{si } s_i \neq 0 \\ 0, & \text{si } s_i = 0 \end{cases}$$

es la Pseudoinversa (inversa generalizada) de Moore-Penrose para X (definida en la sección 4.1.1), asegurándose con esto la existencia de la solución generalizada al problema de mínimos cuadrados dado por (4.10).

Una vez encontrada la solución a (4.10), se requiere de un indicador que diga la precisión o exactitud con la cual es calculado el vector de parámetros $\hat{\underline{\beta}}^+$ y qué tanto dista esta solución de la obtenida en aritmética real. En la siguiente sección se da una cota para el cálculo de dicho error.

4.5.1 Análisis de sensibilidad del problema de mínimos cuadrados lineales.

El objetivo de este análisis es ver qué tanto se altera $\hat{\underline{\beta}}^+ = X^+\underline{y}$ bajo perturbaciones en X y $\underline{y}$ en el modelo lineal

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

La razón de ser de este estudio se debe a que el cálculo numérico de $\hat{\underline{\beta}}^+ = X^+\underline{y}$; i. e., la solución del problema de Mínimos Cuadrados:

$$\underset{\tilde{\underline{\beta}}}{Min} \left\| \underline{y} - X\tilde{\underline{\beta}} \right\|_2^2,$$

usando métodos numéricamente estables, nos da una solución aproximada $\hat{\underline{\beta}}_a^+$ (que se espera sea muy próxima a $\hat{\underline{\beta}}^+$), que es la solución exacta del problema de Mínimos Cuadrados perturbado:

$$\underset{\tilde{\underline{\beta}}}{Min} \left\| (\underline{y} - \delta\underline{y}) - (X + E)\tilde{\underline{\beta}} \right\|_2^2,$$

donde

$$\left\| \delta\underline{y} \right\|_2 \leq \varphi(m,n) \left\| \underline{y} \right\|_2 u \quad \text{y} \quad \left\| E \right\|_2 \leq \varphi(m,n) \left\| X \right\|_2 u$$

Teorema 4.5.1 *Sea* X, $X+E \in \Re^{nxp}$, $\underline{y}$, $\delta\underline{y} \in \Re^n$, $\hat{\underline{\beta}}^+ = X^+\underline{y}$ *y* $\hat{\underline{\beta}}_a^+ = (X+E)^+(\underline{y}+\delta\underline{y})$.

Si $rgo(X+E) \leq rgo(X)$ *y* $\|E\|_2\|X^+\|_2 < 1$ *entonces*

$$\frac{\left\|\hat{\underline{\beta}}_a^+ - \hat{\underline{\beta}}^+\right\|_2}{\left\|\hat{\underline{\beta}}^+\right\|_2} \leq \frac{\kappa_2(X)}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\left[2\frac{\|E\|_2}{\|X\|_2} + \frac{\kappa_2(X)\,\|\underline{r}\|_2}{\left\|P_{im(X)}\underline{y}\right\|_2}\frac{\|E\|_2}{\|X\|_2} + \frac{\|\underline{y}\|_2}{\left\|P_{im(X)}\underline{y}\right\|_2}\frac{\|\delta\underline{y}\|_2}{\|\underline{y}\|_2}\right]$$
$$+\kappa_2(X)\frac{\|E\|_2}{\|X\|_2} \qquad (4.12)$$

donde

$$\begin{aligned} \kappa_2(X) &= {}_{def}\ \|X\|_2\left\|X^+\right\|_2 \\ \underline{r} &= \underline{y} - X\hat{\underline{\beta}}^+ \\ P_{im(X)}\underline{y} &= X\hat{\underline{\beta}}^+ \end{aligned}$$

Demostración : *Véase Stewart-Sun [17].*

Comentarios:

1. La cota de perturbación relativa de $\hat{\underline{\beta}}^+$ está dominada por su segundo término:

$$\frac{[\kappa_2(X)]^2}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\frac{\|\underline{r}\|_2}{\left\|P_{im(X)}\underline{y}\right\|_2}\frac{\|E\|_2}{\|X\|_2}.$$

Es importante remarcar la aparición de $[\kappa_2(X)]^2$ como factor en dicho término. Recuerde que este mismo factor dominaba la cota de perturbación relativa para $\hat{\underline{\beta}}$ cuando se calculaba vía las ecuaciones normales.

2. La cota de perturbación relativa de $\hat{\underline{\beta}}^+$ dada por el teorema 4.5.1 se

puede rescribir como sigue:

$$\frac{\left\|\hat{\underline{\beta}}_a^+ - \hat{\underline{\beta}}^+\right\|_2}{\left\|\hat{\underline{\beta}}^+\right\|_2} \leq \frac{\kappa_2(X)}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\left[2\frac{\|E\|_2}{\|X\|_2} + \tan(\theta)\kappa_2(X)\frac{\|E\|_2}{\|X\|_2} + \sec(\theta)\frac{\|\delta\underline{y}\|_2}{\|\underline{y}\|_2}\right]$$
$$+\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}$$

donde θ =ángulo que forma $\underline{y}$ con $im(X)$ (véase figura 4.1).

3. La cota de perturbación relativa de $\hat{\underline{\beta}}^+$ dada por el teorema se puede rescribir como:

$$\frac{\left\|\hat{\underline{\beta}}_a^+ - \hat{\underline{\beta}}^+\right\|_2}{\left\|\hat{\underline{\beta}}^+\right\|_2} \leq \frac{\kappa_2(X)}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\left[2\frac{\|E\|_2}{\|X\|_2} + \frac{\sqrt{1-R^2}}{|R|}\kappa_2(X)\frac{\|E\|_2}{\|X\|_2} + \frac{1}{|R|}\frac{\|\delta\underline{y}\|_2}{\|\underline{y}\|_2}\right]$$
$$+\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}$$

o bien como

$$\frac{\left\|\hat{\underline{\beta}}_a^+ - \hat{\underline{\beta}}^+\right\|_2}{\left\|\hat{\underline{\beta}}^+\right\|_2} \leq \frac{\kappa_2(X)}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\left[2\frac{\|E\|_2}{\|X\|_2} + \sqrt{\frac{n-p}{p}}\frac{\kappa_2(X)}{\sqrt{F_{(p,n-p)}}}\frac{\|E\|_2}{\|X\|_2} + \frac{1}{|R|}\frac{\|\delta\underline{y}\|_2}{\|\underline{y}\|_2}\right]$$
$$+\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}$$

De donde se vé que si: $\frac{[\kappa_2(X)]^2}{F_{(p,n-p)}} \approx \frac{p}{n-p}$ entonces la cota de error relativo para $\hat{\underline{\beta}}^+$ se puede simplificar como sigue:

$$\frac{\left\|\hat{\underline{\beta}}_a^+ - \hat{\underline{\beta}}^+\right\|_2}{\left\|\hat{\underline{\beta}}^+\right\|_2} \approx \frac{\kappa_2(X)}{1-\kappa_2(X)\frac{\|E\|_2}{\|X\|_2}}\left[3\frac{\|E\|_2}{\|X\|_2} + \frac{\|\delta\underline{y}\|_2}{\|\underline{y}\|_2}\right] + \kappa_2(X)\frac{\|E\|_2}{\|X\|_2},$$

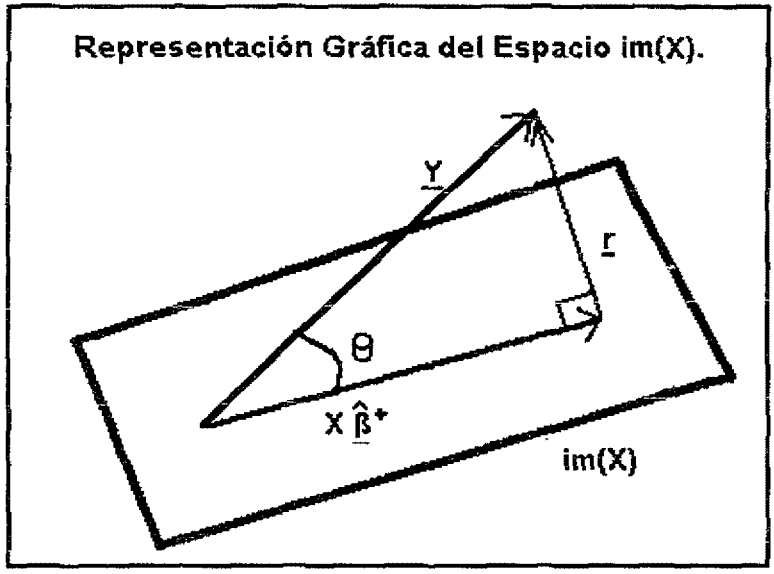


Figura 4.1: Representación Gráfica del Espacio $im\,(X)$.

al considerar que $F_{(p,n-p)} >> 1$ implica que $|R| \approx 1$.

El teorema 4.5.1 permite de esta manera conocer la veracidad que brindará el vector de parámetros estimados $\underset{\sim}{\hat{\beta}}^{+}$, en base al conocimiento previo de la matriz X y el vector de respuesta $\underset{\sim}{y}$. Así también, en el comentario 3 se muestra una expresión obtenida a partir de la teoría del Análisis Numérico que relaciona términos netamente numéricos con términos propios de la teoría estadística del Análisis de Regresión Lineal.

Esto ayuda a visualizar en forma clara que estas dos áreas del estudio científico guardan una gran relación, y que siempre que aparece una, es necesario auxiliarse de la otra para así tener un análisis más profundo de los fenómenos de estudio y de esta manera tener una visión más clara del problema real.

4.5.2 Hipótesis del ángulo agudo.

En esta subsección se tiene por objeto responder a la pregunta: ¿bajo qué condiciones el segundo sumando del lado derecho en (4.12) es del orden del primero?

Esto es, se pide hallar las condiciones bajo las cuales se cumple la siguiente desigualdad:

$$\kappa_2(X)\frac{\|\underline{r}\|_2}{\left\|X\hat{\underline{\beta}}^+\right\|_2}\frac{\|E\|_2}{\|X\|_2} \leq \frac{\|E\|_2}{\|X\|_2}. \tag{4.13}$$

Esto es equivalente a:

$$\kappa_2(X)\frac{\|\underline{r}\|_2}{\left\|X\hat{\underline{\beta}}^+\right\|_2} \leq 1,$$

equivalente a:

$$\kappa_2(X)\|\underline{r}\|_2 \leq \left\|X\hat{\underline{\beta}}^+\right\|_2 \leq \|X\|_2\left\|\hat{\underline{\beta}}^+\right\|_2 \tag{4.14}$$

Como $\kappa_2(X) = \|X\|_2\|X^+\|_2$, (4.14) se simplifica a

$$\|X^+\|_2\|\underline{r}\|_2 \leq \left\|\hat{\underline{\beta}}^+\right\|_2 \tag{4.15}$$

Por otro lado:

$$\hat{\underline{\beta}}^+ = X^+\underline{y} = X^+P_{im(X)}\underline{y}$$

entonces

$$\left\|\hat{\underline{\beta}}^+\right\|_2 \leq \|X^+\|_2\|P_{im(X)}\underline{y}\|_2$$

o sea que:

$$\left\|\hat{\underline{\beta}}^+\right\|_2^2 \leq \|X^+\|_2^2\|P_{im(X)}\underline{y}\|_2^2 = \|X^+\|_2^2\left(\|\underline{y}\|_2^2 - \|\underline{r}\|_2^2\right) \tag{4.16}$$

De (4.15) y (4.16) se tiene

$$\|X^+\|_2^2\|\underline{r}\|_2^2 \leq \|X^+\|_2^2\left(\|\underline{y}\|_2^2 - \|\underline{r}\|_2^2\right)$$

es decir

$$\frac{\|\underline{r}\|_2^2}{\|\underline{y}\|_2^2} \leq \frac{1}{2}$$

entonces (de la figura 4.1)

$$sen\,(\theta) = \frac{\|\underline{r}\|_2}{\|\underline{y}\|_2} \leq \frac{1}{\sqrt{2}}$$

Por lo tanto el primer término de la desigualdad en (4.13) domina al segundo si

$$\theta = \measuredangle\left(\underline{y}, im\,(X)\right) \leq 45^o$$

El anterior razonamiento es conocido como hipótesis del ángulo agudo y es precisamente la condición requerida para garantizar que la solución generalizada bajo perturbaciones $\underline{\hat{\beta}}_a^+$ no distará en mucho de la solución generalizada real $\underline{\hat{\beta}}^+$.

Capítulo 5

Manual del usuario.

Introducción.

En este capítulo se presenta el objetivo fundamental de la realización de este trabajo: *un sistema amigable y transparente al usuario*, que es de muy fácil manejo y evita tareas engorrosas como son la edición, compilación y ligado de programas. Además, se ofrece una breve explicación de lo que es el sistema así como de cada uno de los resultados que presenta cada opción, esto para facilitar la interpretación de las estadísticas a aquellas personas que no sean expertas en la materia. Las tareas que se pueden realizar con el sistema son:

1. La estimación de los parámetros $\underline{\beta}$ del modelo lineal

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon},$$

   mediante el método de Mínimos Cuadrados.

2. Obtener las estadísticas necesarias para llevar a cabo un análisis del modelo de regresión lineal ajustado, tales como realizar pruebas de hipótesis, el cálculo de indicadores que permiten efectuar diagnósticos sobre el comportamiento del modelo (correlación entre regresores, colinealidad y observaciones aberrantes).

3. Llevar a cabo un análisis tanto gráfico como estadístico de los datos, así como la verificación de los supuestos planteados sobre el vector de errores $\underline{\varepsilon}$ no observable (independencia de los errores, homocedasticidad, etc.).

5.1 El sistema ANA_RELI.M VER. 2.0.

El objetivo principal de este sistema llamado ANA_RELI.M VER. 2.0 (Análisis de Regresión Lineal), es brindar al usuario una herramienta amigable y de fácil manejo, de tal manera que no se tenga ningún problema al estar realizando el análisis gráfico y/o estadístico del modelo de regresión.

Con este sistema se trata de resolver eficazmente, desde el punto de vista numérico, el modelo lineal

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{5.1}$$

bajo las siguientes condiciones

$$\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I) \quad \text{ó} \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I) \tag{5.2}$$

$$X \in \Re^{n\times p},\ n > p,\ \text{con } rgo(X) = p \tag{5.3}$$

Bajo las condiciones (5.2), (5.3) para el modelo (5.1), y bajo los supuestos:

1. que X no es mal-condicionada (i. e. X no es de rango deficiente), y
2. que se satisface la hipótesis del ángulo agudo (véase sección 4.5.2) para el modelo (5.1);

se tiene completa garantía sobre la confiabilidad de los resultados que el sistema proporciona.

5.2 Análisis estadístico y gráfico que efectúa el sistema.

Las funciones básicas que proporciona el sistema sirven para llevar a cabo un análisis estadístico del modelo bajo estudio, así como de los datos del mismo.

Las tareas que puede realizar el sistema son:

1. Estandarización y/o centralización del modelo.
2. El cálculo de las estadísticas básicas tales como media, varianza, desviación estándar, máximo y mínimo de cada regresor, así como de la variable de respuesta.

3. El cálculo de la estimación del vector de parámetros $\underline{\beta}$, así como la desviación estándar $\left(se\left(\hat{\beta}_i\right)\right)$ para cada uno de los parámetros, así como su respectivo cuantil $t-$student para la realización de la prueba de hipótesis:

$$H_0 : \beta_i = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0 \quad (i = 1, \ldots, p-1)$$

4. El cálculo de la estimación de la varianza insesgada para los residuales.
5. El coeficiente de determinación múltiple y el coeficiente ajustado.
6. El cálculo de la tabla ANOVA para llevar a cabo la validación del modelo

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0 \quad \text{p. a. } i = 1, \ldots, p-1$$

7. El cálculo de la matriz de correlaciones entre los distintos regresores y la variable de respuesta.
8. El cálculo de residuales estándarizados, estudentizados, leverage de las observaciones y distancias de Cook.
9. La evaluación del estadístico Durbin-Watson para autocorrelación.
10. La detección de colinealidad.
11. La detección de observaciones aberrantes ("outliers").
12. El cálculo de predicciones.
13. La graficación de y Vs. $\hat{y}$.
14. La graficación de los Residuales estándarizados.
15. La graficación de los Residuales estándarizados Vs. y.
16. La graficación de los Residuales estándarizados Vs. $\hat{y}$.
17. La graficación de probabilidad Normal.

Entre otras tareas estadísticas y gráficas, las cuales son calculadas con los mejores métodos numéricos conocidos, y se indica el significado estadístico de cada una de ellas, para que **no únicamente** personas con conocimientos en la materia puedan interpretarlas.

En la sección 5.4, *Diseño del Sistema* se mencionará en qué opción se realiza cada una de las tareas antes mencionadas.

5.3 Estructura modular del sistema.

La integración general del sistema ANA_RELI.M VER. 2.0 está basada en módulos, los cuales tienen características y tareas bien definidas y que hacen que se diferencien entre sí, los módulos por los que está comprendido son los siguientes:

1. Módulo de Presentación del Sistema.
2. Módulo de Presentación del Demo del Sistema.
3. Módulo de In eracción con el Usuario (ó de Entrada).
4. Módulo de Cálculos Numéricos.
5. Módulo de Presentación de Estadísticas y Gráficas.
6. Módulo de Salidas.

Módulo de Presentación del Sistema.

En este se da la bienvenida al usuario al sistema y aparecen opciones que permiten entrar a los siguientes módulos, estas opciones son:

1. **Regresión:** permite el acceso al módulo de Presentación de Estadísticas y Gráficas, donde se encuentran todas y cada una de las opciones con que cuenta el sistema.
2. **Demo:** permite la entrada al módulo de Demostración del Sistema, en éste se da una breve explicación de lo que cada una de las opciones del sistema presenta, así como un breve resumen del significado de la estadística obtenida.

3. **Info:** presenta una breve información general de lo que es el sistema.

4. **Referencias:** se presenta al usuario una lista de los textos que se pueden consultar en caso de que se desee profundizar más sobre algún tema en particular aquí mostrado

Módulo de Presentación del Demo del Sistema.

Como se mencionó anteriormente, aquí se presentan cada una de las opciones con que cuenta el sistema, así como la opción en la que pueden ser encontradas cada una de las estadísticas y las gráficas, acompañadas con una breve explicación del significado o interpretación de las mismas.

Módulo de Interacción con el Usuario (ó de Entrada).

Localizada en el módulo anterior, se permite la captura de los datos tales como la matriz de observación X y el vector de respuesta $\underline{y}$ del modelo $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$, así como el número de observaciones y el número de regresores en dicho modelo, es aquí donde se pueden guardar los datos en un archivo con el nombre elegido por el usuario.

Módulo de Cálculos Numéricos.

Una vez capturados los datos, se indican las cualidades del mismo e inmediatamente se realizan los cálculos necesarios para obtener todas las estadísticas, este módulo tiene encargada dicha tarea y hace conección directa con el siguiente módulo.

Módulo de Presentación de Estadísticas y Gráficas.

Para llevar a cabo el análisis de la regresión lineal ajustada, en la opción "Regresión", en el módulo de "Presentación", se muestran todas aquellas estadísticas y gráficas necesarias para tal finalidad, además de permitir la entrada al módulo de Interacción con el Usuario.

Módulo de Salidas.

Este permite que los datos calculados en el módulo anterior sean grabados en archivos para posteriormente ser llamados por el módulo de estadísticas y gráficas para presentar resultados en pantalla. Los archivos generados son:

1. **Anova.rmr:** se guardan las estadísticas tales como las Sumas de Cuadrados para la construcción de la tabla de Análisis de Varianza, los Coeficientes de Correlación Múltiple, de Determinación, el Ajustado, y la estimación de la raíz de la varianza de los residuales.

