60
2 Ejem

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

" ANA_RELI. SIS: UN SISTEMA AMIGABLE E INTERACTIVO 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 :

PATRICIA RAMIREZ MARIN

MEXICO, D. F. 1994

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

AGRADECIMIENTOS

Al Dr. Jesús López Estrada por la valiosa dirección de este trabajo de tesis.

A los integrantes del jurado:

Dr. Jesús López Estrada

Dra. Blanca Rosa Pérez Salvador

Mat. Margarita Elvira Chavez Cano

M. en C. José Antonio Flores Díaz

M. en C. María Elena García Alvarez

por su acertada orientación.

A la Universidad y a la Facultad de Ciencias

A los maestros con afecto y gratitud

A mi familia y amistades

Prólogo

"La vida es el arte de sacar conclusiones suficientes a partir de datos insuficientes"
Samuel Butler

El Análisis de Regresión se aplica en diversas áreas del conocimiento ya que son múltiples los ejemplos en que es posible observar y medir variables de interés que se encuentran asociadas a otras y mediante una función que puede ser aproximada por un modelo lineal. Cuando el número de variables que intervienen en el modelo es mayor de cinco se recurre a una computadora para efectuar las operaciones aritméticas, sin embargo su empleo en este caso debe efectuarse considerando los métodos numéricos para garantizar que los resultados serán obtenidos de la manera más precisa, ya que esto es fundamental en la toma de decisiones que se deriva de este análisis; y en caso de no considerar métodos numéricos adecuados el problema que puede presentarse es que el redondeo en las operaciones que realiza la computadora puede provocar tomar decisiones equivocadas, es por ello que en el desarrollo del sistema ANA_RELI.SIS, objeto de este trabajo se utilizan las rutinas del paquete LINPACK, que son sobriamente soportadas en los aspectos teóricos y que han sido verificadas experimentalmente.

El siguiente trabajo tiene como objetivo el desarrollo de un sistema amigable e interactivo para llevar a cabo el **análisis de regresión para un modelo lineal:**

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

el cual va desde la estimación Gauss-Markov para β y σ^2 hasta el análisis de las observaciones así como de las variables explicativas del modelo, pasando por el análisis estadístico del mismo a través de pruebas de hipótesis (validación del modelo y prueba F parcial), como con base en coeficientes de determinación, de inflación de varianzas, etc.

En el capítulo primero se revisan en forma breve los conceptos básicos utilizados en el análisis de regresión lineal.

En el capítulo segundo se presenta una breve discusión sobre las estadísticas y coeficientes (o indicadores) más conocidos para llevar a cabo el análisis estadístico de un modelo lineal. Que va desde ver que tan bien se ajusta el modelo a los datos, hasta la realización de pruebas de hipótesis (validación del modelo y prueba F parcial), y la determinación de la correlación entre las variables entre otros.

En el capítulo tercero se revisa el análisis de residuales por medio de métodos gráficos, presentando así otro criterio además del probabilístico, mediante el cual es posible llegar a ciertas conclusiones con respecto a los supuestos o hipótesis de trabajo para el estudio del modelo. Sin embargo los resultados obtenidos de ésta manera, son de carácter subjetivo.

En el capítulo cuarto se presenta un análisis numérico de las ecuaciones normales, y de su muy preferible alternativa para el cálculo de las estimaciones de Gauss-Markov para β y σ^2 con base en la descomposición QR de la matriz X del modelo lineal .

En el capítulo quinto se presenta el sistema ANA_RELI.SIS, cuyas características y objetivos son el ser un sistema de uso accesible y confiable en el aspecto numérico, para lograr lo primero se decidió que la interfase con el usuario se programará

en lenguaje C, y para alcanzar el segundo objetivo los cálculos numéricos se llevan a cabo mediante el uso de subrutinas del paquete profesional Linpack, considerado el mejor en cuanto a los algoritmos numéricos que utiliza.

La documentación (a la vez Manual Técnico) relacionada al sistema ANA_RELI.SIS mediante el cual se obtendrán las estadísticas mencionados en los capítulos I, II y III se encuentra en el apéndice que aparece al final de este trabajo en el que se detalla el funcionamiento, diseño y mantenimiento del sistema.

La nomenclatura que se utiliza a través del texto se encuentra en el apéndice B.

En el apéndice C se presenta el análisis de error correspondiente a los siguientes algoritmos, al conocido como de dos pasos, al que se menciona en los libros de texto de estadística y al desarrollado por West-Hanson, para obtener la varianza muestral, éste último es el que se emplea en el sistema ANA_RELI.SIS; agradezco al Profr. Jesús López Estrada el haberme permitido incluir su desarrollo del tema "Análisis de Sensibilidad para la Media, Varianza y Desviación estándar Muestrales" así como el haber complementado las demostraciones que se incluye en este apéndice.

INDICE Pág.

C A P I T U L O I

Conceptos Básicos Utilizados en el Análisis de Regresión Lineal

Introducción

En este capítulo se presenta una revisión breve sobre los fundamentos matemáticos del análisis de regresión, tema al cual está enfocado el presente trabajo. Para ello, se utilizarán ejemplos por medio de los cuales se ilustra la aplicación de esta teoría.

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

1.1 Modelos Lineales

Cuando se tiene el problema de necesitar conocer la asociación que existe entre una variable de respuesta **y** y ciertos datos numéricos medibles conocidos como variables de control o independientes $(x_1, x_2, \ldots, x_p)$, se puede recurrir al planteamiento de un modelo lineal

$$y = X\beta + \varepsilon \qquad (1.1.1)$$

En este modelo el investigador incluirá todas aquellas variables que considere son relevantes para determinar el comportamiento de la variable dependiente **y**.

Es frecuente encontrar en prácticamente todas las áreas del conocimiento, como son la física, la biología, la astronomía, la medicina, etc... problemas en que existe una variable de interés

y asociada a una o más variables auxiliares o explicativas $x_1, \ldots x_p$, pero desconocer el grado de asociación, y mediante el empleo de un modelo lineal, es posible estimarlo.

Y si existe una función que describe la asociación de **y** con $x_1, \ldots x_p$ como una aproximación aceptable se puede considerar un hiperplano.

De ésta manera el estudio de **y** se puede enfocar a estudiar la función de asociación de $\mathbf{y} \in \mathbb{R}^p$ con respecto a $X \in \mathbb{R}^{nxp}$ proponiendo un modelo lineal de la forma :

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

donde **y** es el vector de observación, X la matriz de variables explicativas β un vector de parámetros desconocidos y ε el error de observación.

El método que utilizaremos para determinar las β's es el muy conocido método de mínimos cuadrados.

A las variables **y's** y a las **x's** se les conoce como variables dependiente e independientes (o variables de control) respectivamente. La notación matricial de los términos que interviene en el modelo (1.1.1) es la siguiente:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

En éste capítulo, se presentan ejemplos en los que se supone que los datos representados en la matriz X y la respuesta en el vector y están asociadas mediante un modelo lineal $y = X\beta + \varepsilon$, donde ε como se mencionó es el vector de errores observados que son el resultado de procesos no controlables, se considera aleatorio y que se distribuye como una normal multivariada con vector de medias 0 y matriz de varianza-covarianza $\sigma^2 I$, es decir

$$\varepsilon \sim N(0, \sigma^2 I) \qquad (1.1.2)$$

El valor esperado de y es :

$$E(y) = X\beta,$$

pero cuando se observa algún fenómeno, existe siempre cierta diferencia, desviación, entre lo que se observa y lo que teoricamente debe ocurrir, estas diferencias son los errores de observación que mencionamos anteriormente, entre los que se pueden considerar los debidos a: cierta inexactitud de los aparatos con que se efectuan las mediciones, o bien al criterio que se utilizó para determinarlas.

1.2 Mínimos Cuadrados

Uno de los métodos para estimar el vector β es el de mínimos cuadrados; el cual se establece así :

Hallar β que minimice

$$SCE = e'e = (y - X\beta)'(y - X\beta),$$

donde $e = y - X\beta$ es el vector de residuales.

Diferenciando la función $e'e$ con respecto a β e igualando a cero, se llega a que $d/d\beta\ e'e=0$ si

$$X'X\beta = X'y, \qquad (1.2.1)$$

Como se puede ver $\mathbf{e'e}$ como función del vector β es continua y diferenciable sólo si $X'X\ \beta = X'y$, por lo que se pueden utilizar los resultados del cálculo diferencial para encontrar el vector que lo minimice.

A las ecuaciones (1.2.1) se les conoce como las ecuaciones normales.

Si X tiene rango p (que es el número de variable del modelo), entonces $X'X$ es positiva definida, y las ecuaciones normales tienen una única solución

$$\hat{\beta} = (X'X)^{-1}X'y,$$

donde el vector $\hat{\beta}$ se conoce como el estimador de mínimos cuadrados de ß.

Bajo el supuesto que $\varepsilon \sim (0,\sigma^2 I)$, las principales características de los estimadores mínimos cuadrados son:

1.- $E[\hat{\beta}] = \beta$.

Por lo que el estimador β es un estimador insesgado de β. (Seber pag.48)

2.- $\text{var}[\hat{\beta}] = \sigma^2(X'X)^{-1}$

En efecto,

$$\begin{aligned} \text{var}[\hat{\beta}] &= \text{var}[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}\ X'\text{var}[y]\ X(X'X)^{-1} \quad \text{ya que } \text{var}[y]=\sigma^2 I; \\ &= \sigma^2(X'X)^{-1}\ (X'X)(X'X)^{-1} \\ &= \sigma^2\ (X'X)^{-1} \end{aligned}$$

3.- $$s^2 = \frac{(\ y - X\hat{\beta})'\ (\ y - X\hat{\beta})}{n-p} = \frac{SCE}{n-p}$$

es un estimador insesgado de σ^2, el cual es independiente de $\hat{\beta}$.

En efecto, debido a la hipótesis de normalidad, basta demostrar que la covarianza es cero.

$$\begin{aligned} Cov[\hat{\beta}, y-X\hat{\beta}] &= Cov[(X'X)^{-1}X'y,\ (I_p - P)y] \\ &= (X'X)^{-1}X'\,var[y](I_p - P)' \\ &= \sigma^2(X'X)^{-1}\,X'(I_p - P) \\ &= 0 \end{aligned}$$

donde

$P = X(X'X)^{-1}X$ y la matriz I_p es la matriz identidad en $\mathbb{R}^{pxp}$

(Seber pag.55)

4.- Si además se cumple (1.1.2) entonces y se distribuye como una $N(X\beta, \sigma^2 I)$ donde X es de n x p de rango p, entonces se tiene que:

(i) $\hat{\beta}$ se distribuye como una $N(\beta, \sigma^2(X'X)^{-1})$.

(ii) $(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})/\sigma^2$ se distribuye como una χ^2_p.

(iii) $\hat{\beta}$ es independiente de s^2.

(iv) $SCE/\sigma^2 = (n - p)s^2/\sigma^2$ se distribuye como una χ^2_{n-p}.

(Seber 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, efectuar su análisis con base en pruebas de hipótesis, etc.

Algunas estadísticas básicas de mucha utilidad en el análisis de regresión se basan en los siguientes hechos:
1.- Si k-variables aleatorias $X_1, \ldots, X_k$, se distribuyen normal e independientemente con medias μ_i y varianzas σ^2_i, entonces

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

tiene una distribución ji-cuadrada con k grados de libertad. (Mood p.242), como ejemplos de esta distribución tenemos a SCT, SCR y SCE que se definen en la siguiente sección.

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

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

se distribuye como una variable F con m y n grados de libertad. (Mood p.247).

3.- Si $Q_i \sim X_{ri}$ para i=1,2, r1>r2, y $Q = Q_1 - Q_2$ es independiente de Q_2, entonces $Q \sim X_r^2$ donde r = r1-r2 (Seber pag.20)

4.- Si Z es una variable aleatoria normal estándar, 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 pag.250).

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 yi's observadas de sus valores esperados" [Searle pag.92] es conocida como la suma de cuadrados de residuales, y denotada de la siguiente forma:

$$SCE = (y - \hat{y})'(y - \hat{y}) = (y - X\hat{\beta})'(y - X\hat{\beta}) \qquad (1.3.1)$$

La suma de cuadrados total está definida por:

$$SCT = \mathbf{y'y} \qquad (1.3.2)$$

Se tiene que SCE es una medida de la variación en y después de estimar $\hat{\beta}$ y nos indica que 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 β.

La diferencia entre (1.3.2) y (.1.3.1) es conocida como la suma de cuadrados atribuible a la regresión, y es denotada por:

$$SCR = SCT - SCE,$$

y "representa que porción de SCT es atribuible a tener la regresión ajustada" [Searle pag.94] es decir indica la variación de la estimación con respecto al promedio.

Existen dos criterios muy conocidos para trabajar los datos, que son el de centrarlos y el de estandarizarlos, el primero 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. Al estandarizar los datos además de obtener las ventajas de que aporta el centrarlos, 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.

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

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

2o) Obtener la media de la variable de respuesta y ($\bar{y}$);

3o) $X^*_{ij} = X_{ij} - \bar{x}_j$

4o) $y^*_i = y_i - \bar{y}$.

Al centrar los datos los parámetros $\hat{\beta}$ 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 ésta queda determinada por la siguiente expresión:

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

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

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

$$SCT_m = y^{*\prime} y^*$$

donde X^* y y^* son los datos centrados.

El procedimiemto para estandarizar los datos es el siguiente:

A partir del modelo

$y = X\beta + \varepsilon$ se construye un nuevo modelo

$y^* = X^*\beta^* + \varepsilon$

donde :

$$x^*_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

$$s_j^2 = \frac{\Sigma (x_{ij} - \bar{x}_j)^2}{n-1}$$

$$y_i = \frac{y_i - \bar{y}}{s_y}$$

$$s_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$

donde $\bar{x}_i$ i=1,..p, son las medias de las columnas de la matriz X y $\bar{y}$ es la media de la variable de respuesta **y**.

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

$$\hat{\beta}_j = \hat{\beta}_j^* \ (s_y^2/s_j^2)^{1/2}$$

y

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1} \hat{\beta}_j \bar{x}_j$$

1.4 Estadística F

Las sumas SCE/σ^2, SCR/σ^2, y $SCRm/\sigma^2$ tienen distribuciones ji-cuadradas y se tiene que los siguientes cocientes tienen distribuciones F, al estar formados por ji-cuadradas divididas entre sus grados de libertad, es decir

$$F = \frac{SCR/p}{SCE/(n-p)} \qquad (1.4.1)$$

$$F(R_m) = \frac{SCR/(p-1)}{SCE/(n-p)} \qquad (1.4.2)$$

Si se desea una explicación teórica del orígen de la distribución F, ésta se puede encontrar en la referencia [8] y en el Mood.

En el Análisis de la varianza se compara a las estadísticas F y $F(R_m)$ con el cuantil $(1-\alpha)$ que hay en las tablas de la distribución F con p, n-p y con p-1,n-p grados de libertad respectivamente.

Cuando el valor de la estadística F determinada por (1.4.1) o (1.4.2) es mayor que el del cuantil encontrado en la tablas de la distribución F a un nivel $(1-\alpha)$, se rechaza la hipótesis de que $\beta = 0$ al nivel de significancia α.

Ejemplo 1.4.1 [Searle]

Supongase que el ingreso anual de un hombre (y) está asociado a su grado de escolaridad (x_1) y a su edad (x_2) mediante un modelo de regresión lineal, entonces su ingreso esperado es:.

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (1)$$

Hombre i	Ingreso y_i	Años de escolaridad x_{i1}	Edad x_{i2}
1	10	6	28
2	20	12	40
3	17	10	32
4	12	8	36
5	11	9	34

A partir de estos datos el método de mínimos cuadrados proporciona la estimación

$$\hat{E}(y) = 56/24 + (50/24)x_1 - (5/24)x_2$$

La estadística F = 5.43 , y el valor en tablas de la distribución $F_{(2,2)}$ = 19 con una significancia de 0.05 es de 19.15 y dado que F = 19 > 5.43 concluímos que el modelo no explica la variación de la variable **y** con un nivel de significancia de 0.05., por lo tanto se debe rechazar la hipótesis de que todos los β_i son distintos de cero es decir, que exista una relación entre las variables de control y la variable independiente, y el nivel de significancia de la prueba es de .05.

1.5 Intervalos de Confianza

El método de mínimos cuadrados proporciona estimadores puntuales para los parámetros del modelo lineal β, σ^2 y E(**y**). Dados éstos estimadores se puede encontrar un intervalo alrededor de ellos que con una confianza fija pueden contener al parámetrc.

Para las coordenadas de β se tiene que:

$\hat{\beta}_i \sim N(\beta_i, \sigma^2 a_{ii})$ entonces

$P(|\beta_i - \hat{\beta}_i| < t_{\alpha/2}\, s) = 1-\alpha$

$P(\hat{\beta}_i - t_{\alpha/2}\, s < \beta_i < \hat{\beta}_i + t_{\alpha/2}\, s)$

Es por esto que los intervalos de confianza simétricos con respecto al valor estimado de β_i o también conocidos como intervalos de tolerancia, se encuentran dados para cada parámetro $\hat{\beta}_i$ por:

$$\beta_i \in (\hat{\beta}_i \pm st_{(n-p,\alpha/2)} * \sqrt{a_{ii}})$$

Donde $\sqrt{a_{ii}}$ es el i-ésimo elemento de la diagonal de la matriz $(X'X)^{-1}$, s es la raíz cuadrada del valor estimado de la varianza, $t_{(n-p,\alpha/2)}$ es el cuantil $(1 - \alpha/2)$ de una distribución t con n - p grados de libertad.

CAPITULO II

Revisión de Estadísticas que Permiten la Validación del Modelo de Regresión Lineal

Introducción

En este capítulo se mencionarán los supuestos básicos para llevar a cabo un Análisis de Regresión, y la obtención de estadísticas que permiten evaluar el modelo de regresión y a las variables que intervienen en el mismo.

2.1 Supuestos Básicos en un Modelo de Regresión Lineal

Para poder aplicar de manera conveniente el método de regresión en un modelo lineal múltiple ó de regresión lineal general

$$y = X \beta + \varepsilon \qquad (2.1.1)$$

se debe observar que se cumplan los siguientes supuestos:

1) El modelo debe ser lineal en los parámetros β_i's

2) Los errores en las observaciones deben ser estocásticamente independientes, y tener una distribución idéntica de media cero y la misma varianza constante, esto es $\varepsilon \sim (0,\sigma^2 I)$. Con frecuencia es adecuado suponer que la distribución de los errores en las observaciones es normal, para que con ello sea posible efectuar pruebas de hipótesis, esto es

$$\varepsilon \sim N(0,\sigma^2 I).$$

3) X es una matriz de entradas reales de n x p donde n > p de rango p , el que el rango de la matriz sea p, significa que las columnas de X son linealmente independientes.

A continuación se plantea un problema que da lugar a un modelo lineal con cuatro variables para estudiar el consumo de combustible.