2. **Matcor.rmr:** se guardan los coeficientes de correlación simple entre cada regresor x_i con x_j, y entre cada regresor con la variable de respuesta y

3. **Outliers.rmr:** se guardan los posibles outliers u observaciones aberrantes del modelo, en caso de no registrarse outliers se guarda un indicador que especifica la ausencia de ellos.

4. **Residual.rmr:** se guardan los valores de los distintos residuales, tales como los residuales comunes, los estándarizados, los estudentizados; además de los leverages de cada observación (h_{ii}) y las distancias de Cook (D_i), y por las dimensiones de este archivo se guarda también el vector de valores ajustados por la regresión $\hat{\underline{y}}$.

5. **Resumreg.rmr:** aquí se guardan las estadísticas básicas tales como media y varianzas de los regresores x_i y la variable de respuesta y, su desviación estándar y los valores máximo y mínimo de cada uno de ellos; también se guardan el vector de parámetros estimados $\hat{\underline{\beta}}$, la desviación estándar de cada estimación, su respectivo cuantil $t-$student, los Coeficientes de Inflación de la Varianza; finalmente el estadístico Durbin-Watson para autocorrelación.

Cada uno de los módulos anteriores fueron programados en Matlab Ver. 5 0. así como algunos procedimientos con algoritmos que permiten la reducción de los errores por redondeo y que son numéricamente estables, como son el cálculo de la media y la varianza (calculadas en MEDYVAR.M), y la actualización de las matrices Q y R en la descomposición QR de la matriz X cuando es eliminada una observación, esto para evitar nuevamente el cálculo de dicha descomposición al quitar una observación en la matriz de datos X (calculada en QRDRMROW.M, auxiliada de las funciones ROTGEN.M y ROTAPP.M), y para optimizar los cálculos que se efectúan se utilizan las rutinas desarrolladas por Stewart [16].

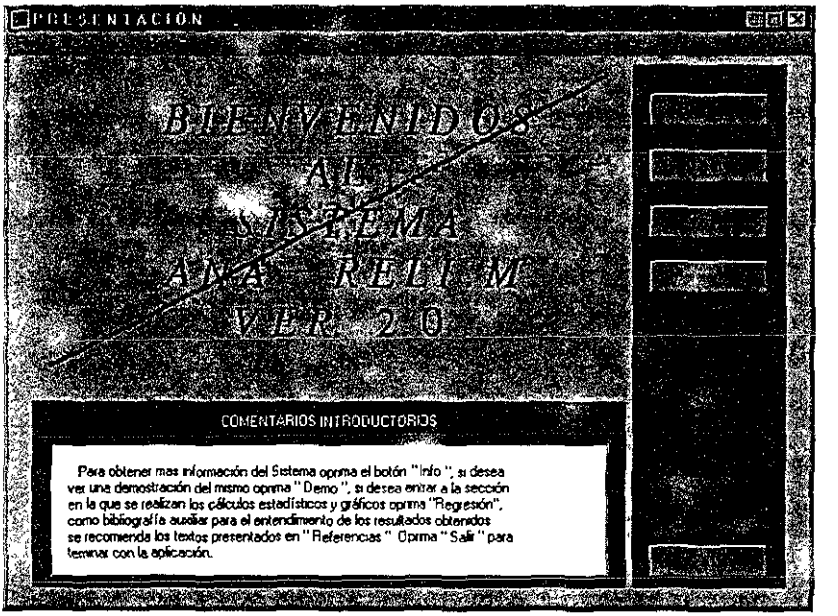


Figura 5.1: Presentación del Sistema ANA_RELI.M VER 2.0

5.4 Diseño del sistema.

A continuación se presenta la conformación del sistema detallando las funciones que realiza cada menú y sus correspondientes opciones.

5.4.1 Ventana de presentación.

En esta se presenta la bienvenida al usuario y un cuadro con comentarios donde se indica el botón a oprimir para el inicio del siguiente procedimiento. Las opciones presentadas son:

1. **Regresión:** abre una nueva ventana donde se presentan las opciones para el análisis de los datos, de las variables y del modelo, más adelante se tratará a fondo cada parte de esta opción.

2. **Demo:** abre la ventana en la cual se presenta una breve demostración de las opciones que conforman el sistema, donde se pueden localizar y lo que significan, cada una en forma brevemente explicada.

3. **Info:** abre la ventana en la que se presenta información general del sistema, los objetivos del mismo, alcances y ventajas que posee respecto a los procedimientos que realiza.

4. **Referencias:** abre la ventana en donde se presenta la bibliografía de algunos textos propuestos en caso de que el analista desee profundizar un poco más en cuanto a la ayuda prestada por el sistema sobre las estadísticas y gráficas aqui calculadas.

Estas opciones son presentadas en la figura 5.1, que es la pantalla de "Presentación" del sistema Ana_Reli.M Ver. 2.0.

5.4.2 Opción de "Regresión".

Con esta opción se abre una nueva ventana donde se presentan las opciones para la introducción de datos, las características del modelo; así como el análisis de los datos, de las variables que intervienen en el modelo y de los supuestos hechos en el planteamiento del modelo.Las siguientes, son las opciones que se incluyen en "Regresión", y la ventana de esta opción está dada en la figura 5.2.

Datos.

En esta opción se presentan dos opciones que son las formas de cómo serán introducidos los datos para que posteriormente el sistema realice los cálculos numéricos convenientes. Estas opciones son:

Entrada Teclado. Abre una pantalla para que el usuario introduzca en forma directa el número de regresores que tendrá el modelo a ajustar, el número de observaciones, la matriz de datos X y el vector de respuesta $\underline{y}$. Una vez capturados estos datos, automáticamente el sistema guarda los datos en un archivo con nombre "datos.arl", o bien, el usuario puede darle un nombre en particular al conjunto de datos.

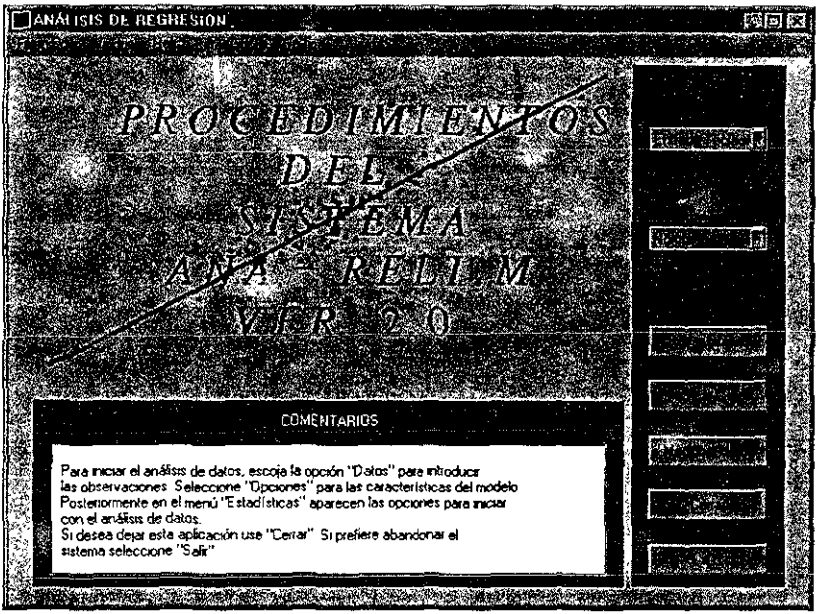


Figura 5 2: Ventana de Regresión del Sistema ANA_RELI.M VER.2.0

Archivo. El sistema presenta una caja de diálogo desde la cual el usuario busca el archivo donde se encuentran los datos que serán leidos por el sistema, siempre y cuando hayan sido grabados con el formato establecido por el sistema.

Opciones.

Con esta se eligen las características del modelo de regresión lineal a ser ajustado, las alternativas son:

Normal. Con esta opción el modelo de regresión a ajustar presentará la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

calculándose todos los parámetros sin modificación alguna.

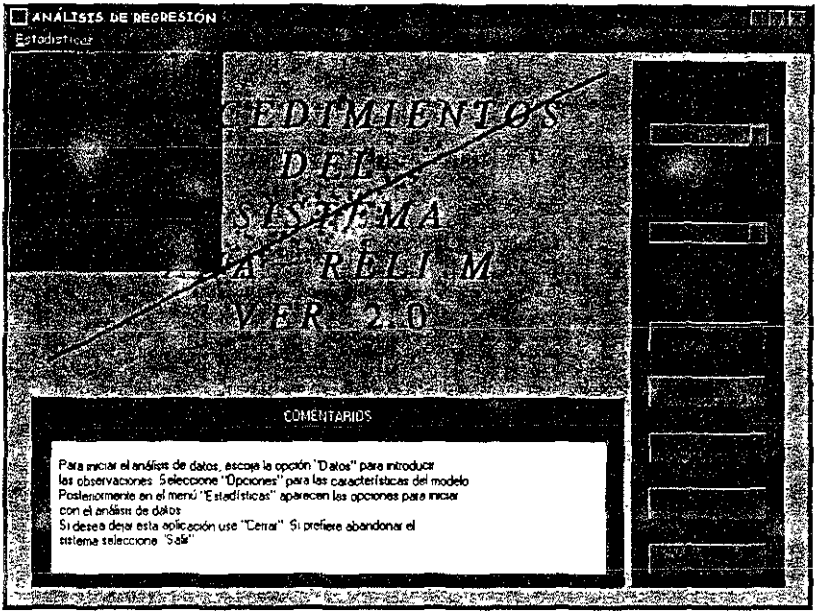


Figura 5.3: Menú "Estadísticas" del sistema ANA_RELI.M VER 2.0

Centrar. Con esta opción el modelo a ajustar presenta la cualidad de que cada uno de sus regresores estará centrado respecto a su media (véase sección 1.3.1). El modelo a ajustar tendrá la siguiente forma:

$$y_i - \overline{y} = \beta_1 (x_{i1} - \overline{x}_1) + \cdots + \beta_{p-1} (x_{i,p-1} - \overline{x}_{p-1})$$

o sea

$$y' = \beta_1 x'_1 + \cdots + \beta_{p-1} x'_{p-1}$$

y todas las estadísticas a calcular se harán a partir de este modelo.

Estándarizar. Con esta opción cada uno de los regresores del modelo, incluyendo la variable de respuesta, estarán centrados respecto a su media y divididos por su desviación estándar (véase sección 1.3.1).

•**Menú "Estadísticas".**

Al desplegar este menú se presentan las opciones estadísticas con que cuenta el sistema (véase figura 5.3), que permiten entre otras cosas, el análisis de los regresores, el análisis del modelo mediante pruebas estadísticas, y el análisis de los datos, las opciones son las siguientes:

Estadísticas Básicas. La Media muestral para cada regresor $(\overline{x}_i)$, así como de la variable de respuesta $(\overline{y})$, la Varianza muestral $\left(s^2_{x_i}\right)$, la Desviación Estándar (s_{x_i}), el Valor Máximo y Mínimo para cada uno de ellos son mostrados al seleccionar esta opción.

Resumen de la Regresión. Indicadores importantes son mostrados al elegir esta opción pues se presentan el Coeficiente de Correlación Múltiple $\left(r_{y,\hat{y}}\right)$, el Coeficiente de Determinación (R^2) y el Coeficiente de Determinación Ajustado (R^2_a), que permiten verificar la explicación de los regresores sobre la variable de respuesta; se presenta también la estimación insesgada de la Desviación Estándar de los residuales $\left(s = \sqrt{\frac{SCE}{n-p}}\right)$; así como la estimación del Vector de Parámetros $\left(\hat{\underline{\beta}}\right)$, la desviación estándar para cada estimador $\left(se\left(\hat{\beta}_i\right)\right)$, y el cuantil t-student de cada regresor.

Análisis de Varianza. Los resultados mostrados por esta opción son de suma importancia pues con ellos se puede llevar a cabo la validación del modelo mediante la prueba de hipótesis:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0 \quad p.a. \quad i = 1, \ldots, p-1$$

Los datos presentados son las Sumas de Cuadrados de la Regresión (SCR), la de los Errores (SCE) y los Totales (SCT); los grados de libertad de cada una de ellas, los Cuadrados Medios y el valor del estadístico F con el cual se determina la aceptación o no de la hipótesis nula.

Matriz de Correlaciones. Los Coeficientes de Correlación Simple entre cada uno de los distintos regresores (r_{x_i,x_j}) y de los regresores con la variable de respuesta $(r_{x_i,y})$, son presentados con esta opción.

Detección de Colinealidad. Algunas estadísticas para detectar colinealidad son presentadas en esta opción, remarcando que quizá no sean suficientes para su detección o que sólo algunas de ellas la detecten. Se presentan los Coeficientes de Correlación Simple entre cada regresor con la variable de respuesta $(r_{x,y})$, la estimación de cada parámetro del modelo $\left(\hat{\beta}_i\right)$ y los Coeficientes de Inflación de la Varianza (CIV) de cada regresor incluido en el modelo. (*Nota: esta opción únicamente funciona para aquellos modelos* con dos o más regresores).

Análisis de Residuales. Las estadísticas mostradas por esta opción permiten el análisis de los datos. Se presentan los Residuales (e_i), los Residuales estándarizados (rs_i), los Residuales estudentizados (rt_i), el Leverage de cada observación (h_{ii}) y las Distancias de Cook (D_i).

Durbin-Watson. Presenta el valor del estadístico necesario para la verificación de la prueba de Autocorrelación en los residuales, conocido como Estadístico Durbin-Watson (DW).

Outliers. Las observaciones que causan problemas en el ajuste del modelo debido a que tienen valores extremos son presentadas por esta opción, incluyendo su Valor Observado (y), el Valor Ajustado por el modelo de regresión $\left(\hat{y}\right)$, el Valor de su Residual (e_i) y el Residual estándarizado (rs_i). En caso de no haber outliers el sistema enviará un mensaje.

Predicción. Un ventana es abierta para que el usuario introduzca el valor (o vector) del regresor (o regresores) en el cual quiera que se calcule dicha predicción $\left(\hat{y}\right)$ o pronóstico.

•Menú "Gráficas".

El análisis gráfico necesario para la verificación de los supuestos del modelo se puede llevar a cabo desplegando las opciones con que cuenta el menú de Gráficas (véase figura 5.4), las cuales se presentan a continuación:

Histograma de Y. Se grafica el histograma de los Valores Observados (y) para verificar si tienen una tendencia cercana a la distribución Normal.

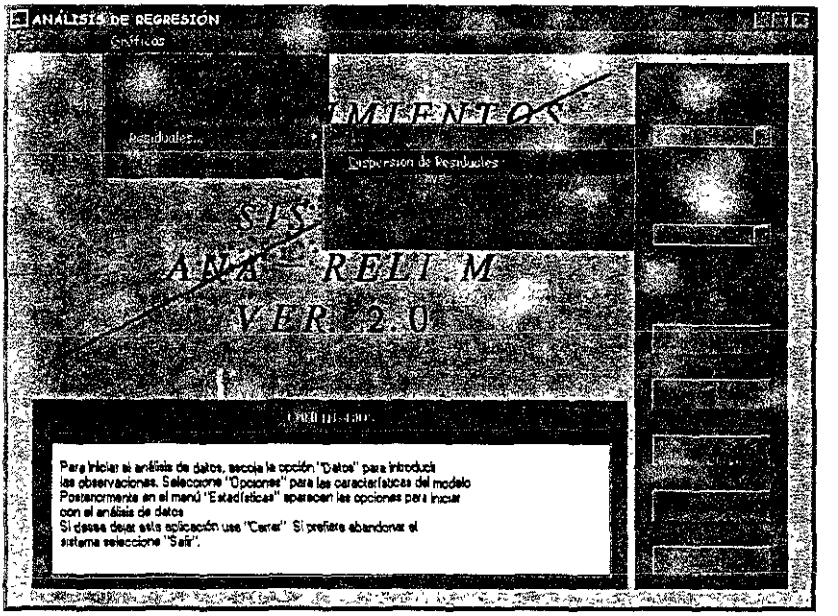


Figura 5.4: Menú "Gráficas" del sistema ANA_RELI.M VER. 2.0

Histograma de Y Estimado. Se grafica el histograma de los Valores Ajustados $\left(\hat{y}\right)$ por el Modelo de regresión para verificar si estos presentan una tendencia cercana a la Normal.

Observados y Predecidos. Se presentan los Valores Observados y los Valores Ajustados por el modelo, si ambas tienen la misma tendencia y no son muy distantes una de la otra entonces, se puede suponer que el modelo ajusta en forma adecuada a los valores observados.

Histograma de Residuales Estándarizados. El supuesto de normalidad en los residuales es sumamente importante pues a partir de éste se genera toda la teoría estadística sobre el modelo de regresión lineal. El Histograma de Residuales estándarizados permite verificar si tal supuesto se cumple, más

adelante se presentan otras gráficas con la misma finalidad.

Dipersión de Residuales. Se grafican los valores de los residuales con la respectiva observación de la que provienen, esto para no perder el orden de dicho residual. La gráfica esencialmente sirve para localizar patrones o tendencias que pudieran tener los residuales y también para verificar si la varianza es constante (homocedasticidad).

Dipersión de Residuales Estándarizados. Se grafican los valores de los Residuales estándarizados con la respectiva observación de la que provienen. La interpretación de esta gráfica es similar a la anterior aunque aquí es permitida la observación de posibles outliers o datos aberrantes, debido a la estándarización de residuales.

Estándarizados Vs. Y. Se grafican los Residuales estándarizados contra los valores observados para la verificación del supuesto de varianza constante en los residuales, así como del supuesto de independencia (o aleatoriedad) entre ellos.

Estándarizados Vs. Y Estimado. Se grafican por un lado los Residuales estándarizados y por el otro los Valores Ajustados (Estimados) por el modelo. La interpretación de esta gráfica es similar a la anterior pero ésta además permite la detección de posibles outliers o datos aberrantes en las observaciones.

Probabilidad Normal. Se grafica por un lado el Cuantil Esperado de la Probabilidad Normal para cada una de las observaciones y por el otro el respectivo valor del Residual estándarizado, para que se cumpla el supuesto de Normalidad en los Residuales los puntos resultantes deben presentar una tendencia lineal o estar más o menos alineados; si los puntos no guardan esta tendencia se considera una violación a dicho supuesto, aquellos puntos que estén alejados de la linea, se consideran posibles outliers.

•Menú "Ayuda...".

En este menú se despliega ayuda básica que auxilia al usuario a entender y manejar en forma más sencilla cada una de las aplicaciones del sistema, y a

encontrar en forma más rápida las estadísticas o gráficas deseadas. Las tres opciones que se presentan son las siguientes

1. **Cómo empezar:** introduce al usuario al manejo inical del sistema.
2. **Opciones Estadísticas:** presenta información y ayuda para cada una de las opciones estadísticas, vistas anteriormente.
3. **Opciones Gráficas:** presenta información y ayuda de cada una de las gráficas mencionadas anteriormente, con que cuenta el sistema.

5.4.3 Opción "Demo".

Una demostración general de las opciones con que cuenta el sistema son presentadas al entrar en la opción "Demo", localizada en la ventana de "Presentación" (véase figura 5.1). Todas las Estadísticas y Gráficas vistas anteriormente se encuentran en éste módulo, mencionándose el nombre con el que aparecen en el menú y una breve explicación de lo que cada opción presenta.

5.4.4 Opción "Info".

Esta opción presenta información general de lo que es en sí el sistema, dando una breve explicación de la importancia del ajuste de modelos de Regresión Lineal mediante el método de Mínimos Cuadrados, los aspectos numéricos que se han desarrollado para la resolución del problema anterior y las ventajas y desventajas que presentan, también se mencionan las ventajas con que cuenta la realización de este proyecto. Opción localizada en la ventana de "Presentación" (véase figura 5.1).