Ejemplo 2.1.1 (Weisberg)

x_1 = TCC = tasa de consumo de combustible 1972, en centavos por galón.
x_2 = PPLC = porcentaje de población con licencia de conductores
x_3 = IP = ingreso promedio (miles de dólares)
x_4 = RM = (miles de millas)
y = CCM = consumo de combustible del motor (galones por persona)

	ESTADO i	x_1 TCC	x_3 IP	x_4 RM	x_2 PPLC	y CCM
1	ME	9.00	3.571	1.976	52.5	541
2	NH	9.00	4.092	1.250	57.2	524
3	VT	9.00	3.865	1.586	58.0	561
4	MA	7.50	4.870	2.351	52.9	414
5	RI	8.00	4.399	.431	54.4	410
6	CN	10.00	5.342	1.333	57.1	457
7	NY	8.00	5.319	11.868	45.1	344
8	NJ	8.00	5.126	2.138	55.3	467
9	PA	8.00	4.447	8.557	52.9	464
10	OH	7.00	4.512	8.507	55.2	498
11	IN	8.00	4.319	5.939	53.0	580
12	IL	7.50	5.126	14.186	52.5	471
13	MI	7.00	4.817	6.930	57.4	525
14	WI	7.00	4.207	6.580	54.5	508
15	MN	7.00	4.332	8.159	60.8	566
16	IA	7.00	4.318	10.340	58.6	635
17	MO	7.00	4.206	8.508	57.2	603
18	ND	7.00	3.178	4.725	54.0	714
19	SD	7.00	4.716	5.915	72.4	865
20	NE	8.50	4.341	6.010	67.7	640
21	KS	7.00	4.593	7.834	66.3	649
22	DE	8.00	4.983	.602	60.2	540
23	MD	9.00	4.897	2.449	51.1	464
24	VA	9.00	4.258	4.686	51.7	547
25	WV	8.50	4.574	2.619	55.1	460
26	NC	9.00	3.721	4.746	54.4	566
27	SC	8.00	3.448	5.399	54.8	577
28	GA	7.50	3.846	9.061	57.9	631
29	FL	8.00	4.188	5.975	56.3	574
30	KY	9.00	3.601	4.650	49.3	534
31	TN	7.00	3.640	6.905	51.8	571
32	AL	7.00	3.333	6.594	51.3	554
33	MS	8.00	3.063	6.524	57.8	577
34	AR	7.50	3.357	4.121	54.7	628
35	LA	8.00	3.528	3.495	48.7	487
36	OK	6.58	3.802	7.834	62.9	644
37	TX	5.00	4.045	17.782	56.6	640
38	MT	7.00	3.897	6.385	58.6	704
39	ID	8.50	3.635	3.274	66.3	648
40	WY	7.00	4.345	3.905	67.2	968
41	CO	7.00	4.449	4.639	62.6	587
42	NM	7.00	3.656	3.985	56.3	699
43	AZ	7.00	4.300	3.635	60.3	632
44	UT	7.00	3.745	2.611	50.8	591
45	NV	6.00	5.215	2.302	67.2	782
46	WN	9.00	4.476	3.942	57.1	510
47	OR	7.00	4.296	4.083	62.3	610
48	CA	7.00	5.002	9.794	57.3	524

Como se mencionó al final de la sección 1.3 en el capítulo anterior es posible centrar los datos con respecto a sus medias, para obtener mejores resultados desde los puntos de vista estadístico y numérico, ya que así las variables son de magnitud parecida, y se evita el manejar números muy grandes y muy pequeños a la vez, además tal centralización no afecta la relación entre las variables; siguiendo el razonamiento anterior en el ejemplo 2.1.1 los resultados que se obtienen son:

$$\hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} -34.790149 \\ 13.364494 \\ -66.588752 \\ -2.425889 \end{bmatrix}$$

El estimador la β_0 es:

$$\hat{\beta}_0 = \hat{y} - \Sigma \hat{\beta}_j^* \bar{x}_j = 377.2911$$

El cuadrado medio del error es :

$$s^2 = \frac{SCE}{n-p'} = \frac{189049.97}{48-(4+1)} = 4396.5$$

tal media es el estimador de la varianza de la distribución.

2.2 Matriz de Correlación

En la matriz de correlación de un modelo de regresión lineal múltiple se encuentran todos los posibles coeficientes de correlación entre parejas de variables x_i incluídas en el modelo, éstos están definidos como:

$$\rho_{i,j} = \frac{cov(x_i, x_j)}{se(x_i) * se(x_j)}$$

al acercarse $|\rho_{i,j}|$ a 1 este coeficiente indica que las variables i y j están altamente correlacionadas de manera lineal y si el coeficiente se acerca a 0 indica que no lo están.

Ejemplo 2.2.1 Matriz de Correlación del ejemplo 2.1.1

TCC	1.0000				
PPLC	-.2880	1.000			
IP	.0127	.1571	1.000		
RM	-.5221	-.0641	.0502	1.000	
CCM	-.4513	.6990	-.2449	.0190	1.000
	TCC	PPLC	IP	RM	CCM

2.3 Errores Estándar

Los errores estándar y las covarianzas estimadas de los β_j's son encontradas a partir de la σ^2 estimada (s^2) y de $(X'X)^{-1}$.

$$se(\hat{\beta}_i) = s\sqrt{a_{ii}}$$

$$cov(\hat{\beta}_i, \hat{\beta}_j) = s^2 a_{ij}$$

donde a_{ii} es el i-ésimo elemento de la diagonal de la matriz $(X'X)^{-1}$ y a_{ij} es el elemento en el i-ésimo renglón y la j-ésima columna de la misma matriz sin considerar en ésta el renglón y la columna correspondiente a la constante.

Para el ejemplo 2.1.1, la matriz $(X'X)^{-1}$ es:

Constante	7.83019	-.42651	-.06110	-.14950	-.07534
TCC	-.42651	.038263	.00221	-.00591	.00571
PPLC	-.06110	.00221	.00084	-.00145	.00041
IP	-.14950	-.00591	-.00145	.06746	-.00154
RM	-.07534	.00571	.00041	-.00015	.00261
	Constante	TCC	PPLC	IP	RM

Para el ejemplo 2.1.1 el error estándar de $\hat{\beta}_1$ y la covarianza de $\hat{\beta}_1$ $\hat{\beta}_2$ son :

$$se(\hat{\beta}_1) = s\sqrt{a_{11}} = s\sqrt{7.89019} = 185.54$$
$$cov(\hat{\beta}_1, \hat{\beta}_2) = s^2 a_{12} = s^2(0.0022158) = 0.1469$$

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

2.4 Estadística t

Con frecuencia interesa saber el peso que cada variable tiene en 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 interviene en el modelo la variable a la que corresponde y si este valor es pequeño, esto sugiere que la variable tiene poco peso en la explicación de la variable de respuesta **y** es por ello que el valor de esta estadística puede sugerir las 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.

Las estadísticas t correspondientes a cada variable se obtienen del cociente del valor estimado para tal variable y su error estándar, es decir

$$t_i = \hat{\beta}_i / se(\hat{\beta}_i)$$

(Montgomery p.125).

En el caso de que no exista multicolinealidad (en caso de que exista, es posible que con la prueba global se rechace Ho, pero ninguna H_i sea rechazada), para determinar si la variable que presenta una estadística t pequeña se puede descartar del modelo, se recurre a la prueba de hipótesis que se menciona a continuación:

Una prueba estadística para la prueba de Hipótesis

$$Ho: \beta_j = 0 \qquad Vs \qquad Ha: \beta_j \neq 0$$

$$t_{j,n-p} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Prueba conocida como: diferencia mínima significativa.

Al suponer que $\beta_j = 0$, se elimina la variable j-ésima del modelo.

La estadística t se compara con el cuantil $(1-\alpha)$ que tiene la distribución t de Student con n-p grados de libertad, y si el valor de la estadística es mayor al valor encontrado en las tablas de la distribución, se rechaza la hipótesis en caso contrario se acepta.

Ejemplo 2.4.1, para el ejemplo 2.1.1 se tienen las siguientes estadísticas:

Variable	Estimación de β	Error Estandar	Valor - t
Constante	377.2911	185.5412	2.03
TCC	-34.79015	12.97020	-2.68
PPLC	13.36449	1.922981	6.95
IP	-66.58875	17.22175	-3.87
RM	-2.425889	3.389174	-.72

s^2 = 4396.511, grados de libertad = 43, R^2 = 0.6787

Como ilustración se considera la siguiente Prueba de Hipótesis:

Ho: $\beta_1 = 0$ vs. Ha: $\beta_1 \neq 0$

el valor de la estadistica t para la variable 1 es -2.68 éste estadística al cuadrado es t^2=7.18, que es un valor menor al correspondiente a la distribución $F_{1,43}$ en tablas que es el de 7.26, por lo que se acepta la hipótesis Ho.

El uso de esta estadística como se mencionó permite resolver un problema que a menudo ocurre y es el de seleccionar el mínimo número de variables que expliquen de forma adecuada la variable de respuesta **y**, para lo cual se siguen los procedimientos que se describen a continuación:

1.- Se ordenan los coeficientes de regresión en forma decreciente con respecto a la magnitud de la estadística $|t_{k,j}|$ donde j=1,2,...,K, 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 para cada p.

En el caso de que haya varias variables $(x_i, x_j, \ldots)$ 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, como se verá en la sección 2.8 (Prueba F parcial) en la que se discute la siguiente prueba de hipótesis:

Ho: $(\beta_i = \beta_j = \ldots = 0)$ vs. H1: Alternativa de que no todas ellas sean iguales a cero

Nota: Como se mencióno previamente esté procedimiento funciona cuando no se presenta multicolinealidad.

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's, ésto se suele hacer en términos del coeficiente de determinación el cual se define como sigue:

Coeficiente de Determinación $R^2 = (SCT-SCE)/SCT = SCR/SCT$

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

En el ejemplo 2.1.1 se puede decir que aproximadamente el 68% de la variabilidad observada en la respuesta es modelada por las x's, ya que $R^2 = 0.6787$.

La distribución probabilística de ésta estadística se desconoce es por esto que con ella no se puede efectuar ninguna prueba a cierto nivel de significancia.

Sin embargo su uso es común y debe ser interpretada con precaución, ya que puede darse el caso de que el coeficiente de determinación encontrado sea cercano a uno y no se puede establecer con base en cierta probabilidad que tan bien se ajusta el modelo a los datos.

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

Como ya se vió en la sección 2.4, el valor de las t's puede sugerir la eliminación de ciertas variables en el modelo lineal general (MLG), o bien el investigador puede decidir descartar ciertas variables en dicho modelo.

En el caso de una sola variable, en la sección 2.4 se vió como se resuelve el problema, en esta sección se revisará la prueba de hipótesis que permite tomar la decisión de eliminar varias variables a la vez que se suponen no relevantes para el modelo.

Si se considera el modelo lineal general (MLG):

$$\mathbf{y} = \mathbf{X}\,\beta + \varepsilon, \quad \varepsilon \sim N(0,\sigma^2 I).$$

bajo el supuesto de la hipótesis nula, se obtiene un modelo de regresión reducido (MLR), esto es un modelo que incluye un subconjunto de las variables que se tienen en el modelo originalmente planteado es decir, el número de parámetros considerados en el MLR es menor que el número de parámetros a ser estimados en el MLG.

En la hipótesis nula se prueban valores específicos para algunos coeficientes de la regresión (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:

Ho : $\beta_1 = \beta_j \ldots = 0$ vs. Ha : $\beta_i \neq 0$ para alguna i

Y el procedimiento con el que se verifica si se cumple la hipótesis nula (i.e. no se rechaza Ho) es el siguiente:
(Pasos que se siguen para realizar una prueba de hipótesis)

1.- Obtener los valores $\hat{Y}$ y $\hat{Y}'$ que son los valores 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 del error o de residuales denotada por SCE(MLG). Esto es:

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

en el caso del MLR tal cantidad es:

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

Para el MLG se supone que se tiene (p+1 si se considera la presencia de una constante en el modelo y p en otro caso) parámetros, y para el MLR, se consideran k parámetros.

3.- Para saber qué tan bien se adecúa el MLR, se compara

SCE(MLR)-SCE(MLG) con SCE(MLG)

utilizando para ello el siguiente cociente:

$$F = \frac{SCE(MLR) - SCE(MLG)/(p+1-k)}{SCE(MLG)/(n-p-1)} \qquad (2.6.1)$$

los divisores del divisor y del dividendo, sirven para compensar el número de parámetros que se involucran en los dos modelos, y asegurar que la estadística de prueba que se obtiene tiene distribución F, con (p+1-k) y (n-p-1) grados de libertad.

4.- El valor resultante del cociente (2.6.1), se compara con el cuantil (1-α) en tablas de una distribución F con (p+1-k) y (n-p-1) grados de libertad, y a un nivel de significancia α, si el cociente es mayor que el cuantil (1-α) de la distribución F, se rechaza la hipótesis nula, y se considera que el MLR explica satisfactoriamente la relación que existe entre las variables independientes y la variable dependiente que se estudia; en tal caso se considera que el resultado es significativo a un nivel α.

Un caso particular de las pruebas de hipótesis son las que se utilizan para la validación del modelo $\mathbf{y} = X\beta + \varepsilon$, y que se veran a continuación.

2.7 Validación del Modelo

En el planteamiento del modelo de regresión se incluyen varias suposiciones (o hipótesis de trabajo) como son:

- los errores se distribuyen normalmente y
- los errores tienen media cero y matriz de varianza covarianza $\sigma^2 I$ (es equivalente a la independencia).

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

Resulta natural la pregunta de verificar si estadísticamente existe una relación lineal entre las variables de control (independientes) y la variable observada (o dependiente).

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

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

Ho: $\beta_1 = \ldots = \beta_p = 0$ vs. Ha : Existe $\beta_i \neq 0$ para alguna i

Si la hipótesis nula (Ho) se acepta, esto indica que estadísticamente no existe evidencia de que las variables $x_1, x_2, \ldots x_p$ del modelo se asocien linealmente a la variable de respuesta **y** y en el caso contrario en el cual se rechaza la hipótesis nula se tiene que estadísticamente si existe evidencia de que las variables explicativas x_i se asocian linealmente a la variable de respuesta.

El procedimiento para llevar a cabo la prueba de hipótesis antes planteado, se efectúa a través del conocido análisis de varianza, con base en la estadística F.

Como ya se mencionó en la prueba de hipótesis para validar el modelo de regresión, se supone que:

1.- Todos los coeficientes de regresión son cero, que es equivalente a suponer que no hay una asociación lineal entre las variables independientes y la dependiente

Ho: $\beta_1 = \ldots = \beta_p = 0$ vs. Ha: Existe $\beta_i \neq 0$ para alguna i

También es posible identificar a esta prueba como un análisis de varianza sobre todo el conjunto, en el cual el modelo completo

$$y = X\beta + \varepsilon$$

es comparado al modelo sin variables , 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 grados de libertad en la SCT menos el número de grados de libertad en SCE dando como resultado p.

Tabla de Análisis de Varianza (Ho: $\beta = 0$)

Fuente de Variación	Grados de libertad	Sumas de Cuadrados	Cuadrados Medios	F
Regresión	p	SCR	SCR/p	$F = \frac{SCR/p}{SCE/(n-p')}$
Residual	n-p'	SCE	SCE/(n-p')	
Total	n-1	SCT		

p' = p+1 si hay en el modelo p variables y una constante.

Si la estadística F es mayor que el cuantil $(1-\alpha)$ de la distribución F obtenida en las tablas al nivel de significancia α, se juzgará que el conocimiento de las x's provee un modelo significativamente mejor que aquel que no las toma en cuenta. Por lo que se rechaza Ho.

Por lo mencionado en los dos párrafos anteriores ésta prueba se utiliza si se desea validar el modelo que se encuentra al estimar las $\hat{\beta}$'s del modelo lineal general, es decir verificar qué tan bien se ajustan los datos al modelo encontrado, y se efectúa un análisis de varianza, el cual se acostumbra presentar en

tablas conocidas como tablas de análisis de varianza.

Ejemplo 2.7.1 Para el ejemplo 1.4.1 se tiene la siguiente tabla:

ANALISIS DE VARIANZA POR AJUSTE DE REGRESION (Ho: $\beta = 0$)

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión	2	SCR = 62.5	31.25	F = 5.43
Residual	2	SCE = 11.5	5.75	
Total	4	SCT = 74.0		

al comparar el valor de F = 5.43 con el de la distribución $F_{2,2}$=19.00, con un nivel de significancia del 5%, se concluye que el modelo no explica la variación de la variable **y**, ésta prueba tiene un nivel de significancia del 5%.

Cuando se rechaza la hipótesis Ho: β_i=0 i=1,...p, donde p es el número de variables, como ocurre en los siguientes ejemplos, lo que se puede pensar es que el modelo lineal es adecuado, o que estadísticamente existe una relación lineal entre las variables de control (independientes) y la variable observada (o dependiente).

Ejemplo 2.7.2 Tabla de Análisis de Varianza para el ejemplo 2.1.1 para la prueba la hipótesis

Ho: β_i=0 i =0,4 vs. Ha: $\beta_j \neq 0$ para alguna j

es la siguiente:

Fuente de Variación	Grados de libertad	Sumas de Cuadrados	Cuadrados Medios	F
Regresión	4	SCR = 399.316	99.829	F = 22.70
Residual	43	SCE = 189.050	4397.000	
Total	47	SCT = 588.366		

Ya que F = 22.7 excede a $F^{.99}_{4,43}$ = 3.79, se puede decir que el consumo de combustible realmente es explicado por alguna(s) variable(s) de la(s) considerada(s) en el modelo.

Una segunda prueba de hipótesis que se mencionará en este trabajo es en la que un subconjunto de los coeficientes de regresión son cero y esta prueba es conocida como prueba F parcial.

2.8 Prueba F Parcial

En algunas ocasiones el observador o investigador considera que un subconjunto de las variables que interviene en el modelo planteado originalmente no lo hacen en forma significativa por lo cual pueden ser descartadas (igualandolas a cero); para verificar esta suposición se recurre a una prueba de hipótesis conocida como Prueba F Parcial.

Al probar que un subconjunto de los coeficientes de la regresión son cero, lo que se trata de hacer es descartar aquellas variables que pareciera que puede no considerárseles en el modelo, es decir que si no se les toma en cuenta, la variable y de cualquier manera es explicada por las variables consideradas; para algunos investigadores esto es de suma importancia ya que consideran que una de las principales metas

del análisis de regresión es "obtener una descripción del fenómeno observado en términos de las menos variables posibles" [Montgomery pag.255], ya que de esta forma únicamente consideran las variables más importantes dentro del modelo, y al manejar uno pequeño es más fácil su comprensión; sin embargo al aplicar este criterio no se debe llegar al grado de que al eliminar variables del modelo éste deje de ser representativo del comportamiento de la variable **y**, es por ello que para determinar si realmente es posible eliminar una variable del modelo existen métodos conocidos como de selección de variables, uno de los cuales será mencionado en el capítulo III.

Después de realizar la prueba de hipótesis anterior y no rechazarse, esto implica que las (p-k) variables que están en el modelo lo explican tan bien como el conjunto inicial de variables que era mayor en número; al proceder con un análisis de residuales del modelo reducido, si no se observa alguna violación a las suposiciones del modelo, es posible concluir que el modelo no se ve afectado al no considerar en él a las variables que supusimos que tomaban el valor de cero en la prueba de hipótesis.

Para el ejemplo 2.1.1 se tiene :

Ho: $\beta_i = 0$: β_i arbitrarias i = 0,2,3,4

Ha: $\beta_i \neq 0$; β_i arbitrarias i = 0,2,3,4

Si se desea saber que tan significativo es TCC, después de ajustar la regresión sobre las variables PPLC, RM e IP se ajusta el modelo que incluye TCC y otro que no lo incluye y se obtienen las sumas de cuadrados de los residuales para ambos, y se obtiene la suma de cuadrados por regresión de TCC después de ajustar las variables incluídas en el modelo, que es la diferencia entre la suma residual de cuadrados para el modelo que

no incluye a TCC, y la suma de cuadrados del modelo que si incluye a TCC. La estadística F es la prueba utilizada para verificar la utilidad de TCC después de que las otras variables han sido incluídas en el modelo, esta prueba no dice nada sobre la utilidad de las otras variables.

Ejemplo 2.8.1

Tabla de análisis de varianza para la Prueba F Parcial (Ho: $\beta_1=0$)

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión sobre (RM ,IP Y PPLC)	3	SCR = 367.684	122.561	
TCC después de otros	1	SCE = 31.632	31.632	7.19
Residual	43	SCT = 189.050	4.397	

Comparando el valor de la estadística F = 7.19 con el de la distribución $F^{99}_{1,43} = 7.26$, TCC parece ser un predictor significativo después de haberse ajustado por las otras variables.

Si la estadística β_1 no es estadísticamente diferente de cero, esto no necesariamente indica que no existe una relación entre las variables X_1 y **y**, ya que ésta al efectuar mediciones inexactas puede quedar oculta.