5.4.5 Opción "Referencias".

En ocasiones el analista desea profundizar más en algún tema en particular o simplemente desea conocer textos para apoyo didáctico, es por ello que en la opción de "Referencias", localizada en la ventana de "Presentación" (véase figura 5.1), se presentan algunos textos bibliográficos, estadísticos y numéricos, en los cuales están basados la mayor parte de los comentarios con los que cuenta el sistema. Los textos consultados son de las ediciones más recientes, para asegurar que los comentarios aquí presentados tienen sus fundamentos en nuevas interpretaciones.

5.5 Manejo del sistema.

La forma de cómo operar el sistema es muy sencilla, la ventaja de estar bajo ambiente Windows lo hace manipulable y de fácil dominio en su manejo. El lector ya se habrá familiarizado con el diseño del sistema ANA_RELI.M VER. 2.0 al observar las gráficas presentadas a lo largo de éste trabajo. Como habrá notado todas ellas están diseñadas a base de botones, ésto lo hace fácil de manejar y rápido de entender. Las únicas interfaces con el usuario son a la hora de introducir la base de datos desde el teclado y al querer calcular alguna predicción de la variable de repuesta.

El usuario sólo tiene que posicionar el cursor del mouse en el lugar donde esté la opción que desea ejecute el sistema y automáticamente sale en pantalla una ventana con los resultados solicitados por el usuario, o bien una caja de diálogo en la cual se permite la continuación de la opción elegida o la detención del proceso.

Una vez introducidos la matriz de datos X y el vector de respuesta $\underline{y}$, el siguiente paso es guardar éstos en un archivo, si así se desea (por default el sistema los guarda en un archivo con nombre "datos.arl"); posteriormente hay que indicar las características del modelo, de lo contrario se presentarán los resultados de la última regresión ajustada.

5.6 Requerimientos de instalación.

Este software fue diseñado para ejecutarse en Matlab bajo ambiente Windows versión 5.0 en adelante.

Para poder ejecutar el sistema es necesario tener dicha versión de Matlab y correrlo como un programa propio de éste, pues hace uso de algunas de sus librerías como son para la obtención de la descomposición QR, para la parte gráfica, entre otras.

En el momento de ejecutar el sistema ANA_RELI.M VER. 2.0, hay que indicar en Matlab la ruta donde se encuentran dichos programas.

Ejemplo 5.6.1 *Supóngase que todas las funciones del sistema ANA_RELI.M VER. 2.0 se encuentran en el directorio* **d : \Programas\Anareli**, *la sintaxis requerida para que el sistema pueda ejecutarse es la siguiente:*

```
path(path,'d : \Programas\Anareli')
```

a continuación teclee el nombre `Anareli`, *inmediatamente después de introducir esta opción aparecerá la pantalla de* "Presentación" *(figura 5.1) del sistema* ANA_RELI.M VER.2.0.

5.7 Conclusiones.

El desarrollo de este sistema trató de conjuntar herramientas propias del Análisis Numérico que pudieron ser aplicadas en forma práctica en el desarrollo de la teoría del Análisis de Regresión Lineal, aunado a ésto, se introdujo un poco de la ciencia computacional de tal manera que se logró crear la primera parte de un proyecto que encuentra sus inicios en el planteamiento del modelo lineal:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

y del cual todavía hay mucho material por desarrollar y que es discutido en forma breve en el Apéndice B

Apéndice A

Manual del Sistema ANA_RELI.M VER 2.0.

Introducción.

En diferentes áreas del conocimiento como son la economía, la biología, la medicina, la física, etc. se presentan problemas que consisten en la observación de un fenómeno y sus posibles causas; lo que los investigadores desean es conocer la relación existente entre el fenómeno (conocido como variable de respuesta o dependiente) y las causas que lo originan (variables de control o independientes), para lograrlo suponen que el comportamiento de la variable de respuesta corresponde a un modelo lineal representado de la siguiente manera:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \tag{A.1}$$

donde X es una matriz que contiene los datos de las observaciones, $\underline{\beta}$ son parámetros desconocidos del modelo, $\underline{\varepsilon}$ son los errores estocásticos, e $\underline{y}$ es el vector de la respuesta observada.

Uno de los supuestos básicos del modelo (A.1) es que $\underline{\varepsilon}$ tiene media cero y matriz de varianza-covarianza $\sigma^2 I$, i. e.

$$\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I)$$

Y para poder llevar a cabo hipótesis estadísticas, se supone adicionalmente normalidad, i. e.

$$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I)$$

La forma inicialmente vista para obtener la estimación $\hat{\underline{\beta}}$ a los parámetros $\underline{\beta}$ es resolviendo el sistema de ecuaciones normales

$$X^t X \,\tilde{\underline{\beta}} = X^t \underline{y}; \tag{A.2}$$

sin embargo, su cálculo directo con frecuencia presenta problemas numéricos, por ello se utiliza la forma alternativa que es la descomposición QR de la matriz X, i e.

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

donde Q es una matriz ortogonal y R es una matriz triangular superior, si además se descompone a $Q = [Q_x, Q_\perp]$ donde Q_x tiene p columnas, el problema planteado en (A.2) se reduce a resolver el sistema triangular superior

$$R \,\hat{\underline{\beta}} = Q_x^t \underline{y}$$

Para la obtención de la descomposición QR de la matriz X, se emplean librerías realizadas en Matlab 5.0, el cual es un paquete numérico que poco a poco va tomando fuerza dentro de la industria del cómputo científico y que ofrece a sus usuarios resultados completamente confiables.

A.1 Estructura del sistema.

Como se mencionó anteriormente, todas las funciones del sistema ANA_RELI.M VER 2.0 fueron programadas en Matlab 5.0, por lo que presentan una estructura similar a los procedimientos realizados por Matlab. A continuación se hace una clasificación de las funciones que conforman el sistema, agrupándose según sus características genéricas en el mismo.

A.1.1 Funciones de cálculos numéricos.

•Calculos.m

Esta función recalcula la matriz X para los casos en que se pidió centralización o estandarización de datos, llama a la función QR.M para obtener la descomposición QR de la matriz X, a partir de la cual ésta función calcula todas las estadísticas; tales como medias y varianzas de cada x_i, coeficientes

de determinación y correlación, estimación de parámetros, sumas de cuadrados, residuales, observaciones aberrantes, etc., necesarias para el análisis estadístico y gráfico del modelo de regresión lineal ajustado. Todos estos resultados son guardados en archivos "*.rmr", de tal manera que únicamente se hacen los procedimientos una sola vez, a excepción de cuando se quitan las observaciones aberrantes del modelo.

Funciones que llama:

De Ana_Reli 2.0: medyvar.m

De Matlab 5.0: clear.m, exist.m, for.m, if.m, inv.m, load.m, nargin.m, ones.m, qr m, save.m, size.m, sqrt.m, strcmp.m

•Medyvar.m

Esta función calcula la media y la varianza ya sea de un vector o de las columnas de una matriz mediante el algoritmo dado por Hanson-West [5], el cual es un método numéricamente estable, pues evita al máximo la cancelación numérica debida a los errores por redondeo en su procedimiento.

Sinopsis: [mediaX,varianzaX] = *medyvar*(X).

Funciones que llama:

De Matlab 5.0: for.m, if.m, size.m, zeros.m

Referencias: Hanson, R. J. [5], Stably updating mean and standard deviation of data, Communications of the ACM

•Rotapp.m

Esta función aplica una rotación al plano q_i, q_j (columnas de la matriz Q) a partir de ciertos ángulos generados por las entradas ij−ésimas de dichos vectores

Algoritmo: Si se tiene el vector $\underline{a}^t = (a_1, a_2)$, la función aplica una rotación de tal manera que se llega al vector $\underline{b}^t = (b_1, 0)$, donde $b_1 = \sqrt{a_1^2 + a_2^2}$, esto es,

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sqrt{a_1^2 + a_2^2} \\ 0 \end{pmatrix}$$

donde $c = \cos\theta$ y $s = sen\theta$, θ es el ángulo formado por el vector $\underline{a}$.

Sinopsis: [vector$X$, vector$Y$] = rotapp(x, y, c, s).

Referencias: Stewart, G. W. [16], Matrix Algorithms.

•Rotgen.m

Esta función genera la rotación del plano mediante los valores a_1 y a_2, sobrescribiendo a_1 con $\sqrt{a_1^2 + a_2^2}$ y a_2 con 0.

Algoritmo: A partir del vector $\underline{a}^t = (a_1, a_2)$ se generan el coseno y el seno para la rotación, éstos están dados por

$$c = \frac{a_1}{\sqrt{a_1^2 + a_2^2}} \quad \text{y} \quad s = \frac{a_2}{\sqrt{a_1^2 + a_2^2}},$$

posteriormente se genera la rotación mandando $a_1' = \sqrt{a_1^2 + a_2^2}$ y $a_2' = 0$.

Sinopsis: [a_1', a_2', c, s] = rotgen(a_1, a_2).

Funciones que llama:

De Matlab 5.0: abs.m, fprintf.m, if.m, size.m, sqrt.m

Referencias: Stewart, G. W. [16], Matrix Algorithms.

•**Qrdelout.m**

Esta función borra aquellas observaciones que se consideran posibles datos aberrantes en la matriz del modelo X y en el vector de respuesta $\underline{y}$. Estos datos tienen la particularidad de que sus residuales estándarizados están fuera del intervalo dado por ± 2. que es aproximadamente el intervalo de confianza al 95% para una distribución normal estándarizada.

Algoritmo: De la matriz X y el vector $\underline{y}$ se remueven aquellos renglones correspondientes a las observaciones aberrantes, quedando $\tilde{X}$ e $\tilde{\underline{y}}$. Se llama a la descomposición QR de la matriz X y en base al llamado de funciones de actualización para esta descomposición, se recalcula la nueva descomposición $\tilde{Q}\tilde{R}$ para la matriz de datos $\tilde{X}$, a partir de la descomposición original.

Sinopsis: $[\tilde{Q}, \tilde{R}, \mathrm{mat}X, \mathrm{mat}Y] = qrdelout(Q, R)$

Funciones que llama:

De Ana_Reli 2.0: qrdrmrow.m

De Matlab 5.0: for.m, if.m, load.m, size.m

Referencias: Stewart, G. W. [16], Matrix Algorithms.

•**Qrdrmrow.m**

Esta función se encarga de eliminar el último renglón de las matrices Q y R en la descomposición QR de la matriz X, y a su vez actualizarlas de tal manera que no se pierda la propiedad de que

$$\tilde{X} = \tilde{Q} \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix},$$

donde $\tilde{X}$ es la matriz del modelo, después de eliminar el último de sus renglones; esto es,

$$X = \begin{bmatrix} \tilde{X} \\ \underline{x}^t \end{bmatrix}.$$

Algoritmo: Mediante rotaciones del plano (rotaciones de Givens), se calcula una matriz P ortogonal y de rotación de tal manera que

$$X = \begin{bmatrix} \tilde{X} \\ \underline{x}^t \end{bmatrix} = Q \begin{bmatrix} R \\ 0 \\ \underline{0}^t \end{bmatrix} = (QP)\, P^t \begin{bmatrix} R \\ 0 \\ \underline{0}^t \end{bmatrix} = \begin{bmatrix} \tilde{Q} & \underline{0} \\ \underline{0}^t & 1 \end{bmatrix} \begin{bmatrix} \tilde{R} \\ 0 \\ \underline{w}^t \end{bmatrix}$$

de donde se sigue que

$$\tilde{X} = \tilde{Q} \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix}.$$

Sinopsis: $[\,\tilde{Q}\,,\,\tilde{R}\,] = \mathrm{qrdrmrow(Q,R)}$

Funciones que llama:

De Ana_Reli 2.0: rotapp.m, rotgen.m

De Matlab 5.0: for.m, size.m

Referencias: Stewart, G. W. [16], Matrix Algorithms.

A.1.2 Funciones de presentación de ventanas.

•Anareli.m

Esta función se encarga de generar la pantalla de "Presentación" del sistema Ana_Reli.M Ver. 2.0 (véase figura 5.1, pág. 99), en la cual se incluyen los botones que permiten la entrada a las opciones "Regresión", donde se realizan las funciones estadísticas y gráficas del sistema; "Demo", donde se presenta la demostración de las opciones con que cuenta el sistema; "Referencias", donde

se presentan textos de ayuda básica en el Análisis de Regresión Lineal; "Info", que presenta información general del sistema, y la opción "Salir" que finaliza con todas las aplicaciones del sistema Ana_Reli.M Ver. 2.0.

Funciones que llama:

De Ana_Reli 2.0: anareg.m, ayudtext.m, demoreg.m

De Matlab 5.0: axes.m, axis.m, cla.m, close.m, figure.m, hold.m, load.m, plot.m, polyfit m, polyval.m, save.m, set.m, text.m, uicontrol.m

•Anareg.m

Esta función se encarga de generar la pantalla de "Procedimientos" del sistema Ana_Reli.M Ver 2.0 (véase figura 5 2, pág. 101), en la cual se incluyen los menús desplegables que permiten la entrada de los datos (módulo de interacción con el usuario) y las características que tendrá el modelo, así como los botones que permiten la eliminación de datos aberrantes u "outliers", también presenta en el menú principal (véanse figuras 5.3, pag 102 y 5.4, pag. 105) las opciones estadísticas, gráficas y la ayuda de las anteriores con que cuenta el sistema

Funciones que llama:

De Ana_Reli 2.0: abrirarc.m, anareg.m, ayudtext m, calculos.m, entradat m, qrdelout.m, regcalc.m

De Matlab 5.0: axes.m, axis.m, cla m, figure m, findobj.m, fix.m, for m, gcf.m, get.m, hold m, if.m, load.m, msgbox.m, nargin.m, num2str.m, strcmp m, plot.m, polyfit.m, polyval.m, questdlg.m, save m, set.m, size m, str2mat m, switch.m, text m, uimenu m

•**Ayudtext.m**

Esta función contiene los textos que presenta la ayuda del sistema, y en combinación con la función ayudfun.m se encargan de desplegar la ventana que presenta: el nombre de la opción de la cual se solicitó la ayuda así como la descripción de ésta, incluyendo botones de cambio de página en caso de que la información sea amplia.

Funciones que llama:

De Ana_Reli 2.0: ayudfun.m

De Matlab 5.0: if.m, nargin.m, strcmp.m

•**Demoreg.m**

Esta función, junto con playdemo.m, presentan la demostración del sistema, en la cual se da a conocer al usuario las opciones estadísticas y gráficas con que cuenta el sistema. En ésta función sólo se presentan los textos que aparecen en el recuadro inferior de dicha ventana, así como los procedimientos para la presentación de las gráficas del demo.

Funciones que llama:

De Ana_Reli 2.0: playdemo.m

De Matlab 5.0: axis.m, cla.m, hist.m, hold.m, load.m, nargout.m, plot.m, polyfit.m, polyval.m, set.m, text.m, xptext.m

•**Entradat.m**

Esta función presenta una ventana desde la cual el usuario puede capturar el número de regresores del modelo de regresión lineal, el número de observaciones, la matriz del modelo X y el vector de respuesta $\underline{y}$, en esta ventana se permite grabar los datos en un archivo " *.arl ", por default el sistema los guarda en el archivo datos.rmr.

Funciones que llar ıa:

De Ana_Reli 2.0: salvarch.m

De Matlab 5.0: char.m, clear.m, close.m, errordlg.m, figure.m, findobj.m, fix.m, fprintf.m, get.m, if.m, load.m, max.m, nargin.m, save.m, size.m, strcmp m, str2num.m, uicontrol.m, warndlg.m, zeros.m

•Regcalc.m

La finalidad de ésta función es presentar en pantalla las ventanas con los resultados estadísticos y gráficos que se han obtenido al haber ajustado el modelo de regresión lineal. La presentación cuenta con el nombre de la opción y cada ventana tiene un botón de información en caso de que se desee saber un poco más de la estadística o el gráfico obtenido.

Funciones que llama:

De Ana_Reli 2.0: ayudtext.m

De Matlab 5.0: axes.m, char.m, erfinv.m, errordlg.m, figure.m, findobj.m, fix.m, for.m, get.m, hist.m, hold.m, if.m, isempty.m, length.m, load.m, msgbox.m, nargin.m, plot.m, polyfit.m, polyval.m, set.m, size.m, sqrt.m, strcmp.m, str2num m, title m, uicontrol.m, xlabel.m, ylabel.m

A.1.3 Funciones auxiliares.

•Ayudfun.m

Esta función trabaja en forma conjunta con ayudtext.m para presentar al usuario la ayuda con la cual cuenta el sistema. Esta función crea la ventana para la presentación de la ayuda colocando el nombre de la ayuda en la posición indicada, el texto correspondiente y el número de hojas con que cuenta la ayuda.

Funciones que llama:

De Matlab 5.0: eval.m, figflag.m, figure.m, for.m, get.m, if.m, nargin.m, num2str.m, set.m, watchoff.m, watchon.m

•Abrirarc.m

Esta función presenta una caja de diálogo donde se escribe el nombre del archivo que se desea abrir, después lee los datos del archivo generado por el sistema Ana_Reli.M Ver. 2.0 (extensión *.arl) y finalmente los coloca en un archivo que cuenta con el formato requerido por la función de cálculos.

Funciones que llama:

De Matlab 5.0: disp.m, fclose.m, find.m, fopen.m, for.m, fread.m, if.m, length.m, load.m, max.m, nargin.m, ones.m, size.m, strcmp.m, str2num.m, uigetfile.m, zeros.m

•Playdemo.m

Esta función se encarga de auxiliar a la función demoreg.m en la presentación de la demostración del sistema, desplegando cada una de las opciones y la información para cada una de ellas, lo puede hacer en forma manual o si el usuario lo desea en forma automática.

Funciones que llama:

De Ana_Reli 2.0: ayudtext.m

De Matlab 5.0: axes.m, disp.m, drawnow.m, eval.m, figure.m, findobj.m, gcf.m, get.m, if m, ishandle.m, length.m, nargin.m, pause.m, set.m, strcmp.m, while m

•Salvarch.m

Esta función se encarga de presentar una caja de diálogo en la cual se introduce el nombre del archivo con el que el usuario desea guardar la matriz

X del modelo y el vector de respuesta $\underline{y}$, la extensión que el sistema escribe a dichos archivos es: " .arl ".

Funciones que llama:

De Matlab 5.0: nargin.m, if.m, uiputfile.m, fopen.m, disp.m, strcmp.m, fprintf.m, save.m, fclose.m

A.2 Actualización de matrices.

Cuando se ajusta un modelo de regresión lineal existen observaciones que con frecuencia alteran las estimaciones de mínimos cuadrados de dichos ajustes, por lo que en ocasiones el analista se ve obligado a eliminar dichas observaciones; esto conllevaría a recalcular todas las estadísticas necesarias para la validación del modelo y el análisis gráfico, pero éste no es el principal problema sino el de recalcular la descomposición QR de la matriz X después de haber eliminado una o más observaciones, la cual a pesar de ser numéricamente estable resulta costosa -en cuanto a tiempo máquina-, es por eso que se han desarrollado diversos algoritmos de actualización de estas matrices a partir de la descomposición original, lo cual resulta mucho más barato.

La actualización de las estadísticas para el análisis de regresión y el análisis gráfico requiere previamente de la descomposición QR de la matriz X, entonces, al haber eliminado uno o varios renglones de la matriz X y del vector de respuesta $\underline{y}$, se obtiene una descomposición QR diferente. Lo que se debe hacer para calcular las estadísticas antes mencionadas es lo siguiente:

A partir de la descomposición QR inicial para la matriz X, se eliminan aquellas observaciones que se considere están causando problemas en el ajuste del modelo -digamos r observaciones-, llegando a una nueva descomposición $\tilde{Q}\tilde{R}$, pero ahora para la matriz $\tilde{X}$ de $(n-r)\times p$, con $n-r \geq p$ y $rgo\left(\tilde{X}\right) = p$.