2.9 Colinealidad

Una de las suposiciones más importantes en el análisis de regresión, es que las variables x_i's son independientes entre sí o que al menos no están fuertemente relacionadas, si tal suposición no se cumple, el cambio que ocurra en alguna variable puede implicar cambios en otra variable, 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 dependiente cuando la variable explicativas x_i es incrementada en una unidad, y las variables restantes permanecen constantes" [Montgomery 110]; y esto como se mencionó no ocurre necesariamente.

Si se observa que no existe alguna relación entre las variables, se dice que son independientes, sin embargo en la práctica en la mayoría de los casos estudiados no son independientes las variables y entonces se considera que existe multicolinealidad o datos colineales, pero esto no llega a afectar seriamente el análisis; para detectar la dependencia que pudiera presentarse en las variables, se tomarán en cuenta los siguientes criterios:

1.- En la matriz $C = (X'X)^{-1}$, los c_{ij}, $i \neq j$, deben ser en valor absoluto lo más lejanos a 1.

2.- El determinante de la matriz $X'X$ debe estar alejado de 0.

3.- Las subrutinas de Linpack proporcionan diagnósticos numéricos sobre que tanto se acerca a la ortogonalidad la matriz de observaciones X.

4.- Grandes cambios en los coeficientes estimados cuando se agrega o se elimina una variable.

5.- Grandes cambios en los coeficientes cuando una observación es alterada o borrada.
Cuando la graficación de los residuales indica que el modelo es el adecuado, la multicolinealidad puede estar presente si:

6.- Los signos algebráicos de los coeficientes estimados no corresponden a los esperados.

7.- Los coeficientes de las variables que se espera sean importantes tienen errores estándar grandes.

2.10 Resumen Estadístico

A continuación se presenta un resumen estadístico básico en el que se pueden observar el promedio, varianza, desviación estándar, los valores mínimos y máximos de cada variable.

Con los datos del ejemplo 2.1.1.

Variable	n	Promedio	Varianza	Desviación Estándar	Valores Mínimos	Máximos
TCC	48	7.6683	00.90396	0.95077	5.000	10.000
PPLC	48	57.033	30.77000	5.54700	45.100	72.400
IP	48	4.2418	00.32904	0.57362	3.063	5.342
RM	48	5.5654	12.19100	3.49150	.431	17.782
CCM	48	576.77	12518.00	111.89000	344.000	968.00

2.10.1 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
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 pag.10, es costoso en tiempo de máquina por el número de operaciones que realiza, 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].

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

$$y = X\beta + \varepsilon.$$

incluye las estadísticas:
- La estimación $\hat{\beta}$
- La desviación estándar de $\hat{\beta}$
- La estadística F
- El Coeficiente de Determinación R^2

- Matriz de $(X'X)^{-1}$
- Matriz de Correlación

En el siguiente ejemplo se vera la aplicación de las estadísticas mencionados en esta sección.

Ejemplo 2.10.1 [Montgomery]

Se analizan las rutas del sistema de distribución de una bebida. Interesa predecir la cantidad de tiempo que requiere el conductor de la ruta para efectuar la entrega. Un ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan al tiempo de entrega son el número de casos que el producto se almacena y la distancia que camina el conductor de la ruta.

El modelo de regresión que se propone es :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Número de Observación	Tiempo de Entrega (Min) y	Numero de Casos x_1	Distancia (Pies) x_2
1	7	560	16.68
2	3	220	11.50
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8.00
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.50
11	16	688	40.33
12	10	215	21.00
13	4	255	13.50
14	6	462	19.75
15	9	448	24.00
16	10	776	29.00
17	6	200	15.35
18	7	132	19.00
19	3	36	9.50
20	17	770	35.10
21	10	140	17.90
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

Matriz $(X'X)^{-1}$

$$\begin{bmatrix} .113215 & -.004448 & -.000083 \\ -.004448 & .002743 & -.000047 \\ -.000083 & -.0004 & .000001 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.34123 \\ 1.61590 \\ 0.01438 \end{bmatrix}$$

$s^2 = 10.6239$ Estadística F : 261.24 $R^2 = 0.9596$

C A P I T U L O III

Análisis de los Datos y las Variables que Intervienen en el Modelo de Regresión Lineal

Introducción

Este capítulo es introductorio a los siguientes temas:

a) El análisis de los residuales que se obtienen al estimar los parámetros β's de el modelo lineal

$$y = X\beta + \varepsilon$$

vía la graficación de los mismos

b) La influencia de las observaciones mediante la estadística conocida como distancia de Cook

c) Los criterios para seleccionar las variables que intervienen en el modelo lineal.

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 no observables son independientes, 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 las variables independientes es lineal.

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

$$e_i = y_i - \hat{y}_i \qquad rs_i = e_i / se(\hat{\beta})$$

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

A continuación se indicarán las gráficas que comúnmente se acostumbra realizar.

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

Al graficar los residuales estandarizados y los valores ajustados por el modelo (i.e. $\hat{y}_i$), se verifica el supuesto de homocedasticidad (i.e. $\varepsilon \sim N(0,\sigma^2 I)$), y se detectan datos aberrantes como en el siguiente párrafo se menciona.

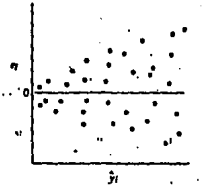
En general, cuando el modelo es correcto los residuales se encuentran distribuídos de manera aleatoria alrededor del cero, y toman valores dentro del intervalo (-2,2); esto tiene su justificación en el hecho de que la probabilidad de que el valor de una variable aleatoria que tiene una distribución normal estándar se encuentre en el intervalo de (-2,2) es del 95%, por lo que al graficar los residuales estandarizados deben estar en este intervalo, si un residual estandarizado presenta un valor significativamente alejado de este intervalo, uno estaría inclinado a pensar que este residual no corresponde a la población y por ello resulta razonable considerarlo como un dato aberrante o "outlier".

Ejemplo 3.1.1.a Gráfica que muestra un modelamiento adecuado

Si existe un patrón sistemático de variación en los residuales, se puede considerar como evidencia de una violación en uno de los supuestos del modelo, que es el de la varianza constante en la distribución de los residuales (i.e. $\varepsilon \sim (0, \sigma^2 I)$).

A la suposición de varianza constante en los errores del modelo lineal, se le conoce como suposición de homocedasticidad; cuando no se cumple tal suposición se dice que hay heterocedasticidad o que los errores son heterocedásticos.

Ejemplo 3.1.1.b Gráfica que muestra un patrón de variación en los residuales



En la gráfica b), se observa que la varianza en 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

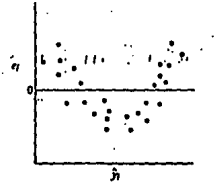
los errores se incrementa al decrecer los valores de la variable de respuesta **y**.

En el caso de que la varianza de los errores no sea constante, los resultados de las pruebas de la teoría no serían de utilidad, ya que si bien los estimadores obtenidos serían insesgados, no serían de los mejores en cuanto a su precisión o varianza.

A continuación se presenta un resumen de las consecuencias que provoca el que la varianza en los errores no sea constante:
1.- Los estimadores tendrán grandes desviaciones estándar.
2.- Como consecuencia de 1, los intervalos de confianza para los parámetros serán grandes.
3.- Las pruebas tendrán baja sensibilidad.

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

Ejemplo 3.1.1.c Gráfica que muestra un patrón de variación en los residuales



En la gráfica c), lo que se puede observar es que al mostrar una tendencia curva los puntos, el comportamiento de la variable **y** no es lineal con respecto a las variables independientes x_1.

En un modelo con una sola variable es recomendable graficar la variable independiente vs. la variable de respuesta para observar si la relación que existe 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 las variables independientes x_1, 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 estandarizados y las variables independientes x_1.

En la siguiente sección se menciona una gráfica que permite verificar si el supuesto de normalidad se cumple.

3.1.2 Graficación en papel Normal de los Residuales

El procedimiento que a continuación se describe se utiliza para verificar el supuesto de que los errores tienen una distribución normal.

1.- Se ordenan los residuales en forma ascendente
2.- Se le asigna a cada residual la probabilidad acumulada $P_i=(i-1/2)/n$ $i=1,n$ donde n es el número de observaciones.
3.- Se grafican en papel normal la probabilidad que se les asignó a los residuales en el paso 2 vs. los residuales.

Si los puntos resultantes se encuentran aproximadamente sobre una línea recta el supuesto de normalidad se cumple.

3.2 Influencia de las Observaciones

Es posible que ciertas observaciones influyan de manera importante en la regresión, por lo tanto 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 leveraje del i-ésimo caso , los residuales studentizados y la distancia de Cook.

El leverage o potencia del i-ésimo caso esta dado por:

$$h_{ii} = x_i' \, (X'X)^{-1} \, x_i$$

e indica el efecto del i-ésimo caso en la regresión.

Los residuales studentizados se encuentran definidos por:

$$rt_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

y esta estadística si el modelo es correcto, deben tomar una distribución t con n-p grados de libertad y si esta cantidad es grande (mayor que 2) la observación se investiga.

Nota: Son diferentes los residuales estandarizados y los Studentizados.

Como se menciona en Montgomery (pag.163) "es conveniente considerar la ubicación del punto y la variable de respuesta al medir la influencia" de los casos, así se origina otra estadística de importancia además de las dos mencionadas previamente que es la distancia de Cook , definida por

$$D_i = \frac{(\hat{\beta}_{(i)}-\hat{\beta})(X'X)(\hat{\beta}_{(i)}-\hat{\beta})}{p\ s^2}$$

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

Para calcular D_i es posible utilizar la siguiente expresión en forma alternativa,

$$D_i = \frac{rt_i^2 * h_{ii}}{p\ (1-h_{ii})}$$

Donde D_i es el producto del cuadrado del i-ésimo residual Studentizado 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 estadísiticas D_i, e_i y h_{ii} para cada caso.

Al decidir eliminar una observación por tener ésta una influencia preponderante con respecto a las demás, se modifican 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.

En el ejemplo 2.1.1 se tiene que para el caso cuarenta que es Wyoming:

$$rt_{40} = 242.6 / (\sqrt{5796.31} * (1-0.0930)) = 3.3453$$

$$D_{40} = ((3.3453)^2/\ 3\ ) * (0.0930\ /\ (\ 1\ -\ 0.0930)) = 0.3826$$

En este caso, se observa que es relativamente importante, ya que su valor de r_i es grande, comparado con los demás.

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

3.3 Selección de Variables

Cuando se tiene un modelo lineal de regresión, existen principalmente dos motivos por los que se efectúa una selección de las variables que deben considerarse en tal modelo. 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 variables que expliquen a la variable de respuesta **y**.

Además de la dos razones mencionadas de una manera informal Montgomery [5] 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. las variables consideradas en el modelo ¿ tienen sentido con respecto al problema ?)

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

3.- ¿ Las β'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 usual 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, y para poder decidir cual es el mejor se utiliza la estadística conocida como C_p de Mallows, la cual tiene su orígen en el criterio siguiente:

Los modelos de regresión que proporcionan mejores predicciones de la variable de respuesta , y por lo tanto residuales más pequeños (esto implica que la estimación de la varianza s^2 es pequeña) son los mejores, es por ello que el cálculo de C_p, se basa en la suma de cuadrados de residuales y en la estimación de la varianza s^2, como es posible observar en las siguientes expresiones que lo definen:

$$\begin{aligned} C_p &= SCE_p/s^2 + 2p - n \\ &= (SCE_p - SCE_{k'})/ s^2 + p - (k' - p) \\ &= (k' - p)(Fp - 1) + p \end{aligned}$$

donde se suponen dos modelos el primero con p variables al que se denomina modelo lineal general (MGL), y otro cuyas variables (k') es un subconjunto de las consideradas en el primero al que se conoce como modelo lineal reducido (MLR), s^2 es la estimación de la varianza del MLG, F_p es la estadística usual F para la hipótesis en que se supone que las k' variables incluídas en el MLR tienen coeficiente cero, y n es el número de observaciones.

La estadística C_p es importante por tener las siguientes características:

a) De la primera expresión se deduce que C_p depende únicamente de cálculos usuales de regresión como son SCE_p, s^2, p , y n, que son

fácilmente calculados. Es por esta razón que C_p es obtenida rápidamente en los algoritmos en que se obtienen todas las posibles regresiones.

b) De la segunda expresión se deduce que C_p mide la diferencia en los residuales entre el modelo completo y los modelos reducidos.

c) En la última expresión se observa que C_p consiste de dos partes una aleatoria F_p y una fija p.

d) Dos modelos reducidos del modelo lineal general pueden ser comparados al considerar sus valores de C_p, Mallows ha sugerido que los buenos modelos tienen $C_p \approx p$. Ya que C_p es una variable aleatoria, dos modelos no son fácilmente distinguibles si sus valores de C_p son muy cercanos. Cualquier modelo donde $F_p<=2$, lo cual implica que $C_p=k'$, será un buen candidato para un mejor modelo de regresión reducido.

Un buen modelo de regresión implica valores mínimos de sumas de cuadrados de residuales y de C_p.

Los valores de C_p primero decrecen cuando los regresores son agregados al modelo, y entonces alcanza un mínimo y empieza a incrementarse.

Los regresores que contribuyen significativamente en el modelo completo tendrán un valor alto de $|t_{k,j}|$ y serán incluídos en el "mejor" modelo de regresión, que es aquel en el que los valores de la suma de cuadrados de residuales y de la estadística C_p son mínimos. Es por esto que ordenando las variables de manera decreciente de acuerdo a los valores de la estadística $|t_{k,j}|$ donde $j=1,2,\ldots,K$, se introduce en el modelo un regresor a la vez en este orden permite encontrar el mejor o uno de los mejores modelos reducidos para cada p.

Daniel y Wood [ver Montgomery] a esta forma de seleccionar las variables que se incluyen en un modelo le llaman procedimiento de búsqueda directa sobre t. El empleo de este procedimiento es frecuentemente efectivo cuando el número de candidatos es relativamente grande por ejemplo K>20.

A continuación se presenta un resumen de los criterios que permiten elegir el mejor modelo lineal reducido con k variables, donde k es determinado en forma arbitraria.

1.- Suma de cuadrados de residuales más pequeña

2.- R^2 el coeficiente de correlación múltiple más cercano a 1

3.- Coeficiente ajustado de correlación múltiple más cercano a 1, definido de la siguiente manera:

$$R^2a = (1 - (n-1)(1-R^2)/(n-p)), \quad \text{donde } p=k+1$$

el cual es un promedio de la varianza sobre los datos observados.

4.- $C_p = SCE/s^2 + 2p - n, \qquad \text{donde } p=k+1$

C A P I T U L O IV

Análisis de Regresión Lineal Vía la Descomposición QR

Introducción

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

$$y = X\beta + \varepsilon \qquad (4.1.1)$$

bajo el supuesto básico de que $X \in \mathbb{R}^{nxp}$ es de rango máximo y que el vector de errores ε tiene media 0, y matriz de varianza - covarianza $\sigma^2 I$.

La herramienta básica es la descomposición QR de X, a partir de la cual se calculan las estimaciones Gauss-Markov $\hat{\beta}$ para β y s^2 para σ^2, así como su correspondiente vector residual $e(= y-X\hat{\beta})$ y su correspondiente matriz de varianzas y covarianzas $\sigma^2(X'X)^{-1}$ [o $s^2(X'X)^{-1}$] entre otros. Con base en estos cálculos se efectúan otros más para efecto de llevar a cabo el análisis estadístico del modelo-datos mediante pruebas de hipótesis ya vistas en los capítulos anteriores.

4.1 Cálculo de los estimadores de los parámetros β's mediante las ecuaciones normales

Como se ha mencionado antes, la estimación $\hat{\beta}$ mínimos cuadrados para β en el modelo lineal (4.1.1) es tal que resuelve el problema :

$$\min_{\tilde{\beta}} \| y - X\tilde{\beta} \|_2$$

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

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

Su correspondiente derivada con respecto a $\tilde{\beta}$ es:

$$-2X'y + 2X'X\beta \qquad (4.1.2)$$

Así si $\hat{\beta}$ es un punto mínimo de $\left\| y - X\tilde{\beta} \right\|_2$ entonces su derivada para $\tilde{\beta} = \hat{\beta}$ debe ser igual a 0. O sea $\hat{\beta}$ debe resolver el sistema de ecuaciones lineales

$$X'X\tilde{\beta} = X'y \qquad (4.1.3)$$

conocido como ecuaciones normales.

El cálculo de $\hat{\beta}$ mediante la solución del sistema de ecuaciones normales (4.1.3) con ayuda de una computadora no es

recomendable, ya que implica el cálculo previo de la matriz (X'X), lo cual aparte de ser costoso, puede resultar inconveniente, pues bien X'X (siendo $X \in \mathbb{R}^{n\times p}$, $n > p$, 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 $fl(1+\varepsilon)=1$). Y que

$$X'X = \begin{bmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{bmatrix} \approx \begin{bmatrix} 1 & 1 & 1 \\ 1 & . & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

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

Por otro lado, es bien sabido que el método ideal para resolver un sistema de ecuaciones (algebráicas) lineales con matriz simétrica y positiva definida, debido a sus propiedades de estabilidad numérica y de economía en cuanto a requerimientos de memoria y de cantidad de cómputo, es el método de Cholesky. Así,

bajo el supuesto de que $X \in \mathbb{R}^{n \times p}$, $n \geq p$, tiene rango máximo, se tiene que $X'X$ es simétrica y positiva definida [8]. Por ello, el método ideal para resolver numéricamente las ecuaciones normales (4.1.3) es el método de Cholesky, el cual consiste en hallar la descomposición de Cholesky de $X'X$, i.e.

$$X' X = R' R$$

siendo R triangular superior con diagonal positiva.

Con base en esta descomposición, la resolución numérica de las ecuaciones normales (4.1.3) se reduce a resolver dos sistemas de ecuaciones triangulares:

$$R'z = X'\underline{y} \qquad \text{y} \qquad R\tilde{\beta} = z$$

El análisis de error retrospectivo para el método de Cholesky [3], indica que

$$\frac{\left\| \hat{\beta}^* - \hat{\beta} \right\|_2}{\left\| \hat{\beta} \right\|_2} \leq C_n K_2 (X'X)\, u + O(u^2) \qquad (4.1.4)$$

en donde $\hat{\beta}^*$ es la solución numérica y $\hat{\beta}$ es la solución exacta de las ecuaciones normales (4.1.3), $C_n \sim n^{3/2}$, u es la unidad de redondeo, y

$$K_2(X'X) = \left|X'X\right|_2 \left|(X'X)^{-1}\right|_2 \qquad (4.1.5)$$

es el número de condición de X'X.

Así pues, la exactitud de la solución numérica $\hat{\beta}^*$ de las ecuaciones normales (4.1.3) está fundamentalmente determinada por el número de condición $K_2(X'X)$.

En términos de la inversa generalizada Moore-Penrose X^+ de X, definida como la solución respecto de $Y \in \mathbb{R}^{pxn}$ del sistema de ecuaciones matriciales:

(MP.1) $XYX = X$,

(MP.2) $YXY = Y$,

(MP.3) $(XY)' = XY$, y

(MP.4) $(YX)' = YX$;

el concepto de número de condición se puede extender; como es el caso de estudio, a matrices rectangulares.

En efecto, dada $X \in \mathbb{R}^{nxp}$ por el número de condición $K_2(X)$ se entiende

$$K_2(X) = \left|X\right|_2 \left| X^+ \right|_2$$

Para la situación bajo discusión ($X \in \mathbb{R}^{nxp}$, $n > p$, rango(X)= p), es directo verificar que

$$X' = (X^+X)^{-1}X^+$$

Ahora, con base en la descomposición en valores singulares de X (vease [3]):

$$X = U\,S\,V'$$

en donde $U \in \mathbb{R}^{nxn}$ y $V \in \mathbb{R}^{pxp}$ son ortogonales, y

$$S = \begin{bmatrix} D \\ \hline 0 \end{bmatrix} \in \mathbb{R}^{nxp}$$

con

$$D_p = \text{diag}(S_1, S_2, \ldots, S_p),$$

$$S_1 \geq S_2 \geq \ldots \geq S_p > 0.$$

Es directo verificar que

$$K_2(X'X) = [K_2(X)]^2 \qquad (4.1.7)$$

De esta igualdad y la relación (4.1.4) se sigue que en una microcomputadora AT de 32 bits con doble precisión (i.e. $\approx 10^{-16}$), si $K_2(X) \propto 10^8$ entonces carece de sentido resolver las ecuaciones normales (4.1.4), pues de (4.1.4) se tiene que

$$\frac{\left\| \hat{\beta}^* - \hat{\beta} \right\|_2}{\left\| \hat{\beta} \right\|_2} \propto 1.$$

A pesar de haber aplicado el método ideal de la descomposición de Cholesky.