Como se mencionó en la sección (1.2) el vector de parámetros estimados $\hat{\underline{\beta}}$ es encontrado al resolver el sistema de ecuaciones

$$\tilde{X}^t \tilde{X} \underline{\beta} = \tilde{X}^t \underline{y}$$

que como se vio en la sección (4.2.1) el problema anterior se reduce a la

del sistema triangular superior

$$\tilde{R}\underline{\tilde{\beta}}=\tilde{Q}_r^t\underline{\tilde{y}}$$

mediante la descomposición $\tilde{Q}\tilde{R}$ de la matriz $\tilde{X}$, obteniéndose así el nuevo vector de parámetros estimados $\underline{\hat{\beta}}$, y a partir de éste las estadísticas requeridas.

Nota: El presente trabajo sólo considera la actualización de la descomposición QR de la matriz X para el caso de eliminación de renglones, esto debido a la complejidad de cálculos numéricos necesarios para la actualización en los casos en que se agregan uno o más renglones (observaciones) o se eliminan o se agregan columnas (regresores) a la matriz de datos X. Es por ello que estos últimos casos serán tratados en trabajos futuros.

A.3 Archivos auxiliares.

Para el buen funcionamiento del sistema, éste crea archivos auxiliares en los cuales se almacenan indicadores que permiten saber el número de datos y regresores con que cuenta el modelo así como las características del mismo ya sea datos centrados o estándarizados, se genera un archivo adicional en el cual se guarda cualquier cambio realizado a la matriz de datos X y al vector de respuesta $\underline{y}$, también se generan archivos en los cuales se guardan las estadísticas que serán presentadas por los procedimientos del sistema y que serán usadas por la parte gráfica del mismo. Estos archivos auxiliares permiten al sistema no perder la información inicial del modelo de regresión lineal y evitan un posible truncamiento en alguno de sus procedimientos. Los archivos auxiliares se dividen en dos grupos: los que guardan la información que es utilizada por el Módulo de Cálculos Numéricos y que no pueden ser visualizados por el usuario y, los que guardan la información que es utilizada por el Módulo de Presentación de Estadísticas y Gráficas y que puede ser visualizada por el usuario en alguna de las opciones del sistema. Estos archivos son presentados a continuación:

1. Los utilizados por el Módulo de Cálculos Numéricos.

   (a) **Datos.rmr:** guarda la matriz original de datos X y el vector de respuesta $\underline{y}$.

(b) **Datos1.mat:** guarda los cambios realizados en la matriz de datos X y en el vector de respuesta $\underline{y}$, tales como centralización y estandarización de datos o eliminación de datos aberrantes, dichos cambios son guardados en código ascii por lo que sólo pueden ser leídos por el sistema.

(c) **Datos2.rmr:** en este archivo se guardan cuatro indicadores importantes; el primero toma valores 0 y 1, si es cero entonces el cálculo de las estadísticas se hará incluyendo todas las observaciones de la matriz X aún las que son posibles datos aberrantes, si es uno entonces se recalcularán las estadísticas sin tomar en cuenta los datos aberrantes localizados por el sistema; el segundo y tercer indicador de dicho archivo guardan el número de regresores y el número de observaciones con que cuenta el modelo de regresión y el cuarto indicador del archivo dice al sistema las características del modelo de regresión lineal a ajustar, tomando los valores 1 para el modelo normal, 2 para el modelo con regresores y variables de respuesta centralizadas y 3 para el modelo con regresores y variable de respuesta estándarizados.

(d) **Qrdex.rmr:** en este archivo se guardan las matrices Q y R en la descomposición QR inicial de la matriz de datos original X.

(e) **Qrsinout rmr:** en este archivo se guarda la descomposición $\tilde{Q}\tilde{R}$ de la matriz $\tilde{X}$ resultante de haber eliminado las observaciones que fueron consideradas como datos aberrantes.

2. Los utilizados por el Módulo de Presentación de Estadísticas y Gráficas.

(a) **Anova.rmr:** se guardan las estadísticas tales como la tabla de Análisis de Varianza, los Coeficientes de Correlación Múltiple, de Determinación, el Ajustado, y la estimación insesgada de la desviación estándar de los residuales.

(b) **Matcor.rmr:** guarda los coeficientes de correlación simple entre los regresores x_i con x_j, y entre los regresores con la variable de respuesta y.

(c) **Outliers.rmr:** guarda los *outliers* u observaciones aberrantes encontrados en el modelo, en caso de no registrarse éstos, se guarda un indicador que especifica la ausencia de ellos.

(d) **Residual.rmr:** guarda los valores de los distintos residuales, tales como los residuales comunes, los estándarizados, los estudentizados; además de los *leverages* de cada observación (h_{ii}) y las distancias de Cook (D_i), y por las dimensiones de este archivo se guarda también el vector de valores ajustados por la regresión $\hat{\underline{y}}$.

(e) **R~~umreg.rmr:** guarda la media y la varianza de cada regresor x_i y de la variable de respuesta y, su desviación estándar y los valores máximo y mínimo de cada uno de ellos; se guarda el vector de parámetros estimados $\hat{\underline{\beta}}$, la desviación estándar de dichas estimaciones y su respectivo cuantil $t-student$, los Coeficientes de Inflación de la Varianza y el estadístico Durbin-Watson para autocorrelación.

Nota: El diseño interno con que cuenta cada uno de los archivos auxiliares es reconocido en forma automática por el sistema, por lo que se garantiza la plena funcionalidad del mismo; cualquier cambio o modificación a estos archivos será responsabilidad directa del usuario.

Apéndice B

Análisis General del Proyecto.

Introducción.

En la vida diaria, usualmente se encuentran problemas o fenómenos que no son fácil de explicar si no se tienen suficientes conocimientos previos de ellos, pero se puede, a partir de ciertos datos conocidos y de experiencias pasadas, intentar formar un modelo matemático que explique de manera confiable y segura el comportamiento empírico que dichos fenómenos tendrían.

Una herramienta muy poderosa en la cual basarse es la Regresión Lineal, éste es un método que centra su atención en la asociación de variables de control o regresores -llamadas x_i-, con una variable de respuesta -llamada y-, llegando a un modelo lineal general que en su forma más simple sería

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \tag{B.1}$$

donde la matriz X ($\in \Re^{n\times p}$) tiene en sus columnas a cada una de las diferentes variables de control o regresores, evaluadas en n diferentes casos; $\underline{y}$ es el vector de los valores de respuesta observados, $\underline{\beta}$ es un vector de parámetros desconocidos y el vector $\underline{\varepsilon}$ son los errores no observables generados a partir de los experimentos realizados y los cuales se desea minimizar para obtener una "buena" estimación al vector de parámetros $\underline{\beta}$ desconocido.

Los supuestos que acompañan el planteamiento del modelo (B.1) son que $rgo(X) = p \leq n$, que los errores son independientes y que se distribuyen idénticamente con media cero y varianza constante σ^2, esto es $\underline{\varepsilon} \sim (\underline{0}, \sigma^2 I)$.

Para encontrar una estimación del vector de parámetros se procede a

resolver el siguiente problema de optimización:

$$\underset{\tilde{\beta}}{Mín} \|\underline{\varepsilon}\|_2^2 = \underset{\tilde{\beta}}{Mín} \left\|\underline{y} - X\tilde{\underline{\beta}}\right\|_2^2 \tag{B.2}$$

a esta forma de proceder se le conoce por *el método de mínimos cuadrados.* Diferenciando esta expresión se llega a que resolver el problema (B.2) es equivalente a resolver el sistema lineal de ecuaciones algebraicas

$$\left(X^t X\right)\tilde{\underline{\beta}} = X^t \underline{y} \tag{B.3}$$

conocido como *Ecuaciones Normales.* Así, si X es de rango máximo entonces el problema (B.2) tiene una única solución dada por

$$\hat{\underline{\beta}} = \left(X^t X\right)^{-1} X^t \underline{y}.$$

Pero desde un punto de vista numérico, siempre que sea posible, se trata de evitar el producto de matrices $X^t X$ en (B.3) debido a dos grandes inconvenientes: el primero, es el gasto computacional requerido para el cálculo de dicha matriz; y el segundo, es que $X^t X$ puede resultar de rango deficiente siendo que X es numéricamente de rango completo

Un método alternativo para la resolución numérica del problema (B.3), es conocido como la descomposición QR de la matriz X; el cual consiste en encontrar dos matrices de tal manera que la matriz X se exprese de la siguiente forma

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix},$$

con $Q \in \Re^{n\times n}$ matriz ortogonal, y $R \in \Re^{p\times p}$ matriz triangular superior.

Mediante esta descomposición de la matriz X, y expresando a la matriz $Q = [Q_x|Q_\perp]$, con $Q_x \in \Re^{n\times p}$, $Q_\perp \in \Re^{n\times(n-p)}$; el problema de resolver las *Ecuaciones Normales* (B.3) se reduce a la resolución del sistema triangular superior

$$R\tilde{\underline{\beta}} = Q_x^t \underline{y}, \tag{B.4}$$

y si además se cumple que $rgo(X) = rgo(R) = p$ entonces la solución a (B.4) está dada por:

$$\hat{\underline{\beta}} = R^{-1} Q_x^t \underline{y}$$

Con este procedimiento se logra conseguir dos ventajas importantes: (i) que la acumulación de errores por redondeo sea amortiguada debido a que Q es ortogonal, y (ii) que con el sistema (B 4) el número de condición sea $\kappa_2(R) = \kappa_2(X)$ en vez de $\kappa_2^2(X)$ como ocurre para las Ecuaciones Normales (B.3).

B.1 Descripción inicial del sistema.

El sistema Ana_Reli.M está diseñado para ser un sistema amigable y de fácil manejo, aún para aquellas personas que no estén muy familiarizadas con el uso de las computadoras, pues no es necesaria la edición de programas o su compilación para poder usarlo El sistema está diseñado para trabajar a base de botones y menús en donde se presentan las opciones con que cuenta el sistema, además, en cada uno de los procedimientos se presenta una breve explicación de manejo y de las opciones seleccionadas como ayuda para aquellas personas que carezcan de conocimientos profundos de los términos estadísticos y gráficos en el Análisis de Regresión Lineal.

El sistema Ana_Reli.M en su Módulo Numérico y Estadístico (véase sección B.2.3) contiene varias subdivisiones (ó submódulos) de tareas estadísticas que son presentadas en las siguientes secciones.

B.1.1 Análisis de Regresión Lineal.

Previo al cálculo de los estimadores de los parámetros del modelo de regresión, se tiene que indicar al sistema algunas cualidades que tendrá el modelo de ajuste, como son estándarizar datos, centrar datos o simplemente que el modelo a ajustar pase por el origen, estas opciones son:

Considerar Constante. Esta opción indica al sistema si hay que calcular la estimación de una constante β_0 en el modelo de regresión lineal a ajustarse o si se considerará que el modelo de regresión pase por el origen, i. e. $\beta_0 = 0$, (véase sección 1.3.1).

Centrar Datos. Esta opción sirve para que el sistema recalcule las observaciones centrándolas alrededor del cero, esto es, que cada uno de los nuevos regresores tenga media cero, i. e. $\overline{x}_i = 0, i = 1, \ldots, p-1$. Para centrar

las observaciones se utilizan las siguientes expresiones:

$$x_{ij}^{*} = x_{ij} - \overline{x}_j \text{ donde } \overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad j = 1, 2, \ldots, p-1$$

$$y_i^{*} = y_i - \overline{y} \text{ donde } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Cuando se elige esta opción, automáticamente "Considerar constante" se desactiva pues $\beta_0 = 0$, (véase sección 1 3.1).

Estándarizar Datos. Esta opción indica al sistema que las observaciones serán centradas con respecto a su media y además serán estándarizadas, esto es, la media de cada regresor será cero y cada uno de ellos tendrá varianza unitaria, i. e. $\overline{x}_i^{*} = 0$ y $s_i^{*} = 1$, $i = 1, \ldots, p-1$. Para estándarizar las observaciones se utilizan las siguientes expresiones·

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \text{ donde } \overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, \quad s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}; \quad j = 1, \ldots, p-1$$

$$y_i^{*} = \frac{y_i - \overline{y}}{s_y} \text{ donde } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}.$$

Al elegir esta opción, automáticamente se desactiva "Considerar Constante" pues $\beta_0 = 0$. (véase sección 1.3.1).

Métodos Estadísticos. Cuando se plantea un modelo de regresión lineal, generalmente los parámetros de dicho modelo son desconocidos por lo que hay que calcular estimaciones confiables a dichos parámetros, pero el problema no se detiene ahí, pues es necesario conocer la veracidad de los resultados obtenidos a partir de dichas estimaciones, es por ello que se deben hacer varios análisis estadísticos tales como pruebas de hipótesis, análisis de las observaciones usadas para la estimación de parámetros, análisis de los regresores que son usados en el estudio de la variable de respuesta, etc., y una vez verificado ésto, hacer estimaciones de la variable de respuesta (predicciones) mediante el uso del modelo ajustado.

Para tales propósitos el sistema cuenta con las siguientes opciones estadísticas.

Estimación de Parámetros. Con esta opción se calculan las estimaciones del vector de parámetros desconocidos $\underset{\sim}{\beta}$, así como sus correspondientes desviaciones estándar $\left(se\left(\hat{\beta}_i\right)\right)$ y la estimación insesgada de la varianza de los errores s^2. Esta estimación es calculada a partir de la descomposición QR de la matriz X, (secciones 1 2 y 4.2). Las expresiones para el cálculo de estas estadísticas son:

$$\hat{\underset{\sim}{\beta}} = R^{-1}Q_x^t\underset{\sim}{y} \quad \text{donde} \quad X = Q\begin{bmatrix} R \\ 0 \end{bmatrix}, \quad Q = [Q_x|Q_\perp]$$

$$s^2 = \frac{1}{n-p}\left\|Q_\perp^t\underset{\sim}{y}\right\|_2^2 \quad \text{y}$$

$$se\left(\hat{\beta}_i\right) = s\sqrt{r_{ii}^{(-1)}} \quad \text{donde} \quad \left(R^tR\right)^{-1} = \left\{r_{ij}^{(-1)}\right\}$$

Análisis de Varianza. Con esta opción se calcula la tabla de Análisis de Varianza (ANOVA) (sección 2.7) en la cual se encuentran los valores de: las Sumas de Cuadrados de los Errores (SCE), de la Regresión (SCR) y los Totales (SCT); los grados de libertad para cada una de ellas, los cuadrados medios y el Estadístico F que permite llevar a cabo la validación del modelo de regresión lineal ajustado mediante el planteamiento del juego de hipótesis:

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \qquad Vs. \qquad H_1 : \beta_i \neq 0 \text{ para algún } i = 1, \ldots, p-1$$

Colinealidad. Con esta opción se hace una revisión de la dependencia lineal que guardan entre sí los regresores, presentando algunas estadísticas que permiten determinar si este problema afecta en forma considerable en la estimación de los parámetros ajustados, (sección 2.8). Tales estadísticas son:

$\hat{\beta}_i$, las estimaciones del vector de parámetros desconocidos

r_{ij}, las correlaciones simples entre los regresores x_i con x_j y

CIV, los Coeficientes de Inflación de la Varianza para cada regresor.

Correlaciones. Esta opción presenta una matriz que en sus entradas cuenta con las correlaciones simples entre los regresores x_i con la variable de

respuesta y de tal manera que se pueda determinar el grado de asociación que tienen, deseándose que sus valores absolutos sean lo más cercanos a uno; y entre cada par de regresores x_i con x_j para tener una idea a priori de los posibles problemas de colinealidad que pudieran existir entre éstos (sección 2.2), en caso de que sus valores absolutos sean muy cercanos a uno. Las expresiones para calcular tales indicadores son:

$$r_{ij} = \frac{cov(x_i, x_j)}{\sqrt{Var(x_i) * Var(x_j)}} = \frac{cov(x_i, x_j)}{se(x_i) * se(x_j)}$$

donde:

$$cov(x_i, x_j) = \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) \quad \text{y} \quad Var(x_i) = \sum_{k=1}^{n} (x_{ki} - \overline{x}_i)^2$$

Análogamente para las correlaciones $r_{x_i y}$, entre regresores y la variable de respuesta y.

Datos Aberrantes. Esta opción presenta las observaciones que son consideradas como posibles puntos extremos (o "outliers") y que pueden afectar en la estimación de parámetros. Estos datos tienen la propiedad de que sus residuales estándarizados se salen de la región ± 2 que es del 95% de confianza aproximadamente, en una distribución Normal estándar, (sección 3.3). Los residuales estándarizados son:

$$rs_i = \frac{e_i}{s}, \quad \text{donde} \quad e_i = y_i - \hat{y}_i \quad \text{y} \quad s^2 = \frac{SCE}{n-p}$$

Estadísticas t-student. Con esta opción se calcula el valor de las estadísticas t-student para cada uno de los regresores del modelo, el cual ayuda a determinar el nivel de significancia que tiene cada uno de ellos en el modelo y su asociación con la variable de respuesta y, a mayor valor absoluto del estadístico t mayor asociación con la variable de respuesta. El valor de esta estadística sirve para una posible selección de variables llevando a cabo el juego de hipótesis

$$H_0 : \beta_i = 0 \qquad Vs \qquad H_1 : \beta_i \neq 0, \ \ i = 1, 2, \ldots, p-1,$$

la cual es muy provechosa cuando no se han registrado problemas de colinealidad en los regresores, (sección 2.4).

Predicción. Con esta opción se permite calcular una estimación futura o predicción de la variable de respuesta y en un determinado valor de $\underline{x}$, a partir del modelo de regresión lineal ajustado. Esto es:

$$y^* = \underline{\hat{\beta}}^t \ \underline{x}^* \quad \text{donde } \underline{x}^* \text{ es conocido.}$$

Residuales. Con esta opción se obtienen distintos tipos de residuales, tales como los estandarizados (rs_i), los estudentizados (rt_i), y algunas otras estadísticas como el leverage de las observaciones (h_{ii}) y las distancias de Cook (D_i), con los cuales se puede hacer un estudio detallado de las observaciones utilizadas para la estimación de los parámetros y el ajuste del modelo de regresión, (sección 3.1). Las expresiones de estas estadísticas son:

$$\begin{aligned}
rs_i &= \frac{e_i}{s}, \quad \text{donde} \quad e_i = y_i - \hat{y}_i \quad \text{y} \quad s^2 = \frac{SCE}{n-p} \\
h_{ii} &= \left(X \left(R^t R \right)^{-1} X^t \right)_{ii}, \\
rt_i &= \frac{e_i}{s\sqrt{1-h_{ii}}} \\
D_i &= \frac{rt_i^2}{p} \frac{h_{ii}}{(1-h_{ii})}
\end{aligned}$$

Métodos Gráficos. La verificación de los supuestos que acompañan el planteamiento del modelo de regresión lineal es importante, pues a partir de la veracidad de éstos se decide la confiabilidad que tienen los estimadores encontrados, un método sencillo y rápido de llevar a cabo tales verificaciones es en forma gráfica, el sistema cuenta con un módulo que contiene las gráficas más comúnmente usadas para tales verificaciones (sección 5.4.2).

Probabilidad Normal de los residuales estandarizados. El supuesto de normalidad en los residuales es el más importante para el caso en que se desean hacer inferencias estadísticas sobre el modelo de regresión ajustado. Con esta opción se presenta tal gráfica en la que, los puntos resultantes deben guardar una tendencia lineal pues por un lado se grafican los residuales estándarizados obtenidos de haber ajustado el modelo y por el otro el cuantil esperado para cada uno de los residules, si son muy parecidos tendrán la tendencia antes mencionada y el supuesto de normalidad se cumplirá, (sección 3.1.2).

Un supuesto no menos importante que el de normalidad en los residuales es el de varianza constante (homocedasticidad) en los mismos. Para dicho propósito el se tienen las siguientes opciones gráficas, (sección 3.1).

Valores estimados vs. residuales estandarizados. Con esta opción se presenta la gráfica que permite la verificación del supuesto de varianza constante, una vez ajustado el modelo de regresión. Por un lado se gráfican los valores estimados (ajustados) por la regresión $\hat{y}_i$ y por el otro sus respectivos residuales estándarizados; el supuesto se cumple si los puntos resultantes no presentan ningún patrón sistemático de alguna tendencia lineal, polinomial o creciente sino una forma completamente aleatoria alrededor del cero, (sección 3.1.1).

Valores observados vs. residuales estandarizados. Con esta gráfica se busca la misma finalidad que en la opción anterior, con la diferencia de que aquí se desea verificar si las observaciones originales presentan el supuesto de varianza constante antes de ser ajustado el modelo de regresión, (sección 3.1.1).

B.1.2 Selección de Variables.

El problema de la Selección de Variables consiste en escoger un subconjunto de regresores que expliquen de manera significativa a la variable de respuesta y, y que además, reduzcan al máximo la Suma de Cuadrados de los Errores (o Residuales) (SCE). Este módulo contará con tres diferentes métodos estadísticos y dos numéricos para la obtención de dicho modelo reducido, y que a continuación se mencionan:

Backward Elimination. Este procedimiento comienza suponiendo un modelo completo (con todos los regresores de estudio), y mediante la verificación de una prueba de hipótesis con la estadística F-parcial para cada regresor, se decide si alguno de ellos se elimina del modelo completo; este procedimiento se repite nuevamente hasta que la estadística F-parcial indique que los regresores seleccionados forman un subconjunto suficiente para la explicación de la respuesta en y, (sección 3.4.2).

Con esta opción se presenta entonces el modelo reducido, sus estadísticas básicas, la tabla ANOVA y los regresores que fueron seleccionados y que el

sistema consideró redundantes en la explicación de la variable de respuesta.

Forward Selection. Este procedimiento comienza suponiendo que, a diferencia del método Backward Elimination, no se tienen regresores significativos en el modelo y mediante la revisión de las correlaciones parciales y una prueba F-parcial, se decide cual de ellos es más significativo en la explicación de la respuesta en y, este procedimiento se repite hasta determinar el número mínimo de regresores considerados significativos para ser parte del modelo de regresión lineal reducido, (sección 3 4.1).

Con esta opción se presentan las estadísticas básicas del modelo reducido, encontrado por el sistema, la tabla ANOVA y los regresores que no fueron seleccionados y que el sistema consideró redundantes en la explicación de la respuesta en y para el modelo lineal reducido.

Estadística C_p de Mallows. Este método de selección de variables, a diferencia de los otros dos, compara directamente el Error Cuadrático Medio de los modelos reducidos obtenidos con los regresores considerados para el estudio de la explicación en y.

El criterio propuesto por Mallows para encontrar el "mejor" modelo reducido es que: aquellos modelos que presenten una estadística $C_p \approx p$, (donde p indica los regresores en el modelo lineal reducido y

$$C_p = \frac{SCE\,(p)}{s^2} - n + 2p$$

es la estadística propuesta por Mallows) son mejores en cuanto a su estimación pues reducen el Error Cuadrático Medio y la Suma de Cuadrados de los Residuales, (sección 3.4.3).

Con esta opción se presenta un cuadro con las estadísticas antes mencionadas, el número de regresores para cada modelo reducido y el valor de los parámetros estimados en los mismos.

Descomposiciones $RRQR$ y ULR. Estos son dos métodos netamente numéricos que, al igual que los métodos estadísticos anteriores, permiten encontrar el "mejor" modelo lineal reducido, (véase Bjork [2], Golub [4], Stewart [17]).

B.1.3 Análisis de Componentes Principales.

Este es un método que mediante una rotación de los ejes coordenados de las variables explicativas o regresores permite la obtención de nuevas variables llamadas Componentes Principales , que tienen la cualidad de ser independientes entre sí, y su variabilidad disminuye conforme se van seleccionando, esto es,

1. no están mútuamente correlacionadas y
2. si z_i y z_j son Componentes Principales del mismo conjunto de regresores y $j > i$, entonces $Var(z_i) > Var(z_j)$,

Lo anterior permite la reducción de dimensión con respecto a las nuevas variables. (Nota: El lector no debe confundir los términos reducir de dimensión y reducir de variables, que en el caso de las Componentes Principales se reduce la dimensión con las variables encontradas, teniendo todavía el problema de la dependencia con los regresores pues las Componentes Principales son una combinación lineal de los regresores originales, lo cual hace imposible que se puedan descartar algunos de éstos en el análisis de regresión). Las opciones estadísticas que presenta este módulo para el análisis de datos son los siguientes:

Aplicación al Análisis de Regresión. Con esta opción se realiza el análisis de regresión entre la variable de respuesta y los regresores mediante el planteamiento del siguiente modelo de regresión lineal:

$$\underline{y} = Z\underline{\alpha} + \underline{\varepsilon} \tag{B 5}$$

donde

$$Z = XU, \quad \underline{\alpha} = U^t\underline{\beta}$$

$$U^tX^tXU = Z^tZ = \Lambda \quad \text{y} \quad \Lambda = diag\,(\lambda_1, \lambda_2, \ldots, \lambda_p)$$

donde $\Lambda \in \Re^{p\times p}$ es una matriz diagonal con los eigenvalores de X^tX y $U \in \Re^{p\times p}$ es una matriz ortogonal que en sus columnas tiene los eigenvectores asociados con $\lambda_1, \lambda_2, \ldots, \lambda_p$. La descomposición espectral $X^tX = U\Lambda U^t$ se calcula via la Descomposición en Valores Singulares de $X = U\Delta V^t$, pues no es recomendable el cálculo explícito de X^tX como ya mencionó antes. Las columnas de $Z = [\underline{z}_1|\underline{z}_2|\cdots|\underline{z}_p]$, definen un conjunto de regresores ortogonales y que están asociados con las Componentes Principales.

El estimador de mínimos cuadrados para el vector de parámetros desconocidos $\underline{\alpha}$ en el modelo dado por B.5 es:

$$\hat{\underline{\alpha}} = \left(Z'Z\right)^{-1} Z'\underline{y} = \Lambda^{-1} Z'\underline{y},$$

y la matriz de varianza-covarianza para el estimador de $\underline{\alpha}$ es:

$$Var\left(\hat{\underline{\alpha}}\right) = \sigma^2 \left(Z'Z\right)^{-1} = \sigma^2 \Lambda^{-1} \quad \text{esto si} \quad Var\left(\underline{y}\right) = \sigma^2 I.$$

Analizando los eigenvalores se tiene que si todos los λ_j son igual a la unidad, los regresores originales (x_i) son ortogonales, mientras que si un λ_j es exactamente igual a cero, implica que existe una relación perfectamente lineal entre los regresores originales. Uno o más de los λ_j cercanos a cero implica que la colinealidad está presente.

Hay que notar que la matriz de varianza-covarianza de los coeficientes de regresión $\hat{\underline{\beta}}$ es:

$$Var\left(\hat{\underline{\beta}}\right) = Var\left(U\,\hat{\underline{\alpha}}\right) = UVar\left(\hat{\underline{\alpha}}\right)U' = \sigma^2 U\Lambda^{-1}U^t = \sigma^2\left(X^tX\right)^{-1}.$$

Correlación con Componentes Principales. Con esta opción se presenta la matriz que tiene en sus entradas los coeficientes de correlación simple entre cada uno de los regresores originales (x_i) y cada una de las Componentes Principales (z_j), para de esta manera conocer cuáles de los regresores originales están mayormente asociados con las Componentes. Esta matriz tiene la siguiente forma:

	z_1	z_2	$\cdots$	z_p
x_1	$r_{x_1z_1}$	$r_{x_1z_2}$	$\cdots$	$r_{x_1z_p}$
x_2	$r_{x_2z_1}$	$r_{x_2z_2}$	$\cdots$	$r_{x_2z_p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_p	$r_{x_pz_1}$	$r_{x_pz_2}$	$\cdots$	$r_{x_pz_p}$

de donde se tiene que:

$$r_{x_iz_j} = \frac{cov\left(x_i, z_j\right)}{\sqrt{Var\left(x_i\right)Var\left(z_j\right)}} \tag{B.6}$$

Para obtener las correlaciones entre los regresores x_i y las componentes z_j, recuérdese que:

$Var(\underline{x}) = \Sigma \quad \{\sigma_{ij}\}$, matriz de varianzas-covarianzas de $\underline{x}$,

$Var(\underline{z}) = \Lambda = diag(\lambda_1, \ldots, \lambda_p)$, matriz de varianzas-covarianzas de $\underline{z}$ y,

$Cov(\underline{x}, \underline{z}) = U\Lambda$, entonces $cov(x_i, z_j) = u_{ij}\lambda_j$.

Por lo que la expresión dada en (B.6) se convierte en:

$$r_{x_i z_j} = \frac{u_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}} = \frac{u_{ij}\sqrt{\lambda_j}}{\sqrt{\sigma_{ii}}}$$

Eigenvalores. Con esta opción se calculan los eigenvalores asociados a la matriz $X'X$, donde cada eigenvalor representa la variabilidad que es explicada por la correspondiente Componente Principal, también se presentan el porcentaje de variabilidad acumulada, para determinar de esta manera cuántas Componentes Principales son necesarias en la explicación del análisis deseado. Ejemplo:

# Eigenvalor	Eigenvalores	% de Variabilidad	% Var. Acumulada.
1	λ_1	$Var(\lambda_1)$	$Acum(\lambda_1)$
2	λ_2	$Var(\lambda_2)$	$Acum(\lambda_2)$
3	λ_3	$Var(\lambda_3)$	$Acum(\lambda_3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
p	λ_p	$Var(\lambda_p)$	100%

donde

$$Var(z_j) = \lambda_j, \quad j = 1, \ldots, p; \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$$

$$Var(\lambda_j) = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} * 100 \quad \text{y} \quad Acum(\lambda_j) = \frac{\sum_{k=1}^{j} \lambda_k}{\sum_{i=1}^{p} \lambda_i} * 100$$

B.1.4 Regresión Ridge.

Es un método alternativo para la estimación del vector de parámetros $\underline{\beta}$ en el modelo:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

bajo presencia de colinealidad; estos estimadores carecen de la cualidad de ser insesgados como los estimadores Mínimo Cuadrados, pero que a diferencia de éstos, poseen la ventaja de tener una menor variablidad y menor Error Cuadrático Medio lo cual los hace más confiables y precisos.

El razonamiento es la introducción de una constante $k \geq 0$ en el sistema de Ecuaciones Normales de la siguiente manera:

$$\left(X^tX + kI\right) \underset{\sim}{\hat{\beta}}_R = X^t y$$

donde la solución a este sistema de ecuaciones lineales algebráicas está determinada por:

$$\underset{\sim}{\hat{\beta}}_R = \left(X^tX + kI\right)^{-1} X^t y, \tag{B.7}$$

siendo $\underset{\sim}{\hat{\beta}}_R$ el estimador ridge de mínimos cuadrados. La dificultad de aplicar este método es que el valor óptimo de k varía de una aplicación a otra y es desconocido. Las opciones estadísticas de este módulo son las siguientes.

Estimación ridge de parámetros. La finalidad de esta opción es el cálculo de la estimación ridge de los parámetros desconocidos $\underline{\beta}$, mediante la resolución del sistema de ecuaciones dado en (B.7), encontrando el valor óptimo del parámetro k que estabilice al máximo las estimaciones. Para tal propósito se tienen las siguientes opciones de determinación (elección) del parámetro ridge k

1. La Traza Ridge.
2. Validación Cruzada Generalizada.
3. Principio de discrepancia
4. Principio de Quasi-optimalidad.
5. Principio de la L-curva.

Predicción. Con esta opción se permite calcular una estimación futura o predicción de la variable de respuesta y en un determinado valor de $\underline{x}$, a partir del modelo de regresión lineal ajustado mediante el estimador ridge (B.7) de mínimos cuadrados. Esto es:

$$y^{\star} = \underline{\hat{\beta}}_{lr}^{t} \underline{x}^{\star} \quad \text{donde } \underline{x}^{\star} \text{ es conocido.}$$

En la siguiente sección se hace una descripción general de lo que es el proyecto Ana_Reli.M, en su concepción original, la forma de cómo está estructurado y las partes que comprenden cada una de sus distintas ramas. Finalmente se especifican los módulos que abarca esta primera versión de dicho proyecto.

B.2 Esquematización general del sistema.

En la figura B.1 se muestra la forma de cómo están estructurados los módulos que comprende el sistema Ana_Reli.M, éste cuenta con 6 módulos principales; cada uno de ellos tiene una finalidad bien definida en el sistema, por lo que, alguna modificación en alguno de los módulos no afecta en absoluto las tareas realizadas por los modulos restantes. A continuación se presenta una breve descripción de las tareas realizadas por cada uno de estos módulos:

B.2.1 Módulo Rector.

Este módulo se encarga de controlar, monitorear, y dirigir todas las tareas asignadas y procedimientos que son realizados en los demás módulos, siendo de esta manera el más importante de los módulos mostrados en el esquema de la figura B.1.

B.2.2 Entrada de Datos.

Este módulo tiene como tarea asignada el cargar la base de datos al sistema Ana_Reli.M para posteriormente ser procesados por los siguientes módulos, las opciones de captura son:

1. Entrada directa desde el Teclado.
2. Entrada desde un Archivo.

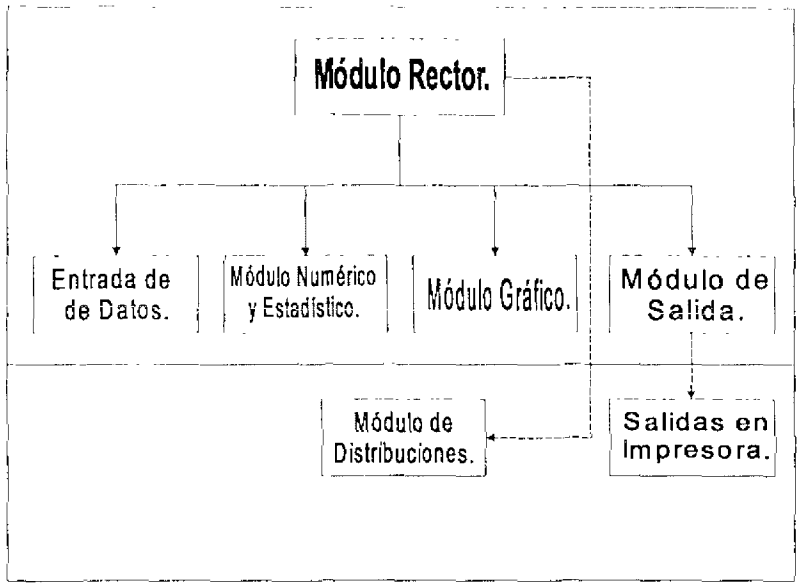


Figura B.1: Esquema General del Proyecto ANA_RELI.M

3. Entrada generada a partir de funciones y/o modelos de regresión preestablecidos por el usuario.

B.2.3 Módulo Numérico y Estadístico.

En este módulo son realizados todos los procedimientos numéricos necesarios para proceder al cálculo de las estadísticas requeridas en el análisis de regresión; tales procedimientos son: la descomposición QR de la matriz X, el cálculo de medias y varianzas muestrales, la centralización y estandarización de las columnas de la matriz X, la actualización de las matrices Q y R al eliminar observaciones aberrantes y en la selección de variables, entre otros; de igual forma se realizan los procedimientos estadísticos, una vez calculados los anteriores, siendo éstos, el cálculo del vector de parámetros β desconocido, la desviación estándar de cada estimación, el cuantil t para cada regresor, la estimación insesgada de la desviación estándar de los residuales, las sumas

de cuadrados para la construcción de la tabla de Análisis de Varianza, los residuales, observaciones aberrantes, etc.

El módulo estadístico, a su vez, tiene varias subdivisiones de aplicación en las que se realizan tareas distintas pero, sobre la misma base de datos Estas subdivisiones (o submódulos) son:

1. Análisis de Regresión Lineal.
2. Selección de Variables.
3. Análisis de Componentes Principales.
4. Regresión Ridge.

Cada uno de ellos es independiente en cuanto a los procedimientos que realiza por lo que, la alteración o modificación de alguno de ellos no influye en los procedimientos realizados por los demás submódulos.

B.2.4 Módulo Gráfico.

Este módulo tiene como finalidad la presentación de salidas gráficas en pantalla, para llevar a cabo la verificación de los supuestos del modelo como homocedasticidad e independencia en los residuales y distribución normal de los mismos, una vez calculadas las estadísticas –tales como residuales, residuales estándarizados, valores $\hat{y}_i$ estimados, cuantiles de la distribución normal estándarizada, etc. , necesarias para cada uno de los gráficos presentados Este módulo requiere de los resultados previamente obtenidos por el Módulo Numérico y Estadístico.

B.2.5 Módulo de Salidas.

Este módulo tiene como tareas asignadas la salida de resultados estadísticos y gráficos en distintas modalidades, éstas pueden ser:

1. Salida en pantalla mediante la presentación de ventanas.
2. Salida a un archivo.
3. Salida a la impresora.

B.2.6 Módulo de Distribuciones.

En este módulo se presentan varias distribuciones de densidad de probabilidad frecuentemente utilizadas en el análisis estadístico de datos. Tales distribuciones son: F, *t student*, *ji cuadrada y Normal*. La finalidad de este módulo es que el usuario introduzca un cierto nivel de significancia (α) para que de esta manera el sistema pueda calcular intervalos de confianza para los parámetros β_i desconocidos, realice pruebas de hipótesis una vez encontrado el cuantil correspondiente y, tome decisiones sobre la validación del modelo general ajustado, así como calcular la probabilidad a partir de la cual una hipótesis nula es rechazada.

Los 6 módulos anteriores cuentan con la cualidad de estar diseñados de tal manera que se puedan modificar en versiones futuras sin necesidad de tener que cambiar la estructura inicial del sistema, ésto, debido a que cada uno de los módulos es independiente de los restantes en cuanto a su configuración interna y algunos de ellos sólo necesitan de las salidas numéricas de los otros módulos.

B.3 Lo que no se alcanzó en esta versión.

El sistema Ana_Reli.M Ver. 2.0 fue creado bajo ambiente Windows en Matlab 5.0 a base de ventanas con botones y menús que lo hacen un sistema amigable, de fácil manejo y rápido de entender; así como confiable en cuanto a los procedimientos numéricos que realiza pues varias de las funciones utilizadas por el sistema son tomadas de Matlab el cual es un paquete con prestigio internacional en el área del Análisis Numérico.

Si además se cumplen las condiciones de que: $X \in \Re^{n\times p}$, $n > p$, $rgo(X) = p$, $\kappa_2(X)$ el número de condición de X no es muy grande, y además que se cumpla la hipótesis de ángulo agudo (véase sección 4.2); entonces se garantiza la plena confiabilidad y precisión de los resultados obtenidos por el sistema Ana_Reli.M Ver 2.0.

Una ventaja de usar algunas de las funciones de Matlab es que si la condición $rgo(X) = p$ no se cumple, se tiene la absoluta confianza de que los procedimientos numéricos y/o estadísticos realizados por el sistema no se verán truncados, debido a que Matlab 5.0 usa el estándar de la **IEEE.** El estándar especifica que todas las operaciones aritméticas son ejecutadas como si se calcularan en precisión infinita y después redondeadas de acuerdo

a uno de los cuatro "modos" (véase Higham, [6])

La aritmética **IEEE** es un sistema cerrado: cada operación aritmética produce un resultado, sea matemáticamente esperado o no, y cada operación genera una señal. Los resultados por default se muestran en la siguiente tabla.