Recapitulando, la resolución numérica de las ecuaciones normales (4.1.3) para calcular la estimación Gauss-Markov $\hat{\beta}$ para $\hat{\beta}$ (estimación bajo la aplicación del criterio de mínimos cuadrados) tiene dos serios inconvenientes:

1o) El cálculo previo (el cual requiere de $O(np^2)$ 'flops' - véa Golub-Van Loan para la definición de flop) de la matriz

$$X'X,$$

la cual puede resultar numéricamente singular aún cuando $X \in \mathbb{R}^{nxp}$ sea numéricamente de rango p. Y,

2o) Si $K_2(X) \approx (\sqrt{u})^{-1}$, 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 (4.1.3), aún cuando estas sean resueltas por el método de Cholesky.

Una alternativa a la resolución del problema (4.1.2) de mínimos cuadrados vía la resolución numérica de sus ecuaciones normales (4.1.3) 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, ella no requiere del cálculo de $X'X$, y bajo ciertas condiciones adicionales muy razonables desde el punto de vista práctico, su sensibilidad numérica queda determinada por $K_2(X)$ en vez de $K_2(X'X)$.

4.2 Descomposición QR de la Matriz X

Son de gran importancia para el tratamiento de éste tema las matrices ortogonales, ya que a partir de cierto tipo de las mismas 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 es una matriz Q de n x n que satisface que $Q'Q = I = QQ'$, donde I es la matriz identidad de n x n, éste tipo de matrices es importante porque preserva la norma euclidiana o norma 2 ya que :

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

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.

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

$$H = I - 2\, v\, v'/v'v \qquad 4.2.1$$

donde v es un vector de n x 1 distinto del vector cero. Estas matrices son simétricas y ortogonales. Esta última característica se verifica a continuación:

$$H'H = HH$$
$$= (I - 2\, v\, v'/v'v)\ (I - 2\, v\, v'/v'v)$$

$$= I - 4\, v\, v'/v'v + 4\, v\, v'/v'v \;\; v\, v'/v'v$$

$$= I - 4\, v\, v'/v'v + 4\, v(\, v'v\,)v'/(v'v)2$$

$$= I - 4\, v\, v'/v'v + 4\, v\, v'/v'v$$

$$= I \quad .$$

En particular se tiene que:

$$H = H' = H^{-1}.$$

Para cualquier vector $a = (a_1, \ldots a_n)'$ no nulo para el cual no existe $t \in \mathbb{R}$ t.q. $a = te_1$, existe un vector v (no nulo) que permite construir una reflexión de Householder que al ser aplicada al vector a, todas las componentes del vector resultante sean cero a excepción de la primera, i.e.

$$Ha = \begin{bmatrix} \alpha \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

y debido a que las transformaciones ortogonales preservan la norma-2, se tiene que $\alpha = \pm \| a \|_2$.

Si a es el vector que se desea transformar lo que se quiere es que:

$$Ha = \alpha e_1$$

o bien que

$$Ha = (I - 2\, v\, v'/v'v)a$$

$$= a - (\, 2v'a/v'v)v$$

$$= \alpha\, e_1$$

o sea que

$$2(v'a/v'v)v = a - \alpha e_1$$

de ésta expresión se deduce que v es un múltiplo de $a - \alpha e_1$, y debido a que se sabe que si v es multiplicada por una constante distinta de cero la reflexión de Householder que se obtiene a partir de v no cambia, esto permite elegir a v como $v = a - \alpha e_1$ y al sustituir a α por $\pm \| a \|_2$ se tiene que:

$$v = a \pm \| a \|_2 e_1$$

en donde el signo se elige como sigue

$+$ si $a_1 > 0$.

$-$ si $a_1 < 0$,

por razones de estabilidad numérica (i.e. se trata de evitar cancelación numérica).

El procedimiento para obtener la factorización QR de la matriz X utilizando las reflexiones de Householder es el siguiente:

1.- Se supone que el vector a es la primera columna de la matriz X de n x p y se construye una primera reflexión de Householder (H_1) que al aplicarla a la matriz X del modelo lineal (4.1) se vuelven cero los elementos del vector a con excepción del primero es decir,

$$H_1X = \begin{bmatrix} \alpha & x \ldots x \\ 0 & x \ldots x \\ . & x \ldots x \\ . & x \ldots x \\ . & x \ldots x \\ 0 & x \ldots x \end{bmatrix}$$

2.- Se supone que el vector a está constituído por los n-1 últimos elementos de la segunda columna de la matriz H_1X y se construye una segunda reflexión de Householder (H_2^*) con la que se forma la matriz

$$H_2 = \begin{bmatrix} 1 & 0 \\ 0 & H_2^* \end{bmatrix}$$

que al aplicarla a la matriz H_1 vuelve cero los elementos del vector a con excepción del primero es decir,

$$H_2 H_1 X = \begin{bmatrix} 1 & 0 \\ 0 & H_2^* \end{bmatrix} H_1X = \begin{bmatrix} \alpha & x & .. & x \\ 0 & \alpha & .. & x \\ . & 0 & x. & x \\ . & . & .. & x \\ . & . & .. & x \\ . & . & .. & x \\ 0 & . & x. & x \end{bmatrix}$$

En el paso i se elige al vector a como el vector constituído por los n-(i+1) últimos elementos de la i-ésima columna de la matriz $H_{(i-1)} \ldots H_1X$, que se obtuvo en el paso anterior y nuevamente se construye una transformación de Householder H_1^* con la que se forma la matriz:

$$H_i = \begin{bmatrix} I_{i-1} & \\ & H_i^* \end{bmatrix}$$

donde I_{i-1} es la matriz identidad (i-1) x (i-1).

que al aplicarla a la matriz que se obtuvo en el paso i-1, se vuelven cero los elementos del vector a con excepción del primero, este procedimento se aplica p ocasiones al cabo de las cuales se reduce a la matriz X a una "triangular superior":

$$\underset{n \times p}{\begin{bmatrix} R \\ 0 \end{bmatrix}} \quad \text{con} \quad R = \begin{bmatrix} \quad \end{bmatrix}$$

Si al producto final de las reflexiones H_i se le llama Q' i.e.

$$Q' = H_p \, H_{p-1} \ldots H_1$$

entonces se concluye que:

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

o bien que :

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

Esta descomposición para X es la bien conocida descomposición QR de la matriz X.

Si se particiona la matriz Q como $Q = (Q_1, Q_2)$ donde Q_1 tiene p columnas, entonces

$$X = \left[Q_1, Q_2\right]\begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R$$

Esto es, se tiene que

$$X = Q_1 R \qquad (4.2.2)$$

Ahora sustituyendo ésta descomposición (4.2.2) en las ecuaciones normales:

$$X'X \tilde{\beta} = X'y$$

Se tiene que

$$R'Q_1Q_1R \tilde{\beta} = R'Q_1'y$$

o bien que

$$R'R \tilde{\beta} = R'Q_1' y$$

así que

$$R \tilde{\beta} = Q_1' y \qquad (4.2.3)$$

En resumen, la resolución del problema de minímos cuadrados

$$\min_{\tilde{\beta}} \frac{1}{2} || y - X\tilde{\beta} ||_2$$

en términos de la descomposición QR de la matriz X,

$$X = Q_1 R,$$

se reduce a efectuar los siguientes dos pasos:

1º . Se calcula:

$$Q_1' y \qquad (4.2.3)$$

2o. Se resuelve el sistema triangular superior

$$R\,\hat{\beta} = Q_1'\,y \qquad (4.2.3)$$

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 (4.2.3) se tiene que :

$$\hat{\beta} = \begin{bmatrix} R^{-1} \\ 0 \end{bmatrix} Q'y = [R^{-1}\ 0]Q'y$$

Esto implica que :

$$X^{+} = \begin{bmatrix} R^{-1} & 0 \end{bmatrix} Q'$$

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

Por lo que, es directo verificar , que

$$K_2(R) = K_2(X)$$

En consecuencia, bajo el supuesto de la hipótesis del ángulo agudo (i.e. que el ángulo formado por y y $im(X)$ es no mayor de $\pi/4$), la cota (a priori) del error relativo está esencialmente dada por

$$\frac{\left\| \hat{\beta}^{*} - \hat{\beta} \right\|_2}{\left\| \hat{\beta} \right\|_2} \leq K_2(R)\,\frac{\left\| \delta R \right\|_2}{\left\| R \right\|_2} \approx K_2(X)\,\frac{\left\| \delta \cdot X \right\|_2}{\left\| X \right\|_2}$$

(vease Bulirsch R. Stoer J., "Introduction to Numerical Analysis", Springer (1980))

Por lo tanto, en general si X de rango máximo 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 (4.1.3) relacionadas al sistema.

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

3) La sensibilidad numérica del sistema (4.2.3) es la de X, y no la de X'X como ocurre con las ecuaciones normales.

4.3 Otros Usos de la Descomposición QR de la Matriz X

El paquete de programas en FORTRAN profesionales LINPACK, cuenta con rutinas que calculan la descomposición QR de la matriz

X, estiman los parámetros $\hat{\beta}$ del modelo lineal, los valores de predicción $\hat{y}$ de las observaciones y el vector de residuales que se denota con el vector **e**.

A partir de los residuales es posible :

obtener:

a) la estimación de s^2 para la varianza, $s^2 = \frac{e'e}{n-p} = \frac{SCE}{n-p}$

b) la estadística F, $F = \frac{SCR/p}{SCE/(n-p)}$

c) el coeficiente de determinación R^2, $R^2 = \frac{SCR}{SCT}$

d) los residuales estandarizados, $rs_i = \frac{e_i}{s}$

e) los residuales studentizados, $rt_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$, donde

$h_{ii} = (X'(R'R)^{-1}X)_{ii}$

f) la estadística C_p de Mallows, $C_p = SCR/s^2 + 2p - n$, donde $p=k+1$

y graficar :

g) los residuales estandarizados vs. las $\hat{y}_i$ para determinar si hay observaciones aberrantes,

h) los residuales estandarizados vs. las $\hat{y}_i$ para determinar si el modelo cumpe con ciertas suposiciones como son la homocedasticidad en los errores, y la distribución normal de los mismos.

Estas rutinas permiten calcular también la matriz $(R'R)^{-1}$ que es exactamente la matriz $(X'X)^{-1}$ debido a que de la descomposición QR de X

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

se tiene que

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

$$= \begin{bmatrix} R' & 0 \end{bmatrix} I \begin{bmatrix} R \\ 0 \end{bmatrix}$$

$$= R'R$$

Por lo tanto

$$(X'X)^{-1} = (R'R)^{1-}$$

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

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

b) el coeficiente de inflación de la varianza para cada $\hat{\beta}_i$, determinado por r_{ii},

c) la potencia de cada observación, $h_{ii} = (X'(R'R)^{-1}X)_{ii}$

d) las distancias de Cook, $D_i = \frac{1}{p} rt_i \frac{h_{ii}}{(1 - h_{ii})}$

Es por estas razones que el sistema ANA_R ELI.SIS 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

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

$$\bar{x} = \sum_{i=1}^{m} x_i \; / \; m \qquad y \qquad s^2 = \frac{1}{m-1} \left(\sum_{i=1}^{m} x_i^2 - m \, \bar{x}^2 \right)$$

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

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

Una alternativa es utilizar algoritmos que sean menos costosos y en los que se pueda conocer una cota del error que se produce en el resultado, uno de ellos es el desarrollado por West-Hanson [1] y que a continuación se menciona.

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

$M_1 = x_1$, $V_1 = 0$

Para i=2,m haz

$$V_i = V_{i-1} + \frac{i-1}{i} (x_i - M_{i-1})^2 \qquad M_i = \frac{(i-1)M_{i-1} + x_i}{i}$$

A continuación se presenta un ejemplo en el que aplican los tres algoritmos mencionados, el análisis de error correspondiente a cada uno de ellos se encuentra en el apéndice C de este trabajo, el cual se decidió incluir debido a la escasa difusión en la literatura.

Ejemplo: En una Aritmética de P.F. de 3 decimales. Sean $x_1=15$, $x_2=16$, $x_3=17$, $x_4=18$ y $x_5=19$; al aplicar los algoritmos anteriores se obtienen los resultados siguientes:

Si se calcula la media como $\bar{x}=\sum_{i=1}^{m} x_i/m$, para el ejemplo, se tiene que $fl(\bar{x}) = \bar{x}$, debido a que la suma de las x_i no excede de 999.

a) Con el algoritmo de dos pasos se tiene lo siguiente:

$s^2_0 = (15-17)^2 = 4$ $\qquad$ $s^2_1 = 4 + (16-17)^2 = 5$

$s^2_2 = 5 + (17-17)^2 = 5$ $\qquad$ $s^2_3 = 5 + (18-17)^2 = 6$

$s^2_4 = 6 + (19-17)^2 = 10$ $\qquad$ $s^2 = 10/4 = 2.5$

b) con el algoritmo de los libros de texto de estadística se tiene que: $s^2 = \frac{1}{m-1}\left(\sum_{i=1}^{m} x_i^2 - m\,\bar{x}^2\right)$ y

$$fl\left(\sum_{i=1}^{5} x_i^2\right) = fl\left(\sum_{i=1}^{4} x_i^2 + x_5\right) = fl(1090 + 361) = 1450$$

y $\bar{x}^2 = 1450$

por lo que

$$s^2 = \frac{1}{m-1} = (1450 - 1450) = 0$$

c) con el algoritmo de West-Hanson $V_i = V_{i-1} + \frac{i-1}{i}(x_i - M_{i-1})^2$

$M_1 = 15$, $V_1 = 0$

$V_2 = V_1 + \frac{1}{2}(x_2 - M_1) = 0.5(16-15)^2 = 0.5$, $\quad M_2 = \frac{1M_1 + 16}{2} = \frac{31}{2} = 15.5$

$V_3 = V_2 + \frac{2}{3}(x_3 - M_2) = 0.5 + .667 \cdot 1.5 = 1.50$, $\quad M_3 = \frac{2M_2 + 17}{3} = \frac{48}{3} = 16$

$$V_4=V_3+\frac{3}{4}(x_4-M_3)=1.50+.75\cdot(2)^2=4.5\ , \qquad M_4=\frac{3M_3+18}{4}=\frac{66}{4}=16.5$$

$$V_5=V_4+\frac{4}{5}(x_5-M_4)=4.50+.8\cdot(2.5)^2=9.50, \qquad M_5=\frac{4M_4+19}{5}=\frac{85}{5}=17$$

Por lo tanto

$$s^2=\frac{9.50}{4}=2.38$$

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	$(m+4)\varepsilon$
Textos de Estadística	$3mK^2\varepsilon + 2\varepsilon$
Hanson	$(2\sqrt{m}+\frac{8}{3}\sqrt{2m})K\varepsilon + (m+4)\varepsilon$

Donde

$X = (x_1, x_2, \ldots, x_m)^t$

m = número datos

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

k = número de condición de $\frac{1}{\sqrt{m-1}}\frac{\|x\|}{s}$

con

$\| X \|$ = norma de X

$s = [\sum_{i=1}^{m}(x_i - \bar{x})^2] / (m-1)$

C A P I T U L O V

Sistema ANA_RELI.SIS Para el Análisis de Regresión Lineal

Introducción

En este capítulo se presenta el objetivo del presente trabajo: un sistema amigable para computadora que hace transparente al usuario varias tareas engorrosas como es la edición, compilación y ligado de programas. Además le ofrece en cada opción una breve explicación que permite que personas no expertas en la materia puedan interpretar los resultados, las tareas qu se peden realizar mediante este sistema son:

a) Estimar los parámetros β's del modelo lineal

$$y = X\beta + \varepsilon ,$$

mediante el método de mínimos cuadrados.

b) Obtener estadísticas para el análisis del modelo mediante la realización de pruebas de hipótesis y el calcular indicadores que permitan al usuario efectuar diagnósticos sobre el comportamiento de su modelo (correlación entre variables independientes, colinealidad u observaciones aberrantes).

c) Efectuar un análisis tanto gráfico como estadístico de los datos (detección y eliminación de observaciones aberrantes por ejemplo), como de los supuestos sobre el vector de errores ε no observable (independencia en los errores, homocedasticidad, etc.)

5.1 El Sistema: ANA_RELI.SIS

El objetivo de este sistema llamado ANA_RELI.SIS (Análisis de Regresión Lineal) , es brindar al usuario una herramienta amigable y muy fácil de usar.

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

$$y = X\beta + \varepsilon \qquad (5.1.1)$$

bajo los siguientes supuestos:

$$\varepsilon \sim (0, \sigma^2 I) \text{ o } \varepsilon \sim N(0, \sigma^2 I) \qquad (5.1.2)$$

$$X \in \mathbb{R}^{nxp} \quad n > p \text{ de rango } p \qquad (5.1.3)$$

NOTA 4.1.1: El sistema únicamente proporcionará resultados confiables si se cumplen los incisos (5.1.1), (5.1.2), (5.1.3) y la matriz X no es numéricamente de rango deficiente (i.e. no es mal condicionada), de no ser así los resultados obtenidos no deberán considerarse confiables en forma alguna.

5.2 Análisis Estadístico que Efectúa el Sistema

Las funciones básicas de este sistema son proporcionar las siguientes tareas para que el usuario pueda llevar a cabo un análisis estadístico de su modelo bajo estudio, como de los datos del mismo.

Las tareas que puede realizar el sistema son:

- Realización de Pruebas de Hipótesis donde la hipótesis nula es:

a) Ho : todos las β's estimadas son cero

b) Ho : algunas de las β's estimadas son cero (Prueba F Parcial)

- Estimación de la varianza de $\hat{\beta}_i$
- Estimación del error estándar de se(β_i)
- Residuales studentizados
- Cálculo del coeficiente de determinación R^2
- Correlación múltiple
- Análisis de varianza

- Análisis de los residuales o errores estadísticos mediante la graficación de los mismos.
- Cálculo de los factores de inflación de varianza C_{ii}
- Distancia de Cook
- Potencial del i-ésimo caso (h_{ii})
- Cálculo de predicciones

Las estadísticas que se usan para llevar a cabo las tareas antes enlistadas se calculan 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 de Diseño del Sistema se menciona en que opción se realiza cada una de las tareas antes mencionadas.

5.3 Estructura Modular del Sistema

El programa ANA_RELI.SIS, está constituido por los cuatro siguientes módulos:

- MIU Módulo de Interacción con el Usuario
- MCN Módulo de Cálculos Numéricos
- MG Módulo Gráfico
- MOPSA Módulo de Salidas

En el MIU se incluyen:

- Información general sobre el sistema
- Demostración del sistema (DEMO)
- Captura de los datos (X,y) del modelo $y = X\beta + \varepsilon$
- Breves explicaciones en las opciones de los análisis estadísticos y gráficos que realiza el programa.

En el MCN se efectúan todas los cálculos numéricos basados en la descomposición QR de la matriz X del modelo. En el MOPSA se presentan como posibles salidas de los resultados la pantalla, la impresora o un archivo.

El primer, tercer y cuarto módulos, se programaron en lenguaje C versión 6.0 de Microsoft.

En el primer y cuarto módulos (MIU y MOPSA respectivamente), para la generación y manejo de ventanas, se utilizaron las librerias de " The C Programer's Extended Function Library V. 5.2 " desarrollados por Mike Smedley.

El segundo módulo (MCN) se programó en lenguaje Fortran de Microsoft versión 5.0 y para la obtención de la descomposición QR de la matriz X en la cual se basa todo el proceso de cómputo numérico del modelo lineal

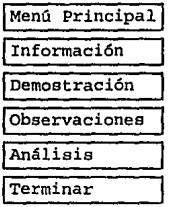
$$y = X\beta + \varepsilon$$

se utilizaron las siguientes subrutinas del paquete de LINPACK [3]: DQRDC, DTRSL, DPODI y DQRSL, del paquete BLAS: DDOT, DNRM, DSWAP, DSCAL, DAXPY, DPODI, DQRSL, DCOPY, DCHEX, DROTG, y DTRSL, y las siguientes rutinas desarrolladas por L. Reichel y W.B. Gragg [7] con el objeto de optimizar los cálculos que se efectúan DINSR, DINSC DDELC y DINSC.

El tercer módulo (MG) se programó en lenguaje C, y su función es graficar las predicciones del modelo $\hat{y}_i$ o una de las variables x_i vs. los residuales estandarizados.

5.4 Diseño del Sistema

A continuación se presenta la conformación del Sistema detallando las funciones que realiza cada menú y sus correspondientes opciónes.



5.4.1 Opción de Información

En la opción de Información, se presentan los objetivos y funciones que realiza el programa y se mencionan los supuestos que se deben cumplir para que el funcionamiento del sistema sea el idóneo.

5.4.2 Opción de Demostración

En la opción de Demostración se muestran al usuario todas las pantallas con que cuenta el programa.

5.4.3 Opción de Observaciones

En la opción de observaciones es posible elegir las siguientes variantes con respecto a los datos del modelo:

Se leen de un archivo las observaciones, que tenga el nombre de obs.dat ó
Si se Generan los valores de las variables al ser estas funciones, por ejemplo $x_i = \text{sen}(t)+t^2$

Además se pregunta al usuario si quiere:

a) Considerar una Constante
b) Centrar las observaciones
c) Estandarizar las observaciones

En ésta opción el usuario debe proporcionar el número de observaciones (renglones) y de variables (columnas) que conforman el modelo lineal $y = X\beta + \varepsilon$.

5.4.4 Opción de Análisis

En la opción de Análisis se presentan los siguientes submenús:

Métodos Estadísticos | Métodos Gráficos | Salidas

5.4.4.a Métodos Estadísticos

Las opciones que se ofrecen en éste submenú son:

En la Opción 1:

- Estimar los parámetros $\hat{\beta}$'s del modelo $y = X\beta + \varepsilon$
- El cálculo de los residuales correspondientes a la estimación de los parámetros $\hat{\beta}$'s del modelo $y = X\beta + \varepsilon$
- La estimación de la varianza s^2
- El cálculo de la desviación estándar de los parámetros β's estimados

En la Opción 2:

- Lo mismo que en la opción anterior
- El cálculo de el coeficiente de determinación R^2
- El cálculo de la matriz de correlación

En la Opción 3:

- Lo mismo que en las opciones anteriores
- El cálculo de las estadísticas t's de cada variable, para conocer la importancia de cada variable dentro del modelo.

En la Opción 4:

- Lo mismo que en las opciones anteriores
- Se efectúa la validación del modelo mediante la prueba de hipótesis

$$Ho: \beta_i = 0 \quad i=1,\ldots,p \qquad Ha: \beta_i \neq 0 \text{ para alguna } i$$

donde p = número de variables que interviene en el modelo

En la Opción 5:

- Lo mismo que en las opciones anteriores

- Se diagnostica la colinealidad con base en los valores que aparecen en la diagonal de la matriz (coeficientes de inflación de la varianza).

En la Opción 6:
- Lo mismo que en las opciones anteriores
- Se evalúa el modelo lineal que se encuentra mediante la regresión en un punto que designa el usuario, obteniéndo así una predicción.

En la Opción 7:
- Lo mismo que en la opción 1
- Se calcula el Coeficiente de Determinación R^2
- Se efectúa una Prueba F Parcial mediante pruebas de hipótesis en las que se supone que algunas variables son cero.

En la Opción 8:
- Lo mismo que en las opciones 1,2, y 3
- Se calculan las estadísticas H_{ii} "potencia de cada observación" y la distancia de Cook, ambas con la finalidad de medir la influencia que cada caso tiene en la regresión.

La Opción 9:
- Permite Agregar o Eliminar Variables

5.4.4.b Métodos Gráficos

Las opciones que ofrece éste submenú son graficar o estimar:

En la Opción 1:
- La variable dependiente y y la independiente del modelo lineal simple.

En la Opción 2:

- Graficar los $\hat{y}$'s estimados por el modelo vs. los residuales estandarizados

En la Opción 3:

- Graficar las variables x_i vs. los residuales estandarizados

La Opción 4:

- Permite Agregar o Eliminar Observaciones

5.4.4.c Salidas

Las opciones que ofrece éste submenú son:

La Opción 1:

- Presenta los resultados de la última opción elegida por el usuario en la Pantalla

La Opción 2:

- Copia los resultados de la última opción elegida por el usuario a un Archivo

La Opción 3:

- Imprime los resultados de la última opción elegida por el usuario

La Opción 4:

- Edita el archivo donde se encuentran las observaciones (Obs.dat)

La Opción 5:

- Sale al Sistema Operativo

La Opción 6:
- Sale del menú que presenta la opción de Análisis

Con la opción de Terminar se finaliza la ejecución del programa.

5.5 Manejo del Sistema

El sistema es muy fácil de operar, en la mayoría de las opciones que se presentan el usuario únicamente debe elegirla e indicar el tipo de salida que desea.

En el menú principal en las opciones de información y demostración lo que el usuario debe hacer para cambiar de pantalla es oprimir cualquier tecla o en la última opción esperar a que el sistema cambie de manera autómatica.

En la opción de Observaciones se pregunta al usuario:
a) Si las observaciones se leerán desde un archivo (*).
a.1) Si las observaciones se centran con respecto a sus medias
a.2) Si se desea considerar una constante en el modelo
a.3) Si las observaciones se pueden representar como funciones básicas como es el caso de x_1 = sen (t), x_2 =cos(t), etc.
Si la respuesta a la pregunta a) es afirmativa, el inciso (b) no se ejecuta.

b) Si las observaciones no se toman de un archivo
b.1) Se preguntará el número de renglones (observaciones) y columnas (variables) que tiene el modelo, y será necesario que el usuario proporcione los coeficientes de la matriz del modelo

$$y = X \beta + c ,$$

que son entradas reales.

b.2) Si a la pregunta de considerar una constante en el modelo, la respuesta fue afirmativa, no es necesario que se proporcione la columna de unos correspondiente a tal constante.

c) Si la(s) variable(s) es (son) funcion(es)
c.1) En este caso aparecerá una ventana con las siguientes funciones: cos, sen, $\sqrt{\ }$, t^2, t y Fin, y al elegir una de ellas distinta de Fin, aparecerá otra ventana con los operadores aritméticos que a continuación se mencionan: (+, -, *, /, (,), Fin), para que eliga el usuario. De esta manera es posible indicar que la variable $X_1 = sen(t) + t^2 + cos(t)$, una vez que se ha proporcionado la función para terminar se elige la opción Fin.

Cuando todas las funciones correspondientes a cada una de las variables se han indicado, es nesario capturar el conjunto de valores en que se evaluarán las funciones posteriormente y así el programa genera la matriz de observaciones del modelo.

5.6 Requerimientos de Instalación

Este programa fue diseñado para ejecutarse bajo ambiente del sistema operativo D.O.S de la versión 5.0 en adelante.

Para poder ejecutar el programa se ocupara el siguiente espacio en disco flexible o duro:

- Para los programas executables 250,000 bytes

- Los archivos de datos ocupan como máximo:
  8 x (n x (p+1) x 13) = # bytes
  donde n = número de observaciones
  p = número de variables incluyendo la constante

Conclusiones :

Se desarrolló un sistema que se basa en el análisis de regresión y el análisis numérico, y que presenta las siguientes características :

1.- Este sistema si bien no se puede equiparar con los que actualmente se han desarrollado para aplicaciones estadísticas como son : el SPSS (Statistical Package for the Social Sciences) y el SAS (Statistical Analysis System, from North Carolina State), presenta la ventaja de ser transparente en su programación (se tiene los programas fuentes), esto permite que crezca o se le adecúe de acuerdo a las necesidades del usuario.

2.- La parte numérica del sistema se basa en la descomposición QR de la matriz X que se realiza con la subrutina DQRDC del paquete LINPACK, y a partir de ella se obtienen todas las estadísticas presentadas en éste trabajo.

3.- Se revisó el funcionamiento del programa y los resultados que se obtuvieron fueron satisfactorios.

Apéndice A

MANUAL DEL SISTEMA ANA_RELI.SIS

Introducción

En diferentes áreas del conocimiento como son la medicina, la biología, la física, la economía, 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 también como variable dependiente) y las causas que lo originan (variables independientes), para lograrlo suponen que el comportamiento de la variable dependiente corresponde a un modelo lineal representado de la siguiente manera :

$$\mathbf{y} = \mathbf{X}\,\beta + \varepsilon \qquad (1)$$

donde $\mathbf{X}$ es una matriz que contiene los datos de las observaciones, β son parámetros desconocidos del modelo, ε son los errores estocásticos, e $\mathbf{y}$ es la respuesta observada.

Uno de los supuestos básicos sobre el vector de errores ε es que tiene media cero y matriz de varianza-covarianza $\sigma^2 I$, i.e

$$\varepsilon \sim (0, \sigma^2 I)$$

Y para efecto de pruebas de hipótesis, se supone adicionalmente normalidad, i.e.

$$\varepsilon \sim N(0, \sigma^2 I)$$

Una forma de obtener los parámetros β's es resolviendo las ecuaciones normales

$$(X'X)\,\hat{\beta} = X'y$$

$$\hat{\beta} = (X'X)^{-1} X'y$$

sin embargo su cálculo directo con frecuencia presenta problemas numéricos; es por ello, que se prefiere la alternativa que a continuación se menciona. Debido a que la matriz X puede ser representada como

$$X = Q\,R$$

donde Q es una matriz ortogonal y R es una matriz triangular superior, a esta forma se le conoce como la descomposición QR de la matriz X, y se utiliza para resolver el sistema

$$X\beta = y \qquad (2)$$

Para obtener la descomposición QR de la matriz X, se emplean reconocidas librerias que actualmente existen en lenguaje fortran y que son las de LINPACK [3] y las desarrolladas por L. Reichel y W.B. Gragg algoritmo en ACM TOMS [7]. Con las subrutinas de

LINPACK se resuelve el sistema lineal (2), y con las de Reichel se actualiza la descomposición QR obtenida al resolver por primera ocasión el sistema lineal (2), ésta actualización se efectúa en los casos de agregar o eliminar observaciones o variables al modelo original.

Requerimientos Computacionales

Este programa fue diseñado para ejecutarse bajo ambiente del sistema operativo D.O.S de la versión 5.0 en adelante.

Para poder ejecutar el programa se ocupara el siguiente espacio en disco flexible o duro:

- Para los programas executables 250,000 bytes
- Los archivos de datos ocupan como máximo:

  8 x (n x (p+1) x 13) bytes

  donde n = número de observaciones y

p = número de variables incluyendo la constante (si la hay)

El sistema ANA_RELI.SIS consta de los siguientes módulos : **MIU**: Este módulo del sistema ANA_RELI.SIS fue programada en lenguaje C y el compilador utilizado fue el C de Microsoft versión 6.00A, para la generación y manejo de ventanas, se recurrió al uso de las librerias de "The C Programer's Extended Function Library V. 5.2" autor Mike Smedley, y las opciones que

lo constituyen son :

Información, mediante ésta opción del programa se presenta una breve explicación de las funciones que éste realiza.

Demostración, ésta opción tiene como finalidad el mostrar al usuario las opciones del programa y un ejemplo de las estadísticas que se manejan evitándo así se tenga que ejecutar cada opción para familiarizarse con el programa.

En la opción de Análisis es donde se determinan las estadísticas y para ello se llama al Módulo de Cálculos Numéricos.

MCN: Este módulo ejecuta las subrutinas en Fortran que se utilizan para resolver el sistema lineal (2) y las estadísticas que se presentan en el programa.

MG: Este módulo grafica los valores que predice el modelo ($\hat{y}$) o los de cada variable vs. los residuales estandarizados.

MOPSA: Este módulo presenta las diferentes opciones de salida de los resultados.

M.1 Estructura del Sistema ANA_RELI.SIS

Programa MCN1.FOR

Este programa conforma el Módulo de Cálculos Númericos que utiliza el programa ANA_RELI.SIS, por lo cual verifica las opciones que eligió el usuario y las ejecuta, utilizando las subrutinas de LINPACK DQRDC y DTRSL para resolver el sistema (2), éstas subrutina manejan de manera auxiliar las subrutinas DPODI, DQRSL, que también pertencen a LINPACK y las subrutinas DDOT, DNORM, DSWAP, DSCAL, DAXPY, DCOPY, DCHEX y DROTG que están en el paquete BLAS.

Para la actualización de la descomposición QR de la matriz X, en los casos de agregar o eliminar variables u observaciones se utilizan las subrutinas desarrolladas por L. Reichel y W.B. Gragg [7].

En la siguiente sección se describirán los algoritmos de cálculo para las opciones que ejecuta el programa ANA_RELI.SIS.

M.2 Algoritmos de Cálculo

Se presenta el algoritmo para la opción en que se obtiene la estadística D_i conocido como distancia de Cook, ya que el sistema fue construido en forma de árbol y ésta opción incluye el mayor número de cálculos por encontrarse en una rama final.

El procedimiento para la opción (no. 21), en la que se determinan las estadísticas D_i o distancias de Cook correspondientes a cada observación, consta de los pasos

siguientes:

1.- Se calcula el promedio y la varianza de las variables (rutina BOYCEN) que intervienen en el modelo lineal

$$y = X\beta + \varepsilon , \qquad \varepsilon \sim N(0,\sigma^2 I) \qquad (1)$$

ambos resultados se obtienen con base en el algoritmo desarrollado por R.J. Hanson [1]. Este paso consta de dos partes:
1.1 Centralización de los datos con respecto a sus medias (rutina CEN)
1.2 Estandarización de los datos (rutina ESTANDAR)

Un análisis numérico de éste algoritmo se encuentra en Tony F. Chan y John Gregg Lewis.(vease [1])

2.- Se calcula la solución del modelo (1)
2.1. Se calcula la descomposición QR de la matriz X del modelo (1) (DQRDC)
2.2 Se Calcula la solución del modelo (1) a partir de la descomposición QR de la matriz X (rutinas DQRDC y DQRSL en LINPACK)
2.2.1 Se calcula la constante del modelo -β_0- si se eligió la centralización de los datos y que en el modelo existiera tal constante.
2.2.1.1 Se calcula la suma de residuales (rutina DSUMRS)
2.2.1.2 Se calcula la varianza estimada s^2 (rutina DS2)

2.2.1.3 Se calcula el Coeficiente de Determinación R^2 (rutina DR2)

2.2.1.4 Se valida el modelo (rutina DVALMD)

2.2.2 Cálculo de los parámetros β's del modelo original si se eligió estandarizar los datos.

2.2.2.1 Se calcula la constante del modelo original β_0 si se eligió estandarizar los datos y que en el modelo existiera tal constante.

2.3 Se calcula la matriz $(X'X)^{-1}$ (rutina DXXINV y DPODI)

3.- Se calcula la suma de residuales (rutina DSUMRS)

4.- Se calculan los errores estandarizados de las estimaciones de los parámetros β's (rutina DERSTD)

5.- Se calcula el Coeficiente de Correlación R^2 (DR2)

6.- Se valida el modelo, presentando una tabla con el correspondiente análisis de varianza (rutina DVALMD)

6.1 Para el caso de efectuar una Prueba F Parcial, se calcula la solución Gauss-Markov para el modelo lineal reducido (MLR)

6.1.1 Se calcula la suma de residuales del MLR (rutina SUMRS)

6.1.2 Se calcula la varianza estimada del MLR (rutina DS2)

6.1.3 Se calculan los errores estandarizados de las estimaciones de los parámetros β's del MLR (rutina DERSTD)

6.1.4 Se calcula el coeficiente de determinación de MLR (rutina DR2)

6.1.5 Se calcula la tabla de análisis de varianza del MLR (rutina DVLAMD)

7.- Se calcula la matriz de correlación (rutian MATCORR)

8.- Se calcula las estadísticas t's para cada variable (rutina TS)

9.- Se calcula las distancias de Cook (D_i)para cada observación (rutina INFLU_CA)

9.1 Se calcula la "potencia" (h_{ii}) de cada observación (rutina INFLU_CA)

9.2 A partir de los "h_{ii}" calculados en 10.1 determinar los D_i (rutina INFLU_CA)

9.3 Se calculan los residuales estandarizados

10.- Se determina la colinealidad entre las variables independientes del modelo.

10.1 Se presenta la diagonal de la matriz $(X'X)^{-1}$ (ya antes calculada por la rutina DXXINV)

11.- Se calculan las predicciones con base a la estimación Gauss-Markov de los parámetros β's para el modelo (1).

M.3 Opciones de Actualización

Con base en los siguientes sistemas lineales

$$X = Q\,R \qquad (1)$$

$$X^* = Q^*\ R^*$$

$$X\beta = y$$

se pueden considerar los casos en los que:

1.- Se elimina una variable

1.1 Para lo cual se calculan a partir de la descomposición QR de la matriz X, Q^* y R^* donde X^* es obtenida de X al eliminar una columna (DDELC)

1.2 Se resuelve el sistema triangular

$$R^* z = Q^* y$$

2.- Se elimina una observación

2.1 Se calcula a partir de la descomposición QR de la matriz X, Q^* y R^* donde X^* es obtenida de X al eliminar un renglón (DDELR)

2.2 Se resuelve el sistema triangular

$$R^* z = Q^* y$$

3.- Se agrega una observación

3.1 Se calcula a partir de la descomposición QR de la matriz X, Q^* y R^* donde X^* es obtenida de X al insertar un renglón (DINSR)

3.2 Se resuelve el sistema triangular

$$R^* z = Q^* y$$

4.- Se agrega una variable

4.1 Se calcula a partir de la descomposición QR de la matriz X, Q^* y R^* donde X^* es obtenida de X al insertar una columna (DINSC)

4.2 Se resuelve el sistema triangular

$$R^* z = Q^* y$$

M.4 Módulo Gráfico

Este módulo fue programado en lenguaje C, y al ejecutarse en el sistema ANA_RELI.SIS ofrece varias opciones de representaciones gráficas para el análisis estadístico de los "datos" del modelo como son las que a continuación se mencionan.

Graficación de las variables que intervienen en el modelo (1) y los residuales estandarizados

La gráfica que se obtiene en este caso, permite verificar si efectivamente la relación que existe entre las variables x_1 y la variable de respuesta **y**, es lineal así como si existen punto aberrantes o "outliers".

El código fuente de éste módulo se encuentra en el archivo MG.C.

Características Computacionales del Módulo de Cálculos Numéricos

Las subrutinas mencionadas fueron compiladas con el compilador Fortran Versión 5.0 de Microsoft, y los cálculos se efectúan en aritmética de doble presición.

El programa MCN1 maneja un arreglo bidimensional para almacenar los datos de la matriz X del modelo (2), que no deben exceder de 50 observaciones y 10 variables. En el caso de que el modelo tenga dimenciones mayores, deberán cambiarse los parámetros que a continuación se mencionan por los necesarios,

MM máximo número de renglones en el modelo

NN máximo número de columnas en el modelo

Parámetros de Entrada del Módulo de Cálculos Numéricos

m, n renglones y columnas de la matriz X

X contiene los coeficientes de la matriz X del modelo lineal

$$X\beta = y \qquad (1)$$

y vector respuesta del modelo (1),

opcion indica la opción que el programa ejecutará opcion=1,...25

opcion2 indica si las variables con conforman el modelo son funciones que deben ser evaluadas ejemplo: x_1=sen(t), x_2=cos(t), etc. En caso de serlo toma el valor de 1 en caso contrario se le asigna el valor de 1.

opcion3 indica si en el modelo se considerará el cálculo de una constante (β_0). Si efectúa dicho cálculo, se le asigna el valor de 1 en caso contrario se le asigna el valor de 1.

opcion4 indica si las observaciones se deben centrar con respecto a las medias de las columnas, en caso de centrar se le signa el valor de 1, en caso contrario se le asigna el valor de 0.

opcion5 indica si las observaciones deben estandarizarse, en caso de estandarizar se le signa el valor de 1, en caso contrario se le asigna el valor de 0.