Tipo de excepción	Ejemplo	Resultado por default
Operación inválida	$0/0$, $0 \times \infty$, $\sqrt{-1}$	`NaN` (Not a Number)
Overflow		$\pm\infty$
División por cero	(constante finita) /0	$\pm\infty$
Underflow		Números subnormales
Inexacto	siempre que $fl(x \text{ op } y) \neq x \text{ op } y$	Resultado bien redondeado

La respuesta por default para una excepción es colocar un banderín y continuar con los procedimientos.

Un `NaN` es generado por operaciones tales como $0/0$, $0\times\infty$, ∞/∞, $(+\infty)+(-\infty)$, y $\sqrt{-1}$. El uso de un `NaN` denota datos no inicializados o falta de los mismos. Las operaciones aritméticas que envuelven un `NaN` regresan un `NaN` como la respuesta. Lo anterior es importante en el desarrollo de los cálculos numéricos y/o estadísticos pues permite al sistema continuar con los procedimientos antes mencionados.

Algunas desventajas. Si bien el sistema Ana Reli.M Ver. 2.0 cuenta con varias ventajas tanto en el aspecto numérico, así como en la presentación de ventanas y facilidad de manejo; también tiene algunos puntos, mencionados en la sección B.2, que no fueron completamente realizados debido a la complejidad numérica y estadística con que cuentan, y que quedan asignados a futuras versiones del presente trabajo; tales puntos son:

1. En el módulo "Entrada de Datos" sólo fueron realizados los puntos correspondientes a:

   (a) entrada de datos a partir del teclado y

   (b) entrada de datos desde un archivo,

   quedando pendiente el correspondiente a entrada de datos a partir de funciones y/o modelos predefinidos.

2. El "Módulo Numérico y Estadístico" en esta versión, sólo comprende la parte básica del Análisis de Regresión Lineal, basada en la descomposición QR, quedando fuera del presente trabajo los submódulos:

   (a) Selección de Variables,

   (b) Análisis de Componentes Principales y

   (c) Regresión Ridge.

   Los requerimientos numéricos (descomposiciones $RRQR, URL$ y SVD) y estadísticos de estos submódulos son complejos, por lo que serán anexados en futuras versiones de este sistema.

3. El "Módulo de Salidas" sólo comprende la parte de Salida en pantalla mediante la presentación de ventanas, quedando pendientes los submódulos:

   (a) Salida a un archivo y

   (b) Salida a la impresora.

4. El "Módulo de Distribuciones" no fue trabajado en esta versión del sistema por lo que el módulo completo quedá pendiente para futuras versiones.

La razón por la cual únicamente se trabajó con el módulo de Análisis de Regresión Lineal es porque la herramienta que proporciona la descomposición QR de la matriz X permite realizar los procedimientos numéricos de manera confiable, pues reduce al máximo los errores por redondeo, al tener que la matriz Q es ortogonal y preserva normas. Para los restantes módulos es necesario el uso de otros métodos numéricos, tales como la descomposición $RRQR$ de la matriz X, la descomposición URL, la descomposición en valores singulares (DVS) de la matriz X, entre otros.

Si bien, la esquematización del proyecto general Ana_Reli.M presenta un sistema computacional robusto, la primera parte de este proyecto –enfocado únicamente al Análisis de Regresión Lineal– muestra en forma tangible que es un proyecto que puede llevarse a cabo hasta sus partes más sencillas. La definición y construcción de cada módulo en forma independiente permite, de esta manera, anexar otros módulos y procedimientos en forma sencilla, sin necesidad de alterar la configuración inicial del mismo.

Apéndice C

Ejemplos estadísticos y gráficos obtenidos con el sistema ANA_RELI.M VER 2.0

Introducción.

En este apéndice se hace una presentación de resultados estadísticos y gráficos obtenidos como salidas del sistema ANA_RELI.M VER 2.0. Los ejercicios usados para ilustrar tales propósitos fueron obtenidos de textos estadísticos que gozan de renombre en las áreas del Análisis de Regresión y el Análisis Numérico; tales como Chatterjee [3], Weisberg [20], entre otros. Los métodos numéricos usados por el sistema garantizan la plena confiabilidad y una alta precisión de los resultados numéricos y estadísticos obtenidos por el mismo pues, los resultados estadísticos mostrados a continuación fueron comparados con los presentados en los libros de texto, mencionados arriba, llegando así a conclusiones satisfactorias.

C.1 Ejemplos.

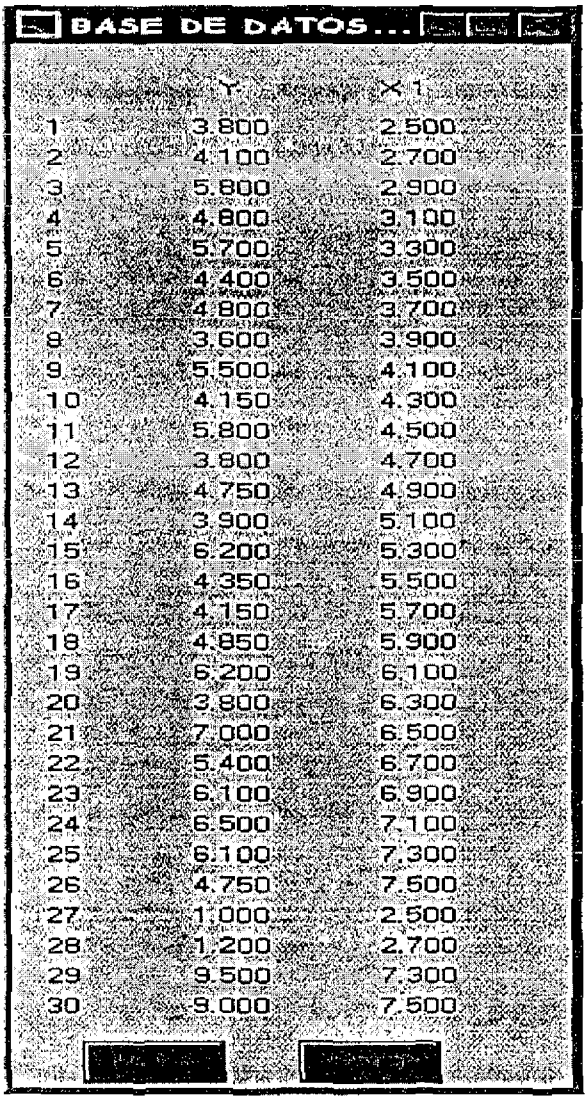
El ejemplo mostrado a continuación fue obtenido de Chatterjee [3], y sirve para ilustrar el procedimiento que realiza el sistema ANA_RELM VER. 2.0 en la eliminación de datos aberrantes, recordando que la característica de estos datos es el tener un residual estándarizado fuera del intervalo $(-2, 2)$, por lo que son consideradas observaciones problemáticas para el modelo que pueden afectar en forma drástica la estimación de los parámetros y, por consiguiente, la veracidad de los resultados estadísticos obtenidos.

Nota: El procedimiento de eliminación de datos aberrantes para el siguiente ejemplo se repitió varias veces, hasta que el sistema no encontró indicios de presencia de datos aberrantes en las observaciones.

Ejemplo C.1.1 (Chatterjee, pag. 20) *El éxito de un programa en la televisión comercial está determinado en parte por un sistema de rating que es una prueba para medir la habilidad de los programas para atraer y tener televidentes. En términos reales, los puntos del rating generan interés al patrocinador y en que brinda renta a la estación. Un director de estación a cargo de los nuevos rating de programas fue contratado para un estudio en el que se desean identificar los factores que afectan a los rating. Además de las variables más comunes, tales como formato, efectos especiales, y el recurso de sujetos para nuevos repartos, se sugirió que hubiera una continuación de los programas, precedido inmediatamente por el noticiero. El rating del noticiero fue parcialmente dependiente del rating del programa que lo precedía. Para cuantificar este efecto, una muestra aleatoria de ratings previos fue tomada a través de regiones para varios puntos en el tiempo, durante los pasados 2 años. Los datos consisten de observaciones en una variable denotado como "y", que es el rating del noticiero; y una segunda variable llamada "x", representando el rating del programa principal. Los datos son mostrados en la figura C.1. El rating es un índice que está entre 1 y 10.*

x = *Rating del programa principal.*

y = *Rating del programa de noticias.*



	Y	X1
1	3.800	2.500
2	4.100	2.700
3	5.800	2.900
4	4.800	3.100
5	5.700	3.300
6	4.400	3.500
7	4.800	3.700
8	3.600	3.900
9	5.500	4.100
10	4.150	4.300
11	5.800	4.500
12	3.800	4.700
13	4.750	4.900
14	3.900	5.100
15	6.200	5.300
16	4.350	5.500
17	4.150	5.700
18	4.850	5.900
19	6.200	6.100
20	3.800	6.300
21	7.000	6.500
22	5.400	6.700
23	6.100	6.900
24	6.500	7.100
25	6.100	7.300
26	4.750	7.500
27	1.000	2.500
28	1.200	2.700
29	9.500	7.300
30	9.000	7.500

Figura C.1: Datos del ejemplo de Rating para programas de televisión.

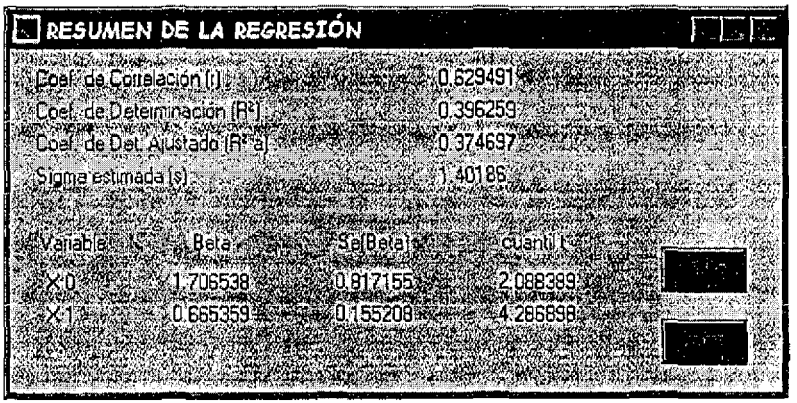


Figura C.2: Resumen de la primera Regresión para los datos del Rating.

Modelo ajustado. $Y = 1.7065 + 0\ 6653 X_1$

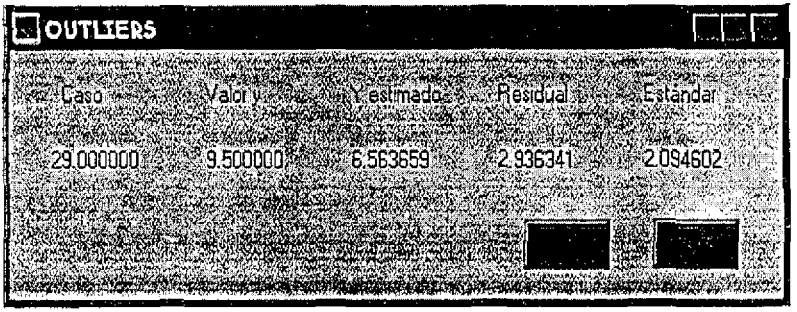
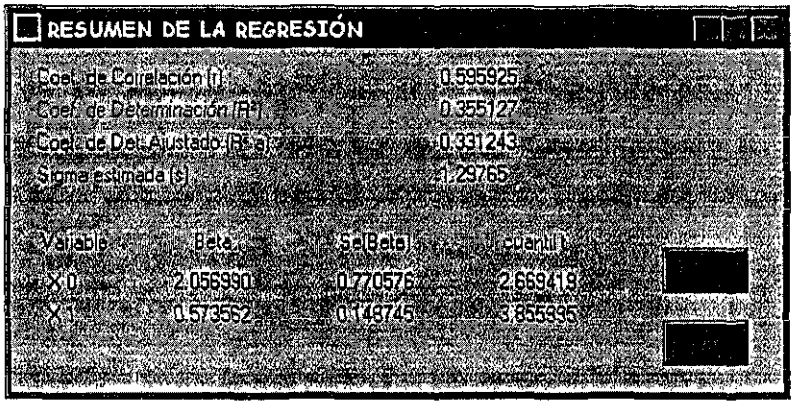


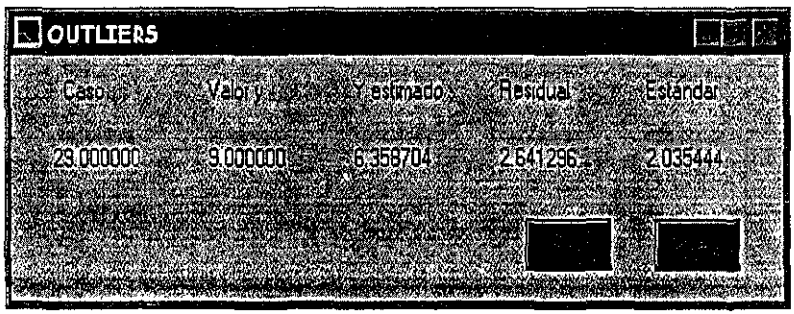
Figura C.3 Detección de Datos Aberrantes ("outliers").

En esta primera regresión el dato 29 resulta ser un punto extremo del conjunto de datos. La decisión de eliminarlo del conjunto de datos queda a elección del analista.



Figura~C 4: Resumen obtenido de la segunda Regresión, después de haber eliminado la observación 29 del conjunto de datos.

Modelo ajustado: $Y = 2.0569 + 0.5735X_1$
Nótese el cambio en los parámetros estimados después de eliminar la observación 29.



Figura~C.5: Dato Aberrante ("outlier") resultante después de eliminar la observación 29 del conjunto de datos.

En esta segunda regresión la observación 29 resultó ser un punto extremo

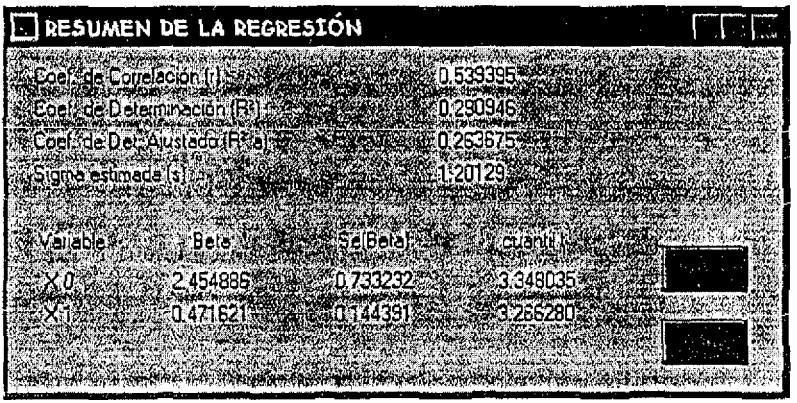


Figura C.6: Resumen obtenido de la tercera Regresión al eliminar la observación 29, resultante después de la primera eliminación de datos.

Modelo ajustado: $Y = 2.4548 + 0.4716X_1$.
Nótese el cambio en los parámetros estimados, después de esta segunda eliminación de datos.

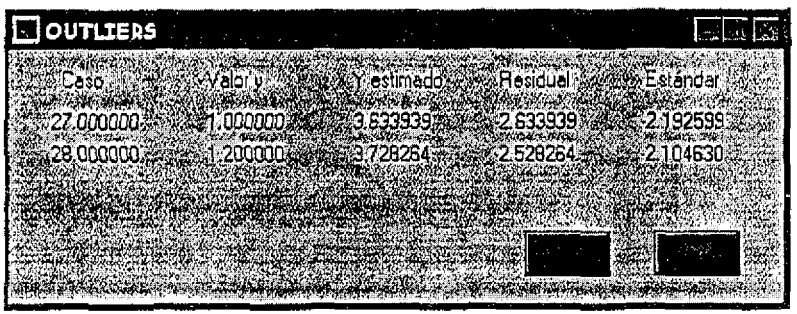


Figura C.7: Datos Aberrantes ("outliers") que aparecieron al eliminar la observación 29, después de la primera eliminación de datos.

Ahora las observaciones 27 y 28 son puntos extremos, después de haber hecho dos eliminaciones de datos.

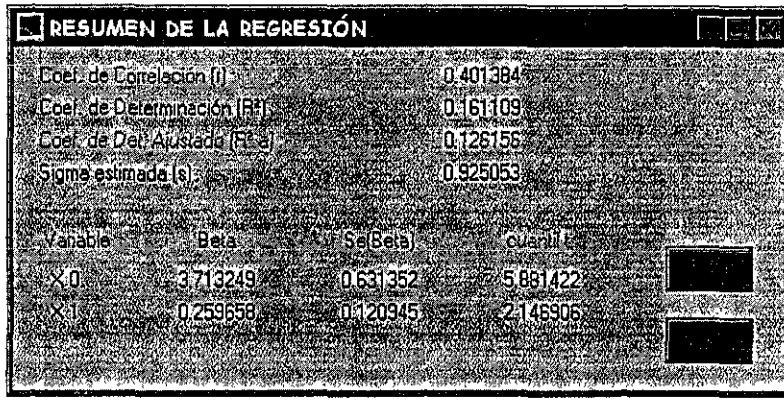


Figura C 8: Resumen obtenido de la cuarta Regresión después de eliminar las observaciones 27 y 28.

Modelo ajustado. $Y = 3.7132 + 0.2596X_1$.

Nota: Los cambios que hubo en los parámetros estimados se pueden considerar significativos, pues del primer modelo a éste último existe una gran diferencia; aunque, cabe mencionar que el coeficiente de determinación para el modelo sin datos aberrantes es de $R^2 = 0.1611$, *siendo este último menor al obtenido en la primera regresión* ($R^2 = 0.3962$), *incluyendo todas las observaciones, por lo que se pueden considerar como datos importantes en el ajuste del modelo de regresión.*

El procedimiento de eliminar datos aberrantes con el sistema ANA_RELIM VER. 2.0, dejó al final 26 observaciones para el estudio del fenómeno. Las observaciones eliminadas del conjunto de datos original fueron la 29, 30, 27 y 28. El analista puede decidir en un momento dado si desea eliminar absolutamente todos los posibles datos aberrantes o si prefiere ajustar el modelo de regresión incluyéndolos.

El siguiente ejemplo ha sido elegido para mostrar los efectos de la colinealidad en la inferencia estadística. Este ejemplo muestra la ambigüedad que quizá resulte cuando se intenta identificar importantes variables explicativas de entre un conjunto de regresores linealmente dependientes.

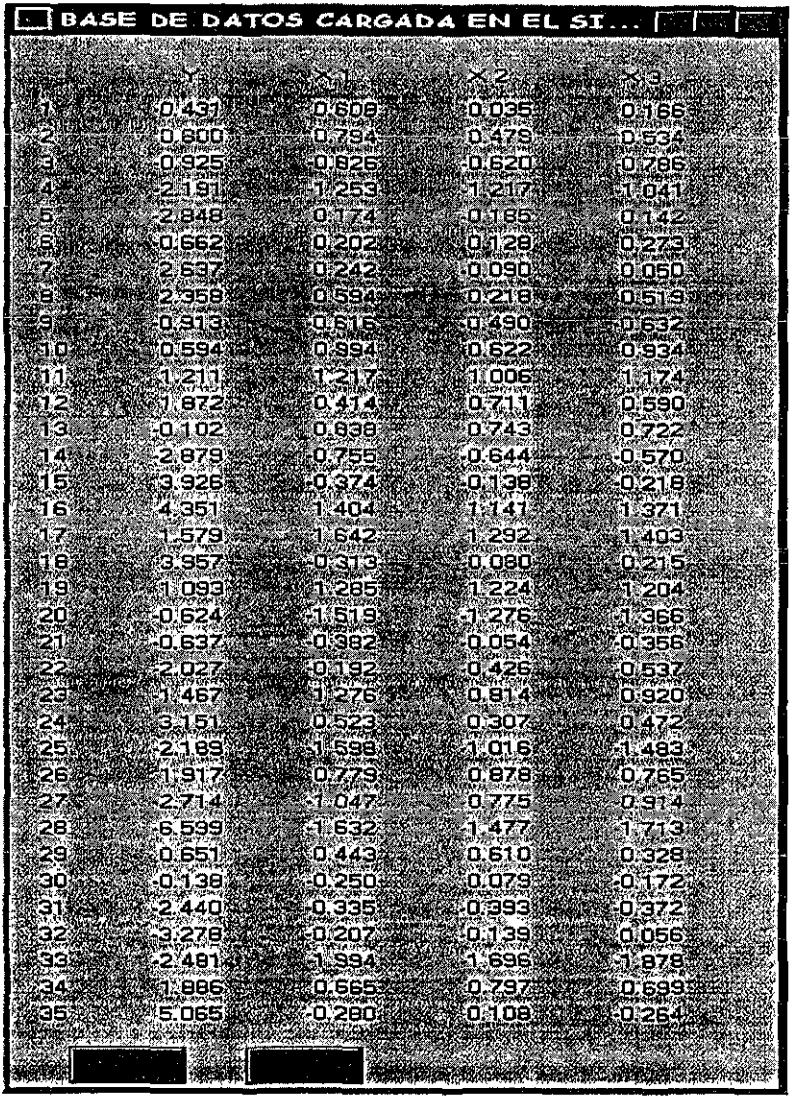
Ejemplo C.1.2 (Chatterjee, pag. 144) *En conjunción con la Comisión de Derechos Civiles en 1964 el Congreso de los Estados Unidos ordenó un examen de estudio ". concerniente a la falta de disponibilidad de oportunidades para igual educación de los individuos por razones de raza, color, religión o nacionalidad en instituciones públicas...". Los datos fueron colectados de una sección de escuelas a través del país. Además del resumen estadístico reportado en variables tales como nivel de logros del estudiante y facilidades de la escuela, el análisis de regresión fue usado para intentar establecer los factores que son más importantes en la determinación de los logros. Los datos para este ejemplo consisten de mediciones tomadas entre 1965 y1970 en escuelas seleccionadas aleatoriamente. Los datos consisten de variables que miden los logros de un estudiante, facilidades de la escuela, y profesorado calificado. El objetivo es evaluar el efecto de la escuela en los logros. Se asume que un índice aceptable ha sido desarrollado para medir estos aspectos del ambiente escolar que sería esperado para afectar los logros. El índice incluye evaluaciones de las áreas físicas, material de enseñanza, programas especiales, entrenamiento y motivación de la facultad. Los logros pueden ser medidos al usar un índice construido de puntajes de pruebas normalizadas. Hay también otras variables que pueden afectar la relación entre la escuela y los logros El mejoramiento de los estudiantes puede ser afectado por el ambiente familiar y la influencia de sus compañeros de escuela. Estas variables deben ser explicadas en el análisis antes de que el efecto de la escuela pueda ser evaluado. Se asume que los índices han sido construidos para estas variables que son satisfactorias para los propósitos del estudio.*

x_1 = *Influencia del ambiente familiar (índice normalizado).*

x_2 = *Influencia de los compañeros de escuela (índice normalizado).*

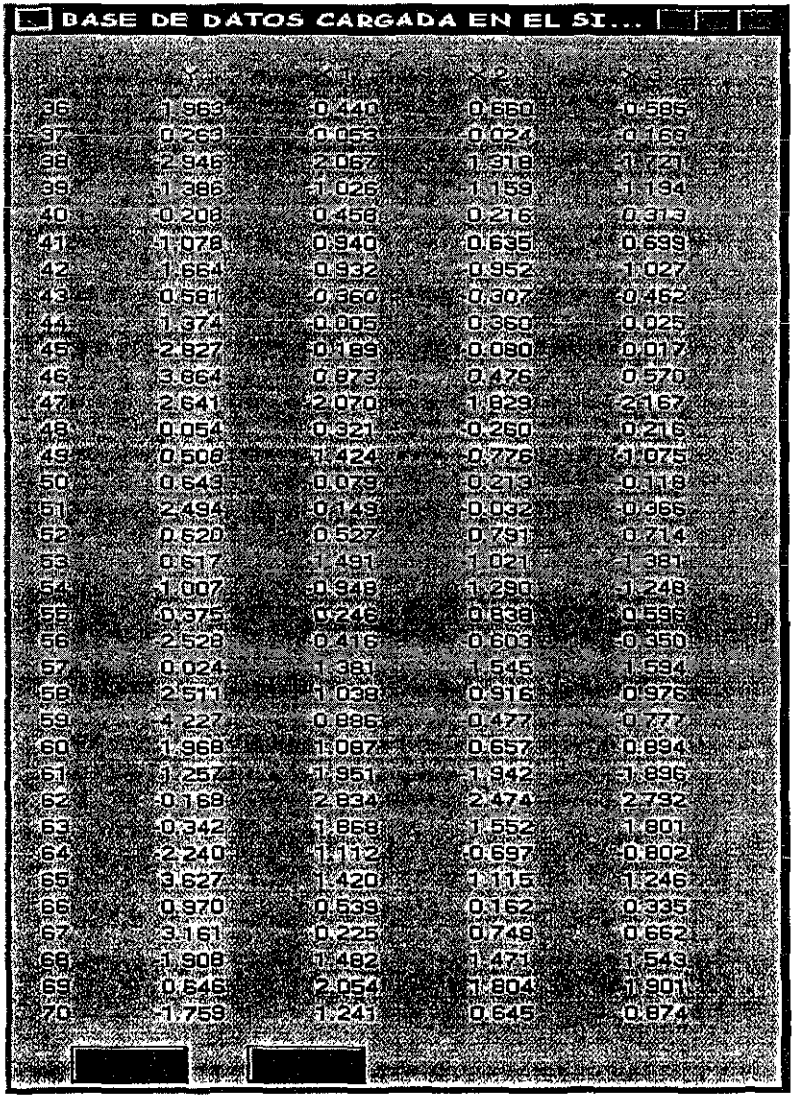
x_3 = *Influencia de la escuela (índice normalizado).*

y = *Logros del estudiante en la escuela (índice normalizado).*


BASE DE DATOS CARGADA EN EL SI...

	Y	X1	X2	X3
1	0.431	0.608	0.035	0.166
2	0.800	0.794	0.478	0.534
3	-0.925	-0.826	-0.620	-0.786
4	-2.191	-1.253	-1.217	-1.041
5	-2.848	0.174	-0.185	0.142
6	-0.662	0.202	0.128	0.273
7	2.637	0.242	-0.090	0.060
8	2.358	0.594	0.218	0.513
9	-0.813	-0.616	-0.490	-0.632
10	0.594	0.894	0.622	0.934
11	1.211	1.217	1.006	1.174
12	1.872	0.414	0.711	0.590
13	-0.102	0.838	0.743	0.722
14	-2.879	-0.755	-0.644	-0.570
15	3.926	-0.374	-0.138	-0.218
16	4.351	1.404	1.141	1.371
17	1.579	1.642	1.292	1.403
18	3.957	-0.313	-0.080	-0.215
19	1.093	1.285	1.224	1.204
20	-0.624	-1.519	-1.276	-1.366
21	-0.637	-0.982	-0.054	-0.356
22	-2.027	-0.192	-0.426	-0.537
23	1.467	1.276	0.814	0.920
24	3.151	0.523	0.307	0.472
25	-2.189	-1.598	-1.016	-1.483
26	1.917	0.779	0.878	0.765
27	-2.714	-1.047	0.775	-0.914
28	-6.599	-1.632	-1.477	-1.713
29	0.651	0.443	0.610	0.328
30	-0.138	-0.250	0.079	-0.172
31	-2.440	-0.335	-0.393	-0.372
32	-3.278	-0.207	-0.139	-0.056
33	-2.481	-1.994	-1.696	-1.878
34	1.886	0.665	0.797	0.699
35	5.065	-0.280	0.108	-0.264

Figura C.9: Datos del ejemplo que estudia los logros de los estudiantes en la escuela.



36	1.963	0.440	0.660	0.585
37	0.269	0.053	0.024	0.168
38	2.946	2.067	1.918	1.721
39	1.386	1.026	1.159	1.194
40	0.208	0.458	0.216	0.313
41	1.078	0.940	0.695	0.699
42	1.664	0.932	0.952	1.027
43	0.581	0.360	0.307	0.462
44	1.374	0.005	0.360	0.025
45	2.827	0.189	0.080	0.017
46	3.664	0.873	0.476	0.570
47	2.641	2.070	1.829	2.167
48	0.054	0.321	0.260	0.216
49	0.508	1.424	0.776	1.075
50	0.643	0.079	0.213	0.118
51	2.494	0.143	0.032	0.366
52	0.620	0.527	0.791	0.714
53	0.517	1.491	1.021	1.381
54	1.007	0.948	1.290	1.248
55	0.375	0.246	0.838	0.596
56	2.528	0.416	0.603	0.350
57	0.024	1.381	1.545	1.594
58	2.511	1.038	0.916	0.976
59	4.227	0.886	0.477	0.777
60	1.968	1.087	0.657	0.894
61	1.257	1.951	1.942	1.896
62	0.168	2.834	2.474	2.792
63	0.342	1.868	1.552	1.801
64	2.240	1.112	0.697	0.802
65	3.627	1.420	1.115	1.246
66	0.970	0.539	0.162	0.335
67	3.161	0.225	0.749	0.662
68	1.908	1.482	1.471	1.543
69	0.646	2.054	1.804	1.901
70	1.759	1.241	0.645	0.874

Figura C.10: Datos del ejemplo que estudia los logros de los estudiantes en la escuela (Continuación).

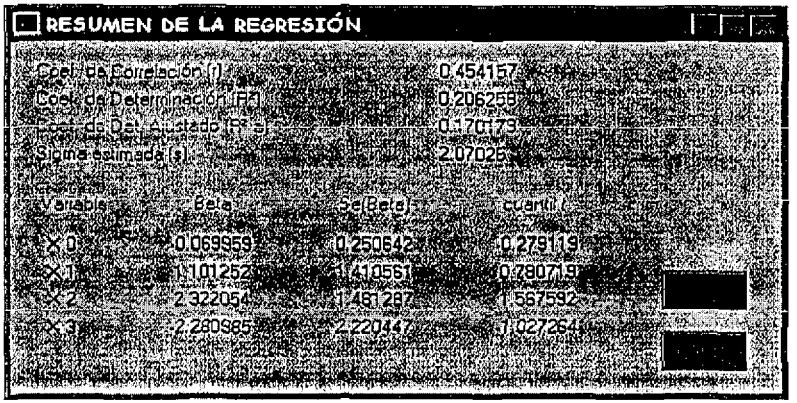


Figura C.11: Resumen de la Regresión.

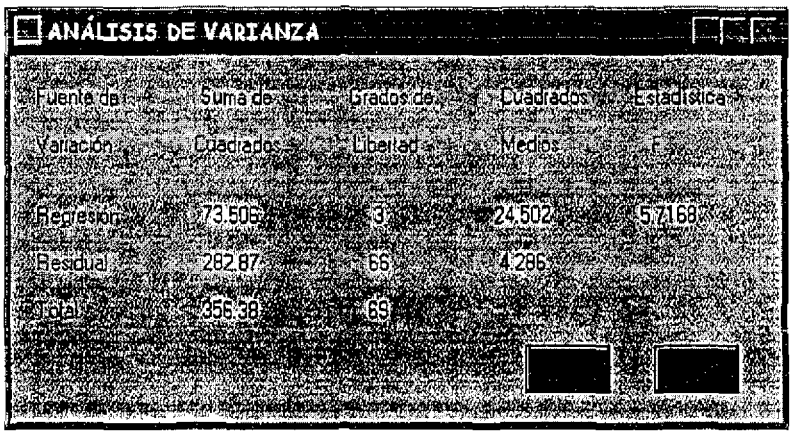


Figura C.12: Tabla de Análisis de Varianza.

Considerando un $\alpha = .01$ *el cuantil* F *es* $F^{99}_{(3,66)} = 4.12$, *siendo menor que la estadística* F *(de la figura anterior) por lo que los regresores aportan información en la explicación de la variable de respuesta* y.

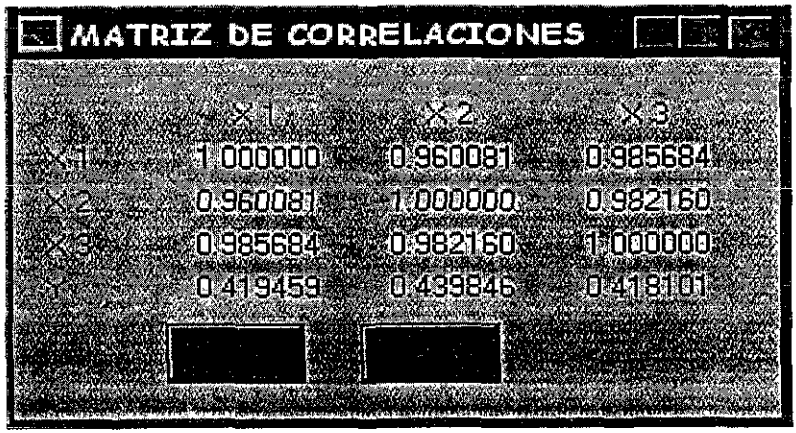


Figura C.13 Matriz de Correlaciones.

Las correlaciones simples entre los regresores (x_i) y la variable de respuesta (y), dadas en la figura anterior, son muy pequeñas por lo que se considera que la relación que tienen éstas no es muy fuerte. En el caso de las correlaciones entre los regresores, se tiene que éstas son muy cercanas a uno por lo que existe una gran dependencia lineal entre éstos; lo anterior hace pensar en la posible presencia del problema de colinealidad en los regresores. Lo más recomendable en este caso sería elegir aquel (o aquellos) regresor(es) que tenga(n) mayor correlación con la variable de respuesta (y) y que, además, en el caso de varios regresores, elimine al máximo el problema de la colinealidad.
Nota: La decisión de eliminar regresores del modelo lineal general queda siempre a consideración del analista.

Figura C.14: Detección de Colinealidad en los Regresores.

En la figura anterior, se presentan algunas estadísticas simples que permiten detectar el problema de colinealidad en los regresores. Para empezar, el signo de algunos de los coeficientes de correlación no coincide con el de su respectivo estimador (tal es el caso de β_0 y β_3), además, los Coeficientes de Inflación de la Varianza (CIV) para cada uno de los regresores son muy grandes (mayores que 10, valor grande estadísticamente hablando), por lo que se tiene evidencia clara de que el problema de colinealidad en los regresores está presente.

Nota: En ocasiones, estos indicadores no son suficientes en la detección del problema de colinealidad en los regresores por lo que, es necesaria la verificación de indicadores más sensibles en la detección de colinealidad.

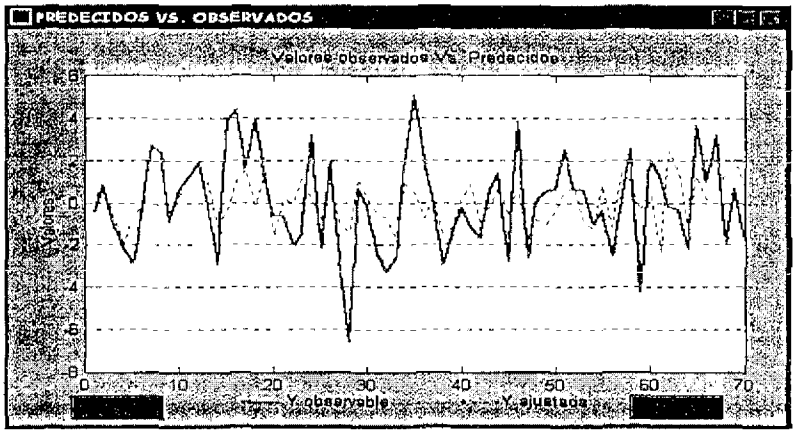


Figura C.15: Valores Predecidos $\left(\hat{y}\right)$ Vs. Observados (y).

Con la gráfica anterior se puede saber en forma rápida si el modelo de regresión ajustado predice en forma adecuada a la variable de respuesta (y). *Para tal propósito se grafican los valores observados* (y) *y los valores ajustados por el modelo de regresión* $\left(\hat{y}\right)$, *cada uno de ellos con su respectivo caso. Se dirá que el modelo ajustado predice en forma adecuada a la variable de respuesta si las tendencias resultantes de ambas gráficas son muy parecidas entre sí y tienen un comportamiento similar, en caso contrario, el modelo ajustado no es del todo confiable.*

En este ejemplo en particular, se tiene que las tendencias resultantes están muy distantes una de la otra y no tienen un comportamiento parecido por lo que es de suponerse que el modelo ajustado no es del todo un buen predictor de la respuesta dada por (y). *Esta misma conclusión puede ser vista con el Coeficiente de Determinación* $R^2 = 0.2062$, *el cual resultó ser muy pequeño.*
Nota: La decisión sobre el ajuste del modelo con este método queda a consideración del analista

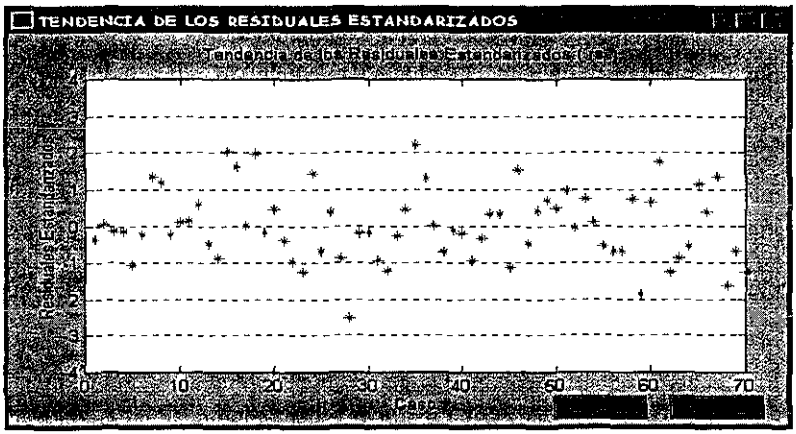


Figura C.16: Tendencia de los Residuales Estándarizados.

Con la gráfica anterior es posible detectar problemas de varianza no-constante en los residuales o, la presencia de alguna tendencia no lineal en el modelo ajustado. Los puntos resultantes, para el caso de varianza constante, deben tener una forma completamente aleatoria, no presentar tendencias crecientes o acumulación de puntos. Para el caso del modelo lineal, los puntos no deben presentar tendencias de alguna forma funcional conocida (cuadrática, cúbica, polinomial, etc.), ya que esto indicaría que un modelo lineal no es el adecuado en el ajuste de los datos. Por otro lado, la ventaja de usar Residuales Estándarizados en esta gráfica es que permite al analista detectar posibles puntos extremos ("outliers") en el conjunto de datos, que son aquellos que se encuentran fuera del intervalo $(-2, 2)$ *de una distribución Normal estándarizada.*

En este ejemplo, los puntos resultantes presentan una forma aleatoria y no hay tendencias de alguna función conocida por lo que se puede considerar que existe varianza constante en los residuales y además, el modelo lineal es el adecuado en el ajuste de los datos. Existen puntos fuera del intervalo $(-2, 2)$ *que son considerados datos aberrantes. Se recomienda al analista verificar si estas observaciones no causan problemas en el ajuste del modelo. Nota: La decisión de los puntos anteriores es muy subjetivo, por lo que queda siempre a consideración del analista.*

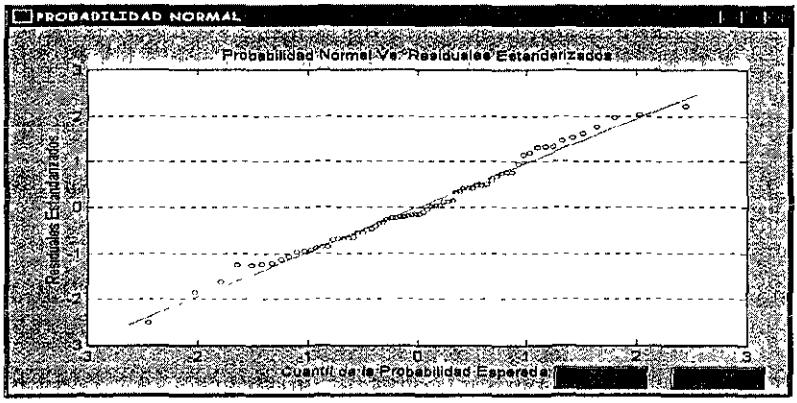


Figura C.17: Gráfica de Probabilidad Normal de los Residuales.