Archivos de entrada

Tarea.dat En este archivo se encuentra la opción que será ejecutada en el formato I2

Obs.dat En este archivo, se encuentra el número de renglones y de columnas así como los indicadores para calcular o no la instersección en el modelo, centrar los datos o estandarizarlos, todos con formato I2 y en el mismo renglón, por ejemplo:

25 3 1 0 1

m, n, opcion3, opcion4, opcion5

renglones, columnas, intersección, centrar, estandarizar.

En este ejemplo se indica al programa MCN1.EXE que son 25 observaciones con tres variables, si se calcula la intersección del modelo, no se centran los datos, y si se estandarizan (de acuerdo a los valores que han tomado y que se explican en los

Parámetros de Entrada.

A partir del segundo renglón comienzan las observaciones y a variable de respuesta, ambos en formato F12.5, ejemplo:

-----2.00000---123.00000------.95000

Nota: - indica un espacio en blanco

Predic.dat en este archivo se encuentran:

n número de variables en el modelo en el formato I2

y vector unidimensional en el que se almacenan los valores valores en que se debe evaluar el modelo lineal $X\hat{\beta}$ para obtener la predicción y*

Mr.dat en este archivo se encuentran:

n número de variables que se omitirán el modelo reducido que interviene en la Prueba F Parcial, con el formato I2

jpvt arreglo unidimensional con los subíndices de variables que se omitirán el modelo reducido que interviene en la Prueba F Parcial, con el formato I2

Act.dat En este archivo se almacenan:

m número de observaciones del modelo (1)

n número de variables del modelo (1)

qraux arreglo bidimensional donde se almacena la matriz X que sale de la subrutina DQRDC

jpvt arreglo unidimensional donde se almacenan los pivotes de las descomposición QR de la matriz X del modelo (1)

bini arreglo unidimensional donde se almacenen los

valores de la respuesta y del modelo lineal (1)

Nota: este archivo es utilizado en el caso de que se quiera agregar o eliminar una observación o una variable al modelo original.

Act.val En este archivo se encuentran:

k indica que la columna o renglón k-ésimo será agregada o eliminada del modelo (1)

kval vector unidimensional que contiene los valores de la columna o renglón que se agregará al modelo.

Archivos de Salida

Mel.dat En éste archivo se presentan las estadísticas que se calculan a partir de la opción elegida.

Nota: Los resultado numéricos se presentan en el formato F12.3 debido a que se observó en los libros que en general se manejan tanto los datos como los resultados con únicamente tres decimales.

Mg.dat En éste archivo se escriben

n número de observaciones del modelo en el formato I2

res residuales estandarizados en el formato F12.5, para ser utilzadas por el programa que ejecuta los Métodos Gráficos contenidos en la parte de Análisis.

Subprogramas Estandares de LINPACK

DQRDC, DTRSL, DDOT, DNRM, DSWAP, DSCAL, DAXPY, DPODI, DQRSL, DCOPY, DCHEX, DROTG,

Subrutinas Estandares Desarrolladas por L. Reichel y W.B. Gragg

DDELC, DDELR, DINSR Y DINSC.

A partir de los siguientes sistemas lineales

$$X = Q\,R \qquad\qquad X^* = Q^*\,R^*$$

DDELC Calcula $Q^*\,R^*$, a partir de Q, R donde X^* es obtenida de X al borrar una columna.

DDELR Calcula $Q^*\,R^*$, a partir de Q, R donde X^* es obtenida de X al borrar un renglón.

DINSC Calcula $Q^*\,R^*$, a partir de Q, R donde X^* es obtenida de X al agregar una columna.

DINSR Calcula $Q^*\,R^*$, a partir de Q, R donde X^* es obtenida de X al agregar un renglón

A su vez estas subrutinas ocupan las siguientes del paquete BLAS: DADFGR, DAPLRC, DAPLRR, DAPLRV, DSHFTD, DSHFTU, DMINRW, DSCAL, DZERO, DAPX, DATPX, DAXPY, DCOPY, DDOT, DNRM2, DSWAP, DPODI, DORTHX Y DORTHO

Subrutinas utilizadas por el programa MCN1.FOR:
HANSN2 (Desarrollada por R.J. Hanson),LEEOPC, LEEOBS2, DQRST, ORDINV, MATCORR, COEFCO, TS, SUMRES, ERRORBS, COLINEAL, VALMOD, PREDICCION, GR_XS_RES, GR_NOR, ESCBYR, INFLU_CA, LEEMR, MULMAT, ORDEN, BOYCEN, CENTRAR, ESTANDAR, CALCBO, CALEBO, GENEOBS, ANIDA1, ANIDA2, EVALUA, EVAL2, QR, ACT, ORDBS, ACT2, ACT3, ESCAQR, LEEAQR, LEEKVAL, LEEK, ACT4.

Los ejemplos con los que se realizaron las pruebas para verificar el funcionamiento del programa , se tomaron de publicaciones reconocidas como son:

- Sanford Weisberg, Applied Linear Regression John Wiley & Sons (1980)
- Douglas C. Montgomery y Elizabeth Peck, Introduction to Linear Regression Analysis John Wyley & Sons (1982)
- Chattiergee, Regression Analysis by Example, John Wyley & Sons (1977)

Resultados Numéricos

Con base en que para obtener los resultados numéricos se utilizaron las Librerias de LINPACK [3], que gozan de reconocido prestigio en el ámbito del Análisis Número, así como las rutinas de actualización de la descomposición QR de la matriz X de observaciones del modelo desarrolladas por L. Reichel y W.B. Gragg y que los resultados de los ejemplos en la bibliografía mencionada coincíden con los obtenidos mediante el programa, es

posible afirmar que los resultados numéricos obtenidos al ejecutar el programa son confiables.

Ejemplo 1 [Searle]

Supongase que el ingreso anual de un hombre (y) está asociado a su grado de escolaridad (x_1) y a su edad (x_2) mediante un modelo de regresión lineal, entonces su ingreso esperado es:.

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (1)$$

Hombre i	Ingreso y_i	Años de escolaridad x_{i1}	Edad x_{i2}
1	10	6	28
2	20	12	40
3	17	10	32
4	12	8	36
5	11	9	34

A partir de estos datos el método de mínimos cuadrados proporciona la estimación

$$\hat{E}(y) = 56/24 + (50/24)x_1 - (5/24)x_2$$

ANALISIS DE VARIANZA POR AJUSTE DE REGRESION (Ho: $\beta = 0$)

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión	2	SCR = 62.5	31.25	F = 5.43
Residual	2	SCE = 11.5	5.75	
Total	5	SCT = 74.0		

Para este ejemplo se obtienen los siguientes resultados con el programa ANA_RELI.SIS.

MATRIZ X Y VECTOR y DEL MODELO $y = X\beta + e$

1	1.000	6.000	28.000	10.000
2	1.000	12.000	40.000	20.000
3	1.000	10.000	32.000	17.000
4	1.000	8.000	36.000	12.000
5	1.000	9.000	34.000	11.000

RESUMEN ESTADISTICO BASICO

N	PROMEDIO	VARIANZA	DESVIACION ESTANDAR	MINIMO	MAXIMO
5	9.000	5.000	2.236	6.000	12.000
5	34.000	20.000	4.472	28.000	40.000
5	14.000	18.500	4.301	10.000	20.000

PARAMETROS DEL VECTOR β EN EL MODELO $y = X\beta + e$

β 0 = 2.33333
β 1 = 2.08333
β 2 = -.20833

RESIDUALES

1	1.00000
2	1.00000
3	.50000
4	.50000
5	-3.00000

inv(X'X)

17.58889	.63889	-.68056
.63889	.13889	-.05556
-.68056	-.05556	.03472

SUMA DE RESIDUALES = 11.50000

S_2 = 5.75000 S = 2.39792

SE(β 0) = 10.05665
SE(β 1) = .89365
SE(β 2) = .44683

R_2 = .84459

MATRIZ DE CORRELACION

1.000	.409	-.871
.409	1.000	-.800
-.871	-.800	1.000

ESTADISTICO t PARA βi ESTIMADA DONDE i=

1 = .359286E+00
2 = .361001E+01
3 =-.722001E+00

VALOR ABS DE t
.359286E+00
.722001E+00
.361001E+01

ANALISIS DE VARIANZA Ho: β=0

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMAS DE CUADRADOS	CUADRADOS MEDIOS	F
REGRESION	2	62.50000	31.25000	5.43478
RESIDUAL	2	11.50000	5.75000	
TOTAL	4	74.00000		

Ejemplo 2 [Montgomery]

Se analizan las rutas del sistema de distribución de una bebida. Interesa predecir la cantidad de tiempo que requiere el conductor de la ruta para efectuar la entrega. Un ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan al tiempo de entrega son el número de casos que el producto se almacena y la distancia que camina el conductor de la ruta.

El modelo de regresión que se propone es :

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

Número de Observación	Tiempo de Entrega (Min) y	Numero de Casos x_1	Distancia (Pies) x_2
1	7	560	16.68
2	3	220	11.50
3	3	340	12.03
4	4	80	14.88
5	6	150	13.75
6	7	330	18.11
7	2	110	8.00
8	7	210	17.83
9	30	1460	79.24
10	5	605	21.50
11	16	688	40.33
12	10	215	21.00
13	4	255	13.50
14	6	462	19.75
15	9	448	24.00
16	10	776	29.00
17	6	200	15.35
18	7	132	19.00
19	3	36	9.50
20	17	770	35.10
21	10	140	17.90
22	26	810	52.32
23	9	450	18.75
24	8	635	19.83
25	4	150	10.75

Matriz $(X'X)^{-1}$

.113215	-.004448	-.000083
-.004448	.002743	-.000047
-.000083	-.0004	.000001

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.34123 \\ 1.61590 \\ 0.01438 \end{bmatrix}$$

$s^2 = 10.6239$ Estadística F : 261.24 $R^2 = 0.9596$

Si la observación 9 se elimina, se obtiene el siguiente resultado:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 4.447 \\ 1.498 \\ 0.010 \end{bmatrix}$$

Si en el modelo de regresión no se considera la observación 22, se obtiene el siguiente resultado:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 1.916 \\ 1.786 \\ 0.012 \end{bmatrix}$$

Para este ejemplo se obtienen los siguientes resultados con el programa ANA_RELI.SIS.

MATRIZ X Y VECTOR y DEL MODELO $y = X\beta + e$

1	1.000	7.000	560.000	16.680
2	1.000	3.000	220.000	11.500
3	1.000	3.000	340.000	12.030
4	1.000	4.000	80.000	14.880
5	1.000	6.000	150.000	13.750
6	1.000	7.000	330.000	18.110
7	1.000	2.000	110.000	8.000
8	1.000	7.000	210.000	17.830
9	1.000	30.000	1460.000	79.240
10	1.000	5.000	605.000	21.500
11	1.000	16.000	688.000	40.330
12	1.000	10.000	215.000	21.000
13	1.000	4.000	255.000	13.500
14	1.000	6.000	462.000	19.750
15	1.000	9.000	448.000	24.000
16	1.000	10.000	776.000	29.000
17	1.000	6.000	200.000	15.350
18	1.000	7.000	132.000	19.000
19	1.000	3.000	36.000	9.500
20	1.000	17.000	770.000	35.100
21	1.000	10.000	140.000	17.900
22	1.000	26.000	810.000	52.320
23	1.000	9.000	450.000	18.750
24	1.000	8.000	635.000	19.830
25	1.000	4.000	150.000	10.750

RESUMEN ESTADISTICO BASICO

N	PROMEDIO	VARIANZA	DESVIACION ESTANDAR	MINIMO	MAXIMO
25	8.760	47.357	6.882	2.000	30.000
25	409.280	105747.293	325.188	36.000	1460.000
25	22.384	241.023	15.525	8.000	79.240

PARAMETROS DEL VECTOR β EN EL MODELO $y = X\beta + e$

β 0 = 2.34123
β 1 = 1.61591
β 2 = .01438

RESIDUALES

CASO	1	2	3	4	5
	-5.02808	1.14639	-.04979	4.92435	-.44440
CASO	6	7	8	9	10
	-.28957	.84462	1.15660	7.41971	2.37641
CASO	11	12	13	14	15
	2.23749	-.59304	1.02701	1.06754	.67120
CASO	16	17	18	19	20
	-.66293	.43636	3.44862	1.79319	-5.78797
CASO	21	22	23	24	25

-2.61418 -3.68653 -4.60757 -4.57285 -.21258

25 OBSERVACIONES

inv(X'X)

.11322	-.00445	-.00008
-.00445	.00274	-.00005
-.00008	-.00005	.00000

SUMA DE RESIDUALES = 233.73168

S_2 = 10.62417 S = 3.25947

SE(β 0) = 1.09673
SE(β 1) = .17073
SE(β 2) = .00361

R_2 = .95959

MATRIZ DE CORRELACION

1.000	-.252	-.224
-.252	1.000	-.824
-.224	-.824	1.000

ESTADISTICO t PARA βi ESTIMADA DONDE i=

1 = .385406E+01
2 = .170871E+02
3 = .718787E+01

VALOR ABS DE t
.385406E+01
.718787E+01
.170871E+02

ESTADISTICOS Di (DISTANCIAS DE COOK PARA MEDIR LA INFLUENCIA DE LOS CASOS OBSERVADOS)

OBS.i	RESIDUALES	hii	ri	Di
1	-5.02808	.10180	-1.62768	.10009
2	1.14639	.07070	.36484	.00338
3	-.04979	.09873	-.01609	.00001
4	4.92435	.08537	1.57972	.07765
5	-.44440	.07501	-.14176	.00054
6	-.28957	.04287	-.09081	.00012
7	.84462	.08180	.27042	.00217
8	1.15660	.06373	.36672	.00305

9	7.41971	.49829	3.21376	3.41932
10	2.37641	.19630	.81325	.05385
11	2.23749	.08613	.71808	.01620
12	-.59304	.11366	-.19326	.00160
13	1.02701	.06112	.32518	.00229
14	1.06754	.07824	.34114	.00329
15	.67120	.04111	.21029	.00063
16	-.66293	.16594	-.22270	.00329
17	.43636	.05943	.13804	.00040
18	3.44862	.09626	1.11295	.04398
19	1.79319	.09645	.57877	.01192
20	-5.78797	.10168	-1.87355	.13244
21	-2.61418	.16528	-.87784	.05086
22	-3.68653	.39158	-1.45000	.45105
23	-4.60757	.04126	-1.44369	.02990
24	-4.57285	.12061	-1.49606	.10232
25	-.21258	.06664	-.06751	.00011

Si la observación 22 se elimina, se obtiene el siguiente resultado con el programa ANA_RELI.SIS:

COEFICIENTES DE REGRESION OBTENIDOS AL ELIMINAR LA

OBSERVACION 9 DEL MODELO INICIAL

4.447199158225843 1.497702961451742 .032392223157713E-002

Si en el modelo de regresión inicial no se considera la observación 22, los resultados que se obtienen con el programa ANA_RELI.SIS son:

MATRIZ X Y VECTOR y DEL MODELO $y = X\beta + e$

1	1.000	7.000	560.000	16.680
2	1.000	3.000	220.000	11.500
3	1.000	3.000	340.000	12.030
4	1.000	4.000	80.000	14.880
5	1.000	6.000	150.000	13.750
6	1.000	7.000	330.000	18.110
7	1.000	2.000	110.000	8.000
8	1.000	7.000	210.000	17.830
9	1.000	30.000	1460.000	79.240
10	1.000	5.000	605.000	21.500
11	1.000	16.000	688.000	40.330
12	1.000	10.000	215.000	21.000
13	1.000	4.000	255.000	13.500
14	1.000	6.000	462.000	19.750

15	1.000	9.000	448.000	24.000
16	1.000	10.000	776.000	29.000
17	1.000	6.000	200.000	15.350
18	1.000	7.000	132.000	19.000
19	1.000	3.000	36.000	9.500
20	1.000	17.000	770.000	35.100
21	1.000	10.000	140.000	17.900
22	1.000	9.000	450.000	18.750
23	1.000	8.000	635.000	19.830
24	1.000	4.000	150.000	10.750

RESUMEN ESTADISTICO BASICO

N	PROMEDIO	VARIANZA	DESVIACION ESTANDAR	MINIMO	MAXIMO
24	8.042	35.955	5.996	2.000	30.000
24	392.583	103072.514	321.049	36.000	1460.000
24	21.137	210.915	14.523	8.000	79.240

PARAMETROS DEL VECTOR β EN EL MODELO $y = X\beta + e$

β 0 = 1.91574
β 1 = 1.78632
β 2 = .01237

RESIDUALES

CASO	1	2	3	4	5
	-4.66671	1.50408	.54979	4.82944	-.73905
CASO	6	7	8	9	10
	-.39181	1.15101	.81248	5.67565	3.16933
CASO	11	12	13	14	15
	1.32313	-1.43834	1.28484	1.40179	.46598
CASO	16	17	18	19	20
	-.37741	.24250	2.94727	1.78000	-6.70746
CASO	21	22	23	24	25
	-3.61065	-4.80875	-4.23072	-.16640	.00000

24 OBSERVACIONES

inv(X'X)

.12132	-.00769	-.00005
-.00769	.00404	-.00006
-.00005	-.00006	.00000

SUMA DE RESIDUALES = 211.39451

S_2 = 10.06641 S = 3.17276

SE(β 0) = 1.10511
SE(β 1) = .20176

SE(β 2) = .00377

Al agregar la observación 22 al modelo se obtiene el siguiente resultado utilizando el programa ANA_RELI.SIS.

810.000000000000000 26.000000000000000 1.000000000000000

COEFICIENTES DE REGRESION OBTENIDOS AL AGREGAR UNA

OBSERVACION EN EL LUGAR 22 AL MODELO

2.341183924266383 1.615895068145520 1.438531761388166E-002

Ejemplo 3 (Weisberg)

x_1 = TCC =tasa de consumo de combustible 1972, en centavos por galón.
x_2 = PPLC = porcentaje de población con licencia de conductores
x_3 = IP = ingreso promedio (miles de dólares)
x_4 = RM = (miles de millas)
y = CCM = consumo de combustible del motor (galones por persona)

ESTADO i	x_1 TCC	x_3 IP	x_4 RM	x_2 PPLC	y CCM
1 ME	9.00	3.571	1.976	52.5	541
2 NH	9.00	4.092	1.250	57.2	524
3 VT	9.00	3.865	1.586	58.0	561
4 MA	7.50	4.870	2.351	52.9	414
5 RI	8.00	4.399	.431	54.4	410
6 CN	10.00	5.342	1.333	57.1	457
7 NY	8.00	5.319	11.868	45.1	344
8 NJ	8.00	5.126	2.138	55.3	467
9 PA	8.00	4.447	8.557	52.9	464
10 OH	7.00	4.512	8.507	55.2	498
11 IN	8.00	4.319	5.939	53.0	580
12 IL	7.50	5.126	14.186	52.5	471
13 MI	7.00	4.817	6.930	57.4	525
14 WI	7.00	4.207	6.580	54.5	508
15 MN	7.00	4.332	8.159	60.8	566
16 IA	7.00	4.318	10.340	58.6	635
17 MO	7.00	4.206	8.508	57.2	603
18 ND	7.00	3.178	4.725	54.0	714
19 SD	7.00	4.716	5.915	72.4	865
20 NE	8.50	4.341	6.010	67.7	640
21 KS	7.00	4.593	7.834	66.3	649
22 DE	8.00	4.983	.602	60.2	540
23 MD	9.00	4.897	2.449	51.1	464
24 VA	9.00	4.258	4.686	51.7	547
25 WV	8.50	4.574	2.619	55.1	460
26 NC	9.00	3.721	4.746	54.4	566
27 SC	8.00	3.448	5.399	54.8	577
28 GA	7.50	3.846	9.061	57.9	631
29 FL	8.00	4.188	5.975	56.3	574
30 KY	9.00	3.601	4.650	49.3	534
31 TN	7.00	3.640	6.905	51.8	571
32 AL	7.00	3.333	6.594	51.3	554
33 MS	8.00	3.063	6.524	57.8	577
34 AR	7.50	3.357	4.121	54.7	628
35 LA	8.00	3.528	3.495	48.7	487
36 OK	6.58	3.802	7.834	62.9	644
37 TX	5.00	4.045	17.782	56.6	640
38 MT	7.00	3.897	6.385	58.6	704
39 ID	8.50	3.635	3.274	66.3	648
40 WY	7.00	4.345	3.905	67.2	968
41 CO	7.00	4.449	4.639	62.6	587
42 NM	7.00	3.656	3.985	56.3	699
43 AZ	7.00	4.300	3.635	60.3	632
44 UT	7.00	3.745	2.611	50.8	591
45 NV	6.00	5.215	2.302	67.2	782
46 WN	9.00	4.476	3.942	57.1	510
47 OR	7.00	4.296	4.083	62.3	610
48 CA	7.00	5.002	9.794	57.3	524

$$\hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} -34.790149 \\ 13.364494 \\ -66.588752 \\ -2.425889 \end{bmatrix}$$

$$\beta_0 = \hat{y} - \Sigma\, \hat{\beta}_j^*\, \bar{x}_j = 377.2911$$

$$s^2 = \frac{SCE}{n-p'} = \frac{189049.97}{48-(4+1)} = 4396.5$$

Matriz de Correlación del ejemplo 3

TCC	1.0000				
PPLC	-.2880	1.000			
IP	.0127	.1571	1.000		
RM	-.5221	-.0641	.0502	1.000	
CCM	-.4513	.6990	-.2449	.0190	1.000
	TCC	PPLC	IP	RM	CCM

Matriz $(X'X)^{-1}$ para el ejemplo 3

Constante	7.83019	-.42651	-.06110	-.14950	-.07534
TCC	-.42651	.038263	.00221	-.00591	.00571
PPLC	-.06110	.00221	.00084	-.00145	.00041
IP	-.14950	-.00591	-.00145	.06746	-.00154
RM	-.07534	.00571	.00041	-.00015	.00261
	Intercept	TCC	PPLC	IP	RM

Para el ejemplo 3 el error estándar de $\hat{\beta}_1$ y la covarianza de $\hat{\beta}_1$ $\hat{\beta}_2$ son :

$$se(\hat{\beta}_1) = s\sqrt{a_{11}} = s\sqrt{7.89019} = 185.54$$

$$cov(\hat{\beta}_1, \hat{\beta}_2) = s^2 a_{12} = (0.0022158) = 0.1469$$

Resumen Estadístico Básico

Variable	Estimación de β	Error Estandar	Valor - t
Constante	377.2911	185.5412	2.03
TCC	-34.79015	12.97020	-2.68
PPLC	13.36449	1.922981	6.95
IP	-66.58875	17.22175	-3.87
RM	-2.425889	3.389174	-.72

s^2 = 4396.511, grados de libertad = 43, R^2 = 0.6787

Tabla de Análisis de Varianza para la prueba la hipótesis

Ho: β_i=0 i =0,4 vs. Ha: $\beta_j \neq 0$ para alguna j

Fuente de Variación	Grados de libertad	Sumas de Cuadrados	Cuadrados Medios	F
Regresión	4	SCR = 399.316	99.829	F = 22.70
Residual	43	SCE = 189.050	4397.000	
Total	47	SCT = 588.366		

Variable	n	Promedio	Varianza	Desviación Estándar	Valores Mínimos	Máximos
TCC	48	7.6683	00.90396	0.95077	5.000	10.000
PPLC	48	57.033	30.77000	5.54700	45.100	72.400
IP	48	4.2418	00.32904	0.57362	3.063	5.342
RM	48	5.5654	12.19100	3.49150	.431	17.782
CCM	48	576.77	12518.	111.89000	344.000	968.000

Tabla de análisis de varianza para la Prueba Parcial F

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrados Medios	F
Regresión sobre (RM ,IP Y PPLC)	3	SCR = 367.684	122.561	
TCC después de otros	1	SCE = 31.632	31.632	7.19
Residual	43	SCT = 189.050	4.397	

Para este ejemplo se obtienen los siguientes resultados con el programa ANA_RELI.SIS archivo 1.chi

MATRIZ X Y VECTOR y DEL MODELO $y = X\beta + e$

1	1.000	9.000	52.500	3.571	1.976	541.000
2	1.000	9.000	57.200	4.092	1.250	524.000
3	1.000	9.000	58.000	3.865	1.586	561.000
4	1.000	7.500	52.900	4.870	2.351	414.000
5	1.000	8.000	54.400	4.399	.431	410.000
6	1.000	10.000	57.100	5.342	1.333	457.000
7	1.000	8.000	45.100	5.319	11.868	344.000
8	1.000	8.000	55.300	5.126	2.138	467.000
9	1.000	8.000	52.900	4.447	8.577	464.000
10	1.000	7.000	55.200	4.512	8.507	498.000
11	1.000	8.000	53.000	4.391	5.939	580.000
12	1.000	7.500	52.500	5.126	14.186	471.000
13	1.000	7.000	57.400	4.817	6.930	525.000
14	1.000	7.000	54.500	4.207	6.580	508.000
15	1.000	7.000	60.800	4.332	8.159	566.000
16	1.000	7.000	58.600	4.318	10.340	635.000
17	1.000	7.000	57.200	4.206	8.508	603.000
18	1.000	7.000	54.000	3.718	4.725	714.000
19	1.000	7.000	72.400	4.716	5.915	865.000
20	1.000	8.500	67.700	4.341	6.010	640.000
21	1.000	7.000	66.300	4.593	7.834	649.000
22	1.000	8.000	60.200	4.983	.602	540.000
23	1.000	9.000	51.100	4.897	2.449	464.000
24	1.000	9.000	51.700	4.258	4.686	547.000
25	1.000	8.500	55.100	4.574	2.619	460.000
26	1.000	9.000	54.400	3.721	4.746	566.000
27	1.000	8.000	54.800	3.448	5.399	577.000
28	1.000	7.500	57.900	3.846	9.061	631.000
29	1.000	8.000	56.300	4.188	5.975	574.000
30	1.000	9.000	49.300	3.601	4.650	534.000
31	1.000	7.000	51.800	3.640	6.905	571.000
32	1.000	7.000	51.300	3.333	6.594	554.000
33	1.000	8.000	57.800	3.063	6.524	577.000
34	1.000	7.500	54.700	3.357	4.121	628.000
35	1.000	8.000	48.700	3.528	3.495	487.000
36	1.000	6.580	62.900	3.802	7.834	644.000
37	1.000	5.000	56.600	4.045	17.782	640.000
38	1.000	7.000	58.600	3.897	6.385	704.000
39	1.000	8.500	66.300	3.635	3.274	648.000
40	1.000	7.000	67.200	4.345	3.905	968.000
41	1.000	7.000	62.600	4.449	4.639	587.000
42	1.000	7.000	56.300	3.656	3.985	699.000
43	1.000	7.500	60.300	4.300	3.635	632.000
44	1.000	7.000	50.800	3.745	2.611	591.000
45	1.000	6.000	67.200	5.215	2.302	782.000
46	1.000	9.000	57.100	4.476	13.942	510.000
47	1.000	7.000	62.300	4.296	4.083	610.000
48	1.000	7.000	59.300	5.002	9.794	524.000

RESUMEN ESTADISTICO BASICO

N	PROMEDIO	VARIANZA	DESVIACION ESTANDAR	MINIMO	MAXIMO
48	7.679	.895	.946	5.000	10.000
48	57.033	30.770	5.547	45.100	72.400
48	4.242	.329	.574	3.063	5.342
48	5.774	13.583	3.686	.431	17.782
48	576.771	12518.436	111.886	344.000	968.000

PARAMETROS DEL VECTOR β EN EL MODELO $y = X\beta + e$

β 0 = 362.38002
β 1 = -33.72069
β 2 = 13.46287
β 3 = -66.43203
β 4 = -2.20227

RESIDUALES

CASO	1	2	3	4	5
	16.88584	-30.37743	-18.48783	-78.95929	-121.81110
CASO	6	7	8	9	10
	20.91239	23.69845	-24.87232	-26.48837	-53.00975
CASO	11	12	13	14	15
	78.63557	26.49632	-38.83928	-58.09128	-73.12599
CASO	16	17	18	19	20
	29.36543	4.73851	118.06969	90.27268	-45.57357
CASO	21	22	23	24	25
	-47.54878	-30.72287	47.86441	85.26309	-47.93059
CASO	26	27	28	29	30
	32.37147	-12.43224	17.47716	14.80166	60.84886
CASO	31	32	33	34	35
	4.30726	-27.04085	-75.91964	14.19389	-19.18727
CASO	36	37	38	39	40
	-73.48544	-7.89687	61.68757	-71.65197	234.20678
CASO	41	42	43	44	45
	-76.33861	66.35661	4.37690	35.28895	68.75171
CASO	46	47	48	49	50
	10.42996	-60.68831	-46.82151	.00000	.00000

48 OBSERVACIONES

inv(X'X)

6.94819	-.05131	-.35895	-.15721	-.05625
-.05131	.00202	.14695	.09894	.00029
-.35895	.14695	.03315	.00382	.00184
-.15721	.09894	.00382	-.00145	-.00141
-.05625	.00029	.00184	-.00141	.00081

SUMA DE RESIDUALES = 188955.40786

S^2 = 4394.31181 S = 66.28961

SE(β 0) = 174.73559
SE(β 1) = 2.98123
SE(β 2) = 12.06952
SE(β 3) = 2.52362
SE(β 4) = 1.89194

R^2 = .67885

ANALISIS DE VARIANZA Ho: β=0

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMAS DE CUADRADOS	CUADRADOS MEDIOS	F
REGRESION	4	399411.07131	99852.76783	22.72319
RESIDUAL	43	188955.40786	4394.31181	
TOTAL	47	588366.47917		

RESULTADOS DEL MODELO y = Xβ + e AL NO CONSIDERAR LA(S) VARIABLE(S) Xi i = 2

1	1.000	52.500	3.571	1.976	541.000
2	1.000	57.200	4.092	1.250	524.000
3	1.000	58.000	3.865	1.586	561.000
4	1.000	52.900	4.870	2.351	414.000
5	1.000	54.400	4.399	.431	410.000
6	1.000	57.100	5.342	1.333	457.000
7	1.000	45.100	5.319	11.868	344.000
8	1.000	55.300	5.126	2.138	467.000
9	1.000	52.900	4.447	8.577	464.000
10	1.000	55.200	4.512	8.507	498.000
11	1.000	53.000	4.391	5.939	580.000
12	1.000	52.500	5.126	14.186	471.000
13	1.000	57.400	4.817	6.930	525.000
14	1.000	54.500	4.207	6.580	508.000
15	1.000	60.800	4.332	8.159	566.000
16	1.000	58.600	4.318	10.340	635.000
17	1.000	57.200	4.206	8.508	603.000
18	1.000	54.000	3.718	4.725	714.000
19	1.000	72.400	4.716	5.915	865.000
20	1.000	67.700	4.341	6.010	640.000
21	1.000	66.300	4.593	7.834	649.000
22	1.000	60.200	4.983	.602	540.000
23	1.000	51.100	4.897	2.449	464.000
24	1.000	51.700	4.258	4.686	547.000
25	1.000	55.100	4.574	2.619	460.000
26	1.000	54.400	3.721	4.746	566.000
27	1.000	54.800	3.448	5.399	577.000

28	1.000	57.900	3.846	9.061	631.000
29	1.000	56.300	4.188	5.975	574.000
30	1.000	49.300	3.601	4.650	534.000
31	1.000	51.800	3.640	6.905	571.000
32	1.000	51.300	3.333	6.594	554.000
33	1.000	57.800	3.063	6.524	577.000
34	1.000	54.700	3.357	4.121	628.000
35	1.000	48.700	3.528	3.495	487.000
36	1.000	62.900	3.802	7.834	644.000
37	1.000	56.600	4.045	17.782	640.000
38	1.000	58.600	3.897	6.385	704.000
39	1.000	66.300	3.635	3.274	648.000
40	1.000	67.200	4.345	3.905	968.000
41	1.000	62.600	4.449	4.639	587.000
42	1.000	56.300	3.656	3.985	699.000
43	1.000	60.300	4.300	3.635	632.000
44	1.000	50.800	3.745	2.611	591.000
45	1.000	67.200	5.215	2.302	782.000
46	1.000	57.100	4.476	13.942	510.000
47	1.000	62.300	4.296	4.083	610.000
48	1.000	59.300	5.002	9.794	524.000

PARAMETROS DEL VECTOR β EN EL MODELO $y = X\beta + e$

β 0 = -2.74617
β 1 = 15.33205
β 2 = -71.81666
β 3 = 1.68229

RESIDUALES

CASO	1	2	3	4	5
	-8.05361	-58.47644	-50.60971	-48.52745	-106.12118
CASO	6	7	8	9	10
	-34.31204	17.29789	-13.58099	-39.37984	-35.85772
CASO	11	12	13	14	15
	75.50310	13.08052	-18.03119	-33.78759	-66.05879
CASO	16	17	18	19	20
	31.99722	16.50059	147.88074	86.44204	-93.58837
CASO	21	22	23	24	25
	-48.09419	-23.39384	30.84443	54.99107	-57.96655
CASO	26	27	28	29	30
	-6.07196	-21.90926	6.98385	4.26798	31.66502
CASO	31	32	33	34	35
	29.34217	-1.51633	-97.44742	26.23860	-9.43532
CASO	36	37	38	39	40
	-57.77219	35.53577	77.41587	-110.22330	245.90615
CASO	41	42	43	44	45
	-58.33227	94.40928	12.91979	79.43873	125.08335
CASO	46	47	48	49	50
	-64.71728	-40.78525	-39.69409	.00000	.00000

48 OBSERVACIONES

SUMA DE RESIDUALES = 223256.17269

S_2 = 5315.62316 S = 72.90832

SE(β 0) = 963.75751
SE(β 1) = 125.88524
SE(β 2) = 253.29238
SE(β 3) = 115.82156

R_2 = .62055

ANALISIS DE VARIANZA Ho: β=0

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMAS DE CUADRADOS	CUADRADOS MEDIOS	F
REGRESION	3	365110.30648	121703.43549	23.98568
RESIDUAL	44	223256.17269	5074.00392	
TOTAL	47	588366.47917		

ANALISIS DE VARIANZA LA PARA PRUEBA PARCIAL F

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMAS DE CUADRADOS	CUADRADOS MEDIOS	F
REG MOD. RED.	3	365110.30648	121703.43549	
REG DESP.OTRA	1	34300.76483	34300.76483	7.80572
RESIDUAL	43	188955.40786	4394.31181	

Actualización de Opciones

Una de las características más importantes del sistema ANA_RELI.SIS es el ser un sistema al que se le pueden agregar o modificar opciones, ya que es transparente en cuanto a su programación porque se tienen los programas fuentes.

Para efectuar actualizaciones es necesario editar el programa ANA_RELI.C Si se desea modificar un menú es necesario localizar el procedimiento donde se encuentra el menú que se quiere modificar y

a) Si la modificación consiste en agregar opciones al menú se debe utilizar la función wmenuitem y se incluye antes de la función menuend. Ejemplo :

```
wmenuitem(7,0,"Opción Nueva",'O',1,M_CLOSE,proce_nuevo,0xd00,0);
```

en este ejemplo la opción se colocará en el séptimo renglón y la primera columna de la ventana en que se encuentra éste menú, con el nombre de Opción Nueva, la letra con que se activará estando dentro del menú es la O y el procedimiento que s ejecutará en ésta opción es proce_nuevo.

b) Si la modificación consiste en cambiar los colores de la ventana en que se activa el menú es necesario modificar los parámetros seis y siete de la función wopen. Ejemplo:

```
wopen(10,10,20,40,0,LCYAN|_BLUE,LGREEN|_BLUE);
```

en éste ejemplo se indica que el borde de la ventana será azul sobre un fondo magenta y el texto que se escriba en la ventana será verde y tendrá un fondo azul.

c) Si se desea cambiar la posición de una ventana es necesario en la función wopen en los primeros cuatro parámetros se indiquen el renglón y la columna en que se quiereque comience la ventana y el renglón y la columna en que debe terminar. Ejemplo:

```
wopen(10,10,20,40,LCYAN|_BLUE,LGREEN|_BLUE);
```

en éste ejemplo se indica que la ventana tiene su esquina superior izquierda en el renglon 10, columna 10 y su esquina inferior derecha en el renglón 20 y la columna 40.

Cuando se han efectuado toas las modificaciones es necesario compilar el programa con el compilador C de Microsoft V. 6.0 y considerar en éste proceso las librerías cxlmss.lib, cxltcs.lib y cxlzts.lib. Para mayores detalles del manejo de las menús y ventanas vease el archivo CXL.DOC.

Apéndice B

Nomenclatura

A continuación se presenta un resumen de las expresiones que se utilizaron en el desarrollo de este trabajo.

$y = X\beta + \varepsilon$	: Modelo de Regresión Lineal
$\varepsilon \sim N(0, \sigma^2 I)$	: ε tiene una distribución normal multivariada, con vector de medias 0 y matriz de varianza-covarianza $\sigma^2 I$
n	: Número de observaciones consideradas en el modelo
p	: Número de variables consideradas en el modelo
X	: Matriz de variables explicativas, $X \in \mathbb{R}^{n \times p}$
β	: parámetros desconocidos, $\beta \in \mathbb{R}^p$
ε	: vector de errores observados, $\varepsilon \in \mathbb{R}^n$
y	: vector de observaciones o variable de respuesta $y \in \mathbb{R}^n$
$\hat{\beta}$	: coeficientes de regresión (estimación de β), se obtienen de la descomposición QR de la matriz X
$\hat{y} = X\hat{\beta}$	: estimación de y

$e = (y - X\hat{\beta})$: vector de residuales

$SCE = e'e$: suma de cuadrados de residuales

$SCT = y'y$: suma de cuadrados total

$SCR = SCT - SCE$: suma de cuadrados atribuible a la regresión

$s^2 = \frac{SCE}{n-p}$: cuadrado medio del error

r_{ij} : elemento del i-ésimo renglón y j-ésima columna de la matriz $(X'X)^{-1} = (R'R)^{-1}$ si $X = QR$, siendo Q una matriz ortogonal

$s^2(X'X)^{-1} = s^2(R'R)^{-1}$: estimación de la matriz de varianza-covarianza de $\hat{\beta}$

$se(\hat{\beta}_i) = s\sqrt{r_{ii}}$: la desviación estándar de $\hat{\beta}_i$

$cov(\hat{\beta}_i, \hat{\beta}_j) = s^2 * r_{ij}$ (sin considerar el renglón y la columna correspondiente a la constante)

$R^2 = \frac{SCR}{SCT}$: coeficiente de determinación

$F = \frac{SCR/p}{SCE/(n-p)}$: estadística F para probar Ho : $\beta = 0$

$t_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$: estadística t con n-p grados de libertad para probar la hipótesis $\beta_i = 0$

$rs_i = \dfrac{e_i}{s}$: residuales estandarizados

$h_{ii} = (X'(R'R)^{-1}X)_{ii}$: potencia de la observación i

$rt_i = \dfrac{e_i}{s\sqrt{1-h_{ii}}}$: residuales studentizados

r_{ii} : coeficiente de inflación de la varianza para cada $\hat{\beta}_i$

$D_i = \dfrac{1}{p}\, rt_i \dfrac{h_{ii}}{(1 - h_{ii})}$: distancias de Cook

$\rho_{i,j} = \dfrac{cov(\hat{\beta}_i, \hat{\beta}_j)}{se(\hat{\beta}_i)\ se(\hat{\beta}_j)}$: coeficiente de correlación

$C_p = \dfrac{SCR}{s^2} + 2p - n$: estadística C_p de Mallows

Si se desea centrar los datos se forma el modelo :

$$y^* = X^*\beta + \varepsilon$$

donde

$$x^*_{i,j} = x_{ij} - \bar{x}_j$$

$$y^*_i = y_i - \bar{y}$$

$$\beta_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j$$

con $\bar{x}, \bar{x}_i = \dfrac{\sum x_i}{n}$ es el vector de medias de las columnas de X

$$\bar{y} = \frac{\sum y_i}{n}$$ es la media de la variable de y

Si se desea estandarizar los datos se forma el modelo :

$$y^* = X^*\beta^* + \varepsilon$$

donde

$$x^*_{i,j} = \frac{x_{i,j} - \bar{x}_j}{s_j}$$

$$s_j^2 = \frac{\Sigma(x_{i,j} - \bar{x}_j)^2}{n-1}$$

$$y_i = \frac{y_i - \bar{y}}{s_y}$$

$$s_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$

$$\hat{\beta}_j = \hat{\beta}_j^* \ (s_y^2/s_j^2)^{1/2}$$

y $$\beta_0 = \bar{y} - \sum_{j=1} \beta_j \bar{x}_j$$

Apéndice C

Análisis de Error para Medias y Varianzas Muestrales

En este apéndice se presenta el análisis de error que realizaron Tony F. Chan y John G. Lewis [1], de los algoritmos para obtener la varianza denominados como: algoritmo de dos pasos, el de los libros de texto y el desarrollado por West en el que utilizan los algoritmos de Hanson para la media aritmética y la varianza. Pero antes es conveniente desarrollar el análisis de sensibilidad numérica para la media, varianza y desviación estándar muestrales.