El supuesto de normalidad en los residuales parece cumplirse, pues los puntos resultantes no se encuentran muy distantes de la recta (arriba), y las barras tienen una tendencia cercana a la normal (abajo), aunque la decisión sobre este supuesto depende, en gran medida, del analista.

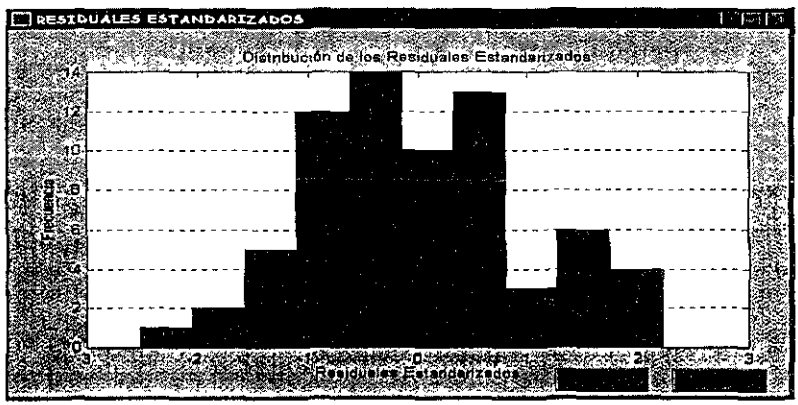


Figura C.18: Histograma de Residuales Estándarizados.

El siguiente ejemplo es utilizado como ilustración para las partes de centralización y estandarización de datos, debido a que éstos en ocasiones son muy grandes (o muy pequeños) y pueden ocasionar problemas por redondeo al momento de realizar los cálculos numéricos y, aunque se utilizan los algoritmos más avanzados, en ocasiones el problema de la obtención de resultados radica principalmente en las observaciones muestrales que fueron utilizadas para dicho proceso, las cuales no pueden ser del todo confiables. Es por esta razón que se presentan estas dos alternativas de estimación

Si desea más información al respecto consulte la sección 1.3.1 o los textos Montgomery [9], Neter [12], Searle [15] y Weisberg [20], entre otros.

Ejemplo C.1.3 (Weisberg, pag. 78) *Los datos dados en la figura C.19 fueron dados por Longley (1967) para demostrar insuficiencias de los programas computacionales en regresión entonces disponibles. Las siete variables son:*

x_1 = *GNP precio, en porcentaje.*
x_2 = *GNP, en millones de dólares.*
x_3 = *Desempleo, en miles de personas.*
x_4 = *Tamaño de la fuerza armada, en miles.*
x_5 = *Población no-institucional de más de 14 años de edad, en miles*
x_6 = *Año.*
y = *Desempleo total derivado en miles.*

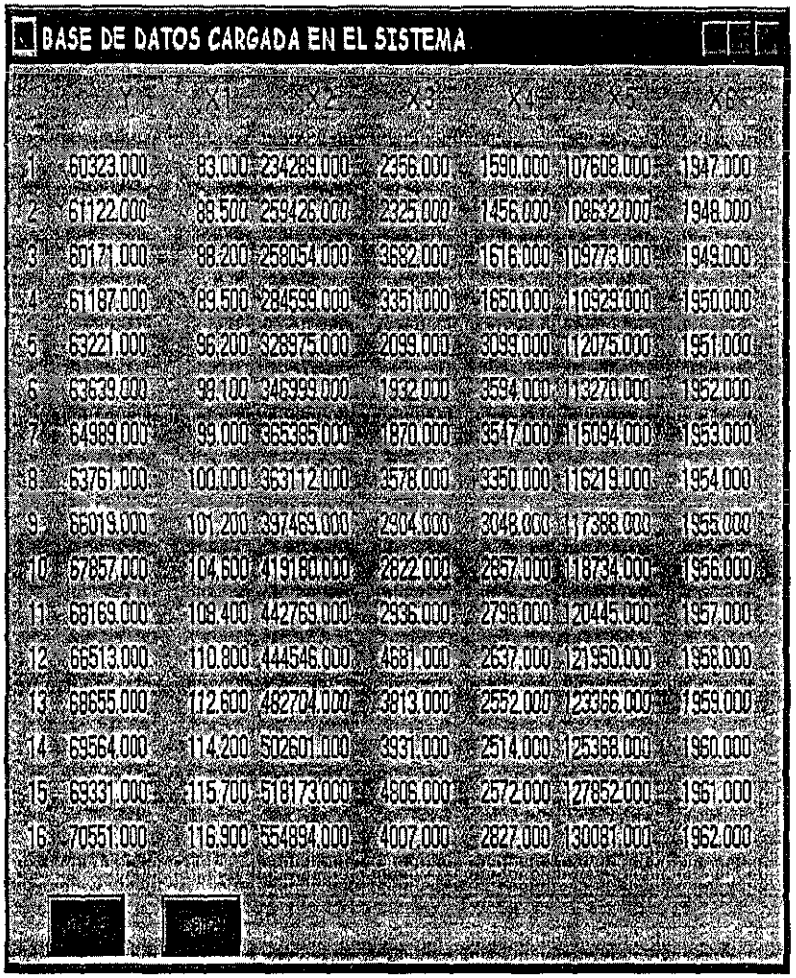


Figura C.19: Datos del ejemplo propuesto por Longley.

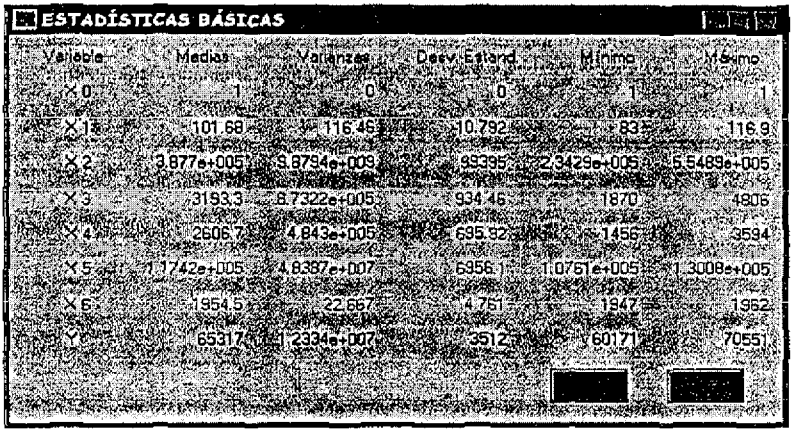


Figura C.20: Estadísticas Básicas para los regresores originales.

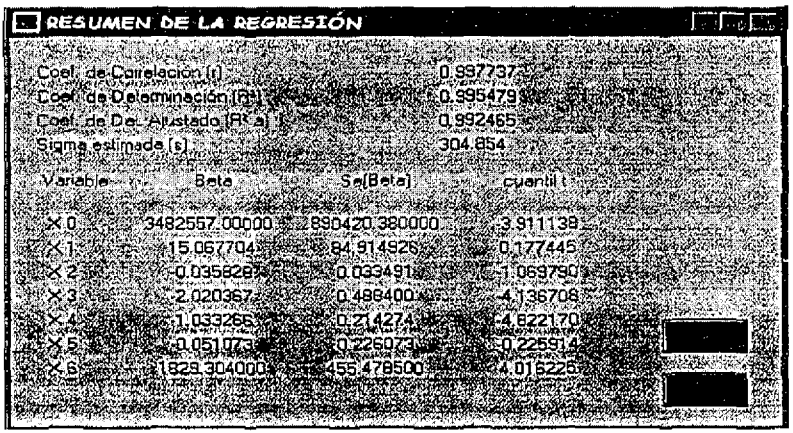


Figura C.21: Resumen de la Regresión obtenida sin cambios al modelo de regresión.

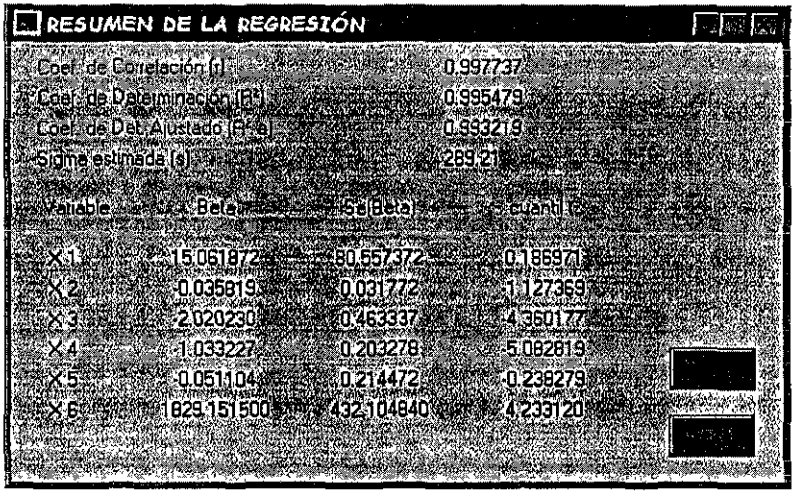


Figura C.22: Resumen de la Regresión para datos centrados respecto a su media.

Al centrar las observaciones respecto a su media el parámetro β_0 desaparece del modelo de regresión. En la sección donde se revisó "centralización de datos", se mencionó que las estimaciones de los parámetros desconocidos para el modelo original y para el modelo centralizado, `no cambiaban`. *Si se revisan los valores que aparecen en la tabla de Resumen de la Regresión, notará que hay pequeñas diferencias en la estimación de los parámetros, lo cual contradice el hecho antes mencionado. Estas variaciones se deben principalmente a la naturaleza de los datos; pues algunos valores, para algunas variables, son muy grandes y esto puede provocar errores por redondeo en los procedimientos utilizados, haciendo que los resultados sean erróneos. Por esta razón se recomienda utilizar datos centrados respecto a su media*

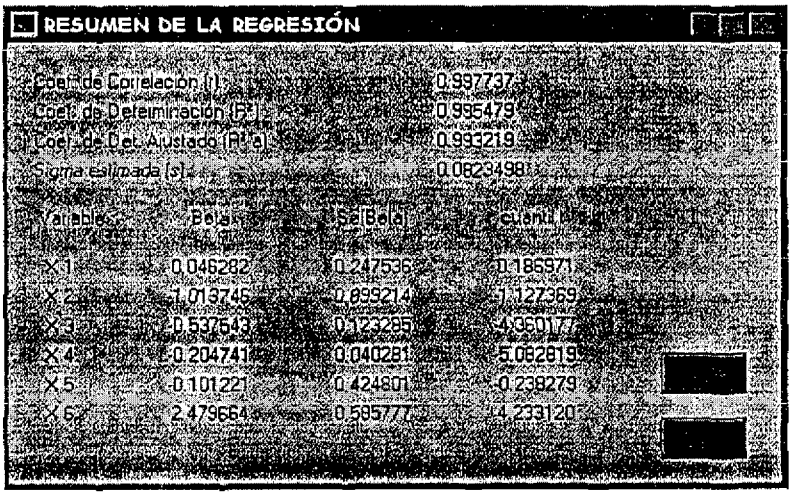


Figura C.23: Resumen de la Regresión para datos Estándarizados.

Para el caso en que se estándarizan los datos, los valores de los parámetros estimados cambian en forma considerable respecto a los valores del modelo original y, en la mayoría de los casos resultan tener pequeñas estimaciones. Este método de presentación de los datos, al igual que el de centralización, ayuda a evitar los errores por redondeo cuando están efectúandose los cálculos numéricos. Este método de ajuste del modelo se recomienda ser usado cuando los valores de los regresores son muy grandes (o muy pequeños), para intentar minimizar al máximo posibles errores por redondeo.

C.2 Actualización y creación de funciones.

Una de las características principales del sistema ANA_RELI.M VER. 2.0 es la facilidad con la cual se le puedan modificar o agregar opciones. Tales opciones pueden ser: algún menú, alguna ventana adicional o simplemente opciones de botones.

A continuación se presenta la sintaxis con la que se deben programar las distintas opciones:

C.2.1 Opción "figure".

La opción que permite la creación de una ventana es "figure", con esta opción se puede establecer un determinado nombre a la ventana, un color específico, el tamaño y un menú (opcional).

Ejemplo C.2.1 *En la expresión (C.1), la propiedad* 'name' *coloca el nombre* ANÁLISIS DE REGRESIÓN *a la ventana generada;* 'numbertitle' *numera las ventanas que se generan, en este caso tal número no aparecerá;* 'color' *pone el color indicado como una mezcla de la terna de rojo, verde y azul [Rojo Verde Azul], donde los valores están entre 0 y 1;* 'units' *especifica las unidades en las que se encontrará el tamaño de la ventana;* 'position' *indica la posición y tamaño de la ventana [Derecha Abajo Ancho Alto] (si la propiedad* 'units' *fue* 'normalized', *los valores dados en* 'position' *deben estar entre 0 y 1; finalmente,* 'menubar' *elimina el menú en la ventana, si esta propiedad no se indica el menú aparecerá en la ventana.*

```
figure('name','ANÁLISIS DE REGRESIÓN',...
    'numbertitle','off',...
    'color',[.55 .85 1],..
    'visible','on',...                                    (C.1)
    'units','normalized',...
    'position',[.05 .08 .9 .81],...
    'menubar','none');
```

C.2.2 Opción "uicontrol".

La opción que permite la creación de botones y texto en la ventana generada es "uicontrol", con esta opción se pueden generar botones, texto estático, texto editable, barras de desplazamiento y menús desplegables.

Ejemplo C.2.2 *En la expresión (C.2); la propiedad* 'style' *indica la característica del resultado dado en* uicontrol, *en este caso* 'push' *se refiere a colocar un botón en la pantalla, con el nombre* 'BorrarOutliers', *que será colocado en la coordenada .8 y .45 (considerando que la ventana normalizada mide* 1×1*) y medirá* $.15 \times .06$ *en tamaño proporcional al de la ventana, la propiedad* 'callback' *llama a la función* borra, *la cual indica dónde se encuentra el procedimiento a realizar (Este botón puede encontrarse en la pantalla de "ANÁLISIS DE REGRESIÓN", figura 5.2).*

```
borra ='anareg("borraout");';
uicontrol('style','push', ..
    'units','normalized',...
    'position',[.8 .45 .15 .06],. .
    'string','Borrar Outliers',...
    'callback', borra);
```
(C.2)

Para la propiedad 'style', en uicontrol, hay varias alternativas de uso, a continuación se presentan estas alternativas y el resultado en pantalla:

1. **edit:** genera texto editable en la ventana.
2. **popup:** genera un menú desplegable en la ventana.
3. **push ó pushbutton:** genera un botón a oprimir en la ventana
4. **slider:** genera una barra de desplazamiento en la ventana.
5. **text:** genera texto estático en la ventana.

C.2.3 Opción "uimenu".

La opción "uimenu" permite al programador adicionar un menú o un submenú a la figura que se ejecute (véase sección C.2.1). Si se desea que el menú aparezca en una determinada ventana entonces la opción 'uimenu' debe aparecer en el mismo programa fuente en donde se encuentra la opción 'figure' que ejecuta la ventana. Esto es, en la figura 5.2 aparecen los menús "Estadísticas", "Gráficas" y "Ayuda", los cuales fueron editados en el mismo programa fuente que crea la ventana de "ANÁLISIS DE REGRESIÓN".

Ejemplo C.2.3 *En la expresión (C.3), la primera parte genera el menú con nombre* 'Estadísticas' *y se indica que aparezca en la ventana como parte del menú principal en la primera posición (en caso de ser varios), la segunda parte de esta expresión es un submenú que aparece en la opción* 'Estadísticas', *llamado* 'Resumen de la Regresión' *que presenta la ventana donde se encuentra el Resumen de la Regresión obtenido por el sistema, después de llamar la función que lo genera,* resume *(Este menú puede encontrarse en la pantalla de "ANÁLISIS DE REGRESIÓN", figura 5.2).*

```
estadis = uimenu(gcf,...
    'label','&Estadísticas',...
    'position',1);

resume ='regcalc("Resreg");';                                    (C.3)
resumen = uimenu(estadis,...
    'label','&Resumen de la Regresión',...
    'callback',resume);
```

Los ejemplos aqui mostrados pueden servir como ayuda al programador en caso de que desee modificar alguna de las presentaciones establecidas en el sistema ANA_RELI.M VER. 2.0. Si este es el caso se le sugiere tener copias de respaldo antes de llevar a cabo cualquier modificación o alteración del mismo.

Si desea profundizar en el tema de programación en Matlab, orientada a la creación de ventanas y botones, se recomienda la revisión de los siguientes textos, enfocados básicamente a este tema, Manual del sistema [8], Nakamura [11].

Bibliografía

[1] Belsley, D. A., Kuh, E. & Welsch, R. E. [1980], *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley & Sons.

[2] Bjork, A. [1996], *Numerical Methods for Least Squares Problems*, SIAM.

[3] Chatterjee, S. & Price, B. [1977], *Regression Analysis by Example*, Wiley & Sons.

[4] Golub, G. H., Van Loan, Ch. F. [1996], *Matrix Computations*, Johns Hopkins. 3rd. Ed

[5] Hanson, R. J. [1975], *Stably updating mean and standard deviation of data*, *Communications of the ACM* 18, 57-58.

[6] Higham, N. J. [1996], *Accuracy and Stability of Numerical Algorithms*, SIAM. Philadelfia.

[7] López, J. & Madrid, H. [Por aparecer], *Colinealidad y Factorización RRQR*.

[8] *MATLAB. Users Guide*, The MathWorks Inc.

[9] Montgomery, D. C. & Peck, E. A. [1992], *Introduction to Linear Regression Analysis*, Wiley & Sons, 2nd Ed.

[10] Mood, A., Graybill, F., & Boes, D. [1974], *Introduction to the Theory of Statistics*, McGraw-Hill, 3rd Ed.

[11] Nakamura, S. [1997], *Análisis Numérico y Visualización Gráfica con MATLAB*, Prentice Hall.

[12] Neter, J., Wasserman, W. & Kutner, M. H. [1990], *Applied Linear Statistical Models*, Irwin, Inc., 3rd Ed.

[13] Ramírez, P. [1994], *Ana_Reli.Sis: Un sistema Amigable e Interactivo para el Análisis de Regresión Lineal*. Facultad de Ciencias.

[14] Ryan, T. P. [1997], *Modern Regression Methods*, Wiley, New York

[15] Searle, S. R. [1971], *Linear Models*, Wiley & Sons.

[16] Seber, G. A. F. [1977], *Linear Regression Analysis*, Wiley & Sons.

[17] Stewart, G. W. [1998], *Matrix Algorithms. Vol. I: Basic Decompositions*, SIAM.

[18] Stewart, G. W. & Sun, J. G. [1990], *Matrix Perturbation Theory*, Academic Press, San Diego.

[19] Thisted, R. A. [1988], *Elements of Statistical Computing*, Chapman and Hall.

[20] Weisberg, S. [1985], *Applied Linear Regression*, Wiley & Sons, 2nd Ed.