C.1 Análisis de Sensibilidad para la Media, Varianza y Desviación Estándar Muestrales.

Sean $x,\ldots,x_m$ valores dados de m variables aleatorias $\alpha_1,\ldots,\alpha_m$ independientes e identicamente distribuida.
La media y la desviación estándar están dadas por:

$$\bar{x} = (\sum_{i=1}^{m} x_i) \ / \ m \qquad (C.1.1)$$

$$s = [(\sum_{i=1}^{m} x_i - \bar{x})^2 \ /(\ m - 1)]^{1/2}$$

En esta sección se estudia el análisis de sensibilidad (de primer orden) para $\bar{x}$ y s.

Para $\bar{x}$:

Tomando la diferencial total para $\bar{x}$, se tiene que

$$\delta \bar{x} = \sum_{i=1}^{m} \partial_{x_i} \bar{x} \cdot \delta x_i$$

con

$$\partial_{x_i} \bar{x} = \frac{1}{m} \qquad \text{para toda i}$$

Por lo que

$$\partial \bar{x} = (\sum_{i=1}^{m} \delta x_i) / m$$

Ahora, si $| \partial \bar{x} | \leq \eta | x_i |$ entonces

$$| \partial \bar{x} | \leq \eta \max_i \{ | x_i | \},$$

y en consecuencia,

$$\frac{| \partial \bar{x} |}{| \bar{x} |} \leq K_{\bar{x}} \eta \qquad \text{(C.1.3)}$$

donde

$$K_{\bar{x}} = \frac{\max\{ | x_i | \}}{| \bar{x} |} \geq (1) \qquad \text{(C.1.4)}$$

es el número de condición para $\bar{x}$.

Luego entonces los datos x_i serán mal comportados cuando algunos datos x_i están muy alejados de $\bar{x}$.

Para s:

Se empieza por el análisis de sensibilidad para la varianza muestral

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

Tomando la derivada total, se tiene que

$$\delta s^2 = \sum_{i=1}^{m} \partial_{x_i} s \cdot \delta x_i$$

$$= \frac{1}{m-1} \sum_{i=1}^{m} 2 (x_i - \bar{x}) \partial_{x_i} (x_i - \bar{x}) \cdot \delta x_i + O(\| \delta x \|^2)$$

donde

$$\| \delta x \|^2 = \sum_{i=1}^{m} | \delta x_i |^2$$

y como

$$\partial_{x_i} (x_i - \bar{x}) = 1 - \frac{1}{m} = \frac{m-1}{m} ,$$

se sigue que

$$\delta s^2 = \frac{2}{m} \sum_{i=1}^{m} (x_i - \bar{x}) \delta x_i + O(\| \delta x \|^2)$$

Por lo cual,

$$| \delta s^2 | \leq \frac{2}{m} | \sum_{i=1}^{m} (x_i - \bar{x}) \delta x_i | + O(\| \delta x \|^2)$$

$$\leq \frac{2}{m} (\sum_{i=1}^{m} (x_i - \bar{x})^2)^{1/2} (\sum_{i=1}^{m} |\delta x_i|^2)^{1/2} + O(\| \delta x \|^2)$$

Así, si se escribe

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

y

$$\delta x = (\delta x_1, \delta x_2, \ldots, \delta x_m)^t$$

entonces se obtiene que

$$| \delta s^2 | \leq \frac{2}{m} \| x - \bar{x} \|_2 \| \delta x \|_2 + O(\| \delta x \|^2)$$

Por lo tanto,

$$\frac{| \delta s^2 |}{| s^2 |} \leq \frac{2\sqrt{m-1}}{m} \left(\frac{\| x \|_2}{s} \right) \frac{\| \delta x \|_2}{\| x \|} + O(\| \delta x \|^2), \qquad (C.1.5)$$

ya que

$$\| x - \bar{x} \|_2 = \sqrt{m-1}\, s$$

y en donde

$$X = (x_1, x_2, \ldots, x_m)^t.$$

Por simplicidad, (C.1.5) la reescribiremos como:

$$\frac{| \delta s^2 |}{| s^2 |} \leq K_{s^2} \frac{\| \delta x \|_2}{\| x \|} + O(\| \delta x \|^2) \qquad (C.1.6)$$

donde

$$K_{s^2} = \frac{2}{\sqrt{m-1}} \frac{\| x \|_2}{s} \qquad (C.1.7)$$

es el número de condición para la variana muestral s^2.

Ahora, considerarel desarrollo de Taylor:

$$\delta(\sqrt{y_0}) \approx \sqrt{y} - \sqrt{y_0} = \frac{1}{2}\frac{1}{\sqrt{y_0}}(y - y_0) + O(|y-y_0|^2)$$

para $y_0 = s^2$, $y - y_0 = \delta s^2$, de (C.1.6) y (C.1.7) se obtiene que

$$\frac{|\delta s|}{|s|} \leq \frac{1}{2}\frac{|\delta s^2|}{|s^2|} + O(|\delta s^2|^2)$$

o sea que

$$\frac{|\delta s|}{|s|} \leq K_S \frac{\|\delta x\|_2}{\|x\|_2} + O(\|\delta x\|^2) \qquad (C.1.8)$$

donde

$$K_S = \frac{1}{\sqrt{m-1}}\frac{\|x\|_2}{s} \qquad (C.1.9)$$

resulta ser el número de condición para la desviación estándar. Nótese que

$$K_S = \frac{[(\sum_{i=1}^{m} x_i^2)^2 / (m-1)]^{1/2}}{[(\sum_{i=1}^{m} x_i^2 - m^{-1}(\sum_{i=1}^{m} x_i)^2) / (m-1)]^{1/2}}$$

$$= \left[\frac{\sum_{i=1}^{m} x_i^2}{\sum_{i=1}^{m} x_i^2 - m^{-1}(\sum_{i=1}^{m} x_i)^2}\right]^{1/2} \geq 1.$$

Así, los datos serán mal comportados para el cálculo de la desviación estándar muestral cuando la desviación s es muy pequeña comparada con los datos x_i. Esto es, cuando $x_i \approx \bar{x}$, para toda i.

C.2 Análisis de Error para el Algoritmo de Dos Pasos

1) Se calcula $\bar{x}$ por algún algoritmo, sea el resultado numérico $\hat{\bar{x}}$, y sea $\Delta\bar{x} = \hat{\bar{x}} - \bar{x}$

2) Se calcula la suma $\sum_{i=1}^{m}(x_i - \hat{\bar{x}})^2$, y se divide por (m-1).

Inicialmente,

$$t_0 = \hat{t}_0 = 0; \quad \Delta t_0 = 0$$

El paso general es

$$\hat{t}_i = fl(\hat{t}_{i-1} + [fl(x_i - \hat{\bar{x}})^2]$$

si

$$\hat{a} = fl(x_i - \hat{\bar{x}})^2 = (x_i - \hat{\bar{x}})^2 (1 + \eta_i)$$

donde $|\eta_i| \leq 3\varepsilon + O(\varepsilon^2)$

$$\hat{t}_i = fl(\hat{t}_{i-1} + \hat{a})$$

$$= (\hat{t}_{i-1} + \hat{a})(1 + \alpha_i), \quad \text{con } |\alpha_i| \leq \varepsilon$$

$$= \hat{t}_{i-1} + (x_i - \hat{\bar{x}})^2 + \alpha_i (\hat{t}_{i-1} + (x_i - \hat{\bar{x}})^2) + \eta_i \cdot (x_i - \hat{\bar{x}})^2 + O(\varepsilon^2)$$

Ahora reemplazando $\hat{t}_{i-1}$ por $t_{i-1} + \Delta t_{i-1}$, $\hat{\bar{x}}$ por $\bar{x} + \Delta\bar{x}$, para obtener

$$\hat{t}_i = [t_{i-1} + (x_i - \bar{x})^2] + \Delta t_{i-1} - 2(x_i - \bar{x})(\Delta\bar{x}) + (\Delta\bar{x})^2 + \alpha_i (t_{i-1} + (x_i - \bar{x})^2) + \alpha_i \Delta t_{i-1} - 2(x_i - \bar{x})(\Delta\bar{x}) \cdot \alpha_i + \alpha_i (\Delta\bar{x})^2 + \eta_i (x_i - \bar{x})^2 - 2\eta_i \cdot \Delta\bar{x} \cdot (x_i - \bar{x})^2 + \eta_i (\Delta\bar{x})^2 + O(\varepsilon^2).$$

Suponiendo que $\Delta\bar{x} = O(\varepsilon)$ y $\Delta t_{i-1} = O(\varepsilon)$, para eliminar los términos $O(\varepsilon^2)$:

$$\hat{t}_i = t_i + \Delta t_{i-1} - 2\Delta\bar{x} \cdot (x_i - \bar{x}) + \alpha_i t_i + \eta_i (x_i - \bar{x})^2 + O(\varepsilon^2)$$

$$\Delta t_i = \Delta t_{i-1} - 2\Delta\bar{x}\,(x_i - \bar{x}) + \alpha_i t_i + \eta_i (x_i - \bar{x})^2 + O(\varepsilon^2).$$

Por inducción,

$$\Delta T_m = -\sum_{j=1}^{m} (2\Delta\bar{x}(x_j - \bar{x})) + \sum_{j=1}^{m} \alpha_j t_j + \sum_{j=1}^{m} \eta_j (x_j - \bar{x})^2 + O(\varepsilon^2).$$

Pero el primer sumando es igual a cero.

$$\left| \sum_{j=1}^{m} \alpha_j t_j \right| \le \sum_{j=1}^{m} \varepsilon \cdot t_m \le m\, t_m\, \varepsilon$$

$$\left| \sum_{j=1}^{m} \eta_j (x_j - \bar{x})^2 \right| \le 3\, \varepsilon \sum_{j=1}^{m} (x_j - \bar{x})^2 \le 3\, \varepsilon\, t_m$$

por lo tanto, $| \Delta t_m | \le (m + 3)\, \varepsilon \cdot t_m + O(\varepsilon^2)$.

El paso final es

$$\hat{s}^2 = fl(\hat{t}_m / (m-1)) = (\hat{t}_m / (m-1))(1 + \delta),$$

así

$$| s^2 | \le (m + 3)\, \varepsilon \cdot s^2 + \varepsilon\, s^2 + O(\varepsilon^2).$$

$$= (m + 4)\, \varepsilon \cdot s^2 + O(\varepsilon^2).$$

El error relativo es

$$\boxed{\frac{| s^2 |}{s^2} \le (m + 4)\varepsilon + O(\varepsilon^2)}$$

que es independiente del número de condición k.

O sea, el algoritmo de dos pasos en una computadora con dígitos de guardia, tiene una cota de error en m solamente; la cota es independiente de k para la varianza muestral y del error al calcular $\bar{x}$, en tanto que el error es mucho menor que $| x |\, \varepsilon$.

C3. Análisis de Error para la fórmula de los libros de Texto de Estadística.

El algoritmo que se sigue consta de dos etapas que son:

1) Se calcula $\sum_{i=1}^{m} x_i$ y $\sum_{i=1}^{m} x_i^2$ y

2) Se calcula $s^2 = (\sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2/m)/(m-1)$

el error que se comete al calcular $t = \sum_{i=1}^{m} x_i$ es $\Delta t = \hat{t} - t$

donde

$$|\Delta t| = |\sum_{i=1}^{m} \eta_i x_i| + O(\varepsilon^2),$$

con $|\eta_i| \leq \begin{cases} (m-1-)\varepsilon & \text{para } i = 1 \\ (m-1-i)\varepsilon & \text{para } i > 1 \end{cases}$

Luego, se sigue que

$$|\Delta t| = \|\eta\| \|x\|$$

donde

$$\eta = \varepsilon[m-1, m-2, m-3, \ldots, 2, 1]^T$$

$$\|\eta\| = ((m-1)^2 + \sum_{i=1}^{m-1} i^2)^{1/2}\varepsilon = (\frac{m^3}{3} + \frac{m^2}{2} - \frac{11m}{6} + 1)^{1/2}\varepsilon$$

Similarmente, si

$$r = \sum_{i=1}^{m} x_i^2$$

$$\hat{r} = \sum_{i=1}^{m} x_i^2 + \sum_{i=1}^{m} \phi_i x_i^2 + O(\varepsilon^2),$$

donde

$$|\phi_i| \leq \begin{cases} (m-1-)\varepsilon \\ (m-1-i)\varepsilon, \quad i \geq 2 \quad i > 1 \end{cases}$$

De aquí, que

$$\hat{\Delta r} = \sum_{i=1}^{n} \phi_i x_i^2 .$$

y por lo tanto

$$| \Delta \hat{r} | \leq | \phi_{max} | \cdot \| x \|^2 \leq (m-1) \| x \|^2 \varepsilon \qquad (C.3.1)$$

Si no asumimos dígitos de guardia se calcula :

$\hat{a} = fl(\hat{t} \cdot \hat{t}) = (\hat{t}^2)(1+\alpha) = (t + \Delta t)^2(1+\alpha)$,

$\hat{b} = fl(\hat{a}/m) = (\hat{a}/m)(1+\beta) = (\hat{t}^2/m)(1+\alpha+\beta) + O(\varepsilon^2)$

$\hat{c} = fl(\hat{r}-\hat{b}) = \hat{r} - \hat{b} + \gamma \cdot \hat{r} - \gamma \cdot \hat{b}$

$$= \hat{r} - \frac{\hat{t}^2}{m} (1 + \alpha + \beta + \gamma) + \gamma \hat{r} + O(\varepsilon^2)$$

Finalmente

$$\hat{s}^2 = fl (\hat{c} / (m-1)) = (\hat{c}/(m-1))(1+\omega)$$

$$= \frac{\hat{r}}{(m-1)} (1+\gamma+\omega) - \frac{\hat{t}^2}{m(m-1)} (1+\alpha+\beta-\gamma+\omega) .$$

Ahora sustituyendo $r+\Delta r$ para t y $t+\Delta t$ para $\hat{t}$, y eliminando términos del orden de $\varepsilon\Delta r$, $\varepsilon\Delta t$ y ε^2:

$$\hat{s}^2 = \frac{r}{(m-1)} (1+\gamma+\omega) + \frac{\Delta r}{(m-1)} - \frac{(t+\Delta t)^2}{m(m-1)} (1+\alpha+\beta+\gamma+\omega) .$$

$$= \frac{r}{(m-1)} + (\gamma+\omega) \frac{r}{(m-1)} + \frac{\Delta r}{(m-1)} - \frac{t^2}{mm(m-1)} - \frac{\alpha+\beta+\gamma+\omega}{m(m-1)} t^2$$

$$- \frac{2t \cdot \Delta t}{m(m-1)} + O(\varepsilon^2) .$$

Entonces

$$\hat{s}^2 = s^2 + \Delta s^2$$

$$= \left(\frac{r - t^2/m}{(m-1)}\right) + \frac{1}{m-1}\left((\gamma+\omega)r + \Delta r - \frac{\alpha+\beta+\gamma+\omega}{m}\, t^2 - \frac{2t\Delta t}{m}\right) + O(\varepsilon^2),$$

por lo que

$$\Delta s^2 = \frac{1}{m-1}\left((\gamma+\omega)r + \Delta r - \frac{\alpha+\beta+\gamma+\omega}{m}\, t^2 - \frac{2t\Delta t}{m}\right)$$

$$= \frac{1}{m-1}\left(\gamma r + \Delta r - \frac{\alpha+\beta+\gamma}{m}\, t^2 - \frac{2t\Delta t}{m}\right) + \omega s^2$$

así que el algoritmo para los axiomas utilizando dígitos de guardia son los mismos hasta la operación fl (a_1+a_2) luego los términos δ y γ anteriores son iguales y lo que resta es como arriba. O sea:

$$\Delta s^2 = (\gamma + \omega)s^2 + \frac{1}{m-1}\left[\frac{(\alpha+\beta)}{m}\, t^2 + \Delta r - \frac{2t\Delta t}{m}\right]$$

$$|\Delta s^2| \le 2\varepsilon s^2 + \frac{2\varepsilon\ (m)\ |x|^2}{m(m-1)} + \frac{\varepsilon\ (m-1)\ |x|^2}{(m-1)}$$

$$+ \frac{2m\ (m-1)}{\sqrt{2}\ m(m-1)}\, |x|^2 \varepsilon + O(\varepsilon^2),$$

para toda $m \ge 8$. Ya que, para $m > 8$:

$$\frac{1}{2}\, m\ (m-1)^2 > \frac{1}{3}\, m^2 + \frac{1}{2}\, m^2 - \frac{11}{6}\, m + 1$$

Combinando términos se obtiene:

$$|\Delta s^2| \le 2\ \varepsilon\ s^2 + \frac{3m}{m-1}\, |x|^2 \varepsilon + O(\varepsilon^2)$$

o bien que

$$\boxed{\frac{|\Delta s^2|}{s^2} < 2\varepsilon + 3mk^2\varepsilon + O(\varepsilon^2).} \qquad (C.3.2)$$

C.4 Análisis de Error para el Algoritmo para la Media de Hanson

El algoritmo es

$$\bar{x}^{(1)} = x_1$$

$$\bar{x}^{(i+1)} = \frac{i\bar{x}^{(i)} + x_{i+1}}{i+1}$$

Numéricamente, esto es

$$M_1 = x_1$$

$$M_i = fl\ (\ fl((i-1)M_{i-1}+x_i)\ /\ i)$$

El primer paso es:

$$\hat{M}_1 = M_1 = x_1, \quad \text{por lo que } \Delta M_1 = 0$$

Para los pasos siguientes en el caso de dígitos de guardia. se calcula

$a = fl\ ((i-1)\cdot \hat{M}_{i-1}) = ((i-1)\ M_{i-1})(1+\alpha_i) \quad , \ |\alpha_i| \le \varepsilon$

$b = fl\ (a + x_i) = (((i-1)\hat{M}_{i-1})(1+\alpha_i)+X_i)(1+\beta_i) \quad , \ |\beta_i| \le \varepsilon$

$c = fl\ (b / i) = (((i-1)\hat{M}_{i-1})(1+\alpha_i)+X_i)(1+\beta_i)/i)\ (i + \gamma_i) \ , |\gamma_i| \le \varepsilon$

entonces :

$$\hat{M}_i = \frac{(i-1)\hat{M}_{i-1}(1+\alpha_i + \beta_i+\gamma_i) + X_i(1+\beta_i + \gamma_i)}{i} + O(\varepsilon^2)$$

$$= \frac{(i-1)(M_{i-1}+\Delta M_{i-1})(1+\alpha_i + \beta_i+\gamma_i) + X_i(1+\beta_i + \gamma_i)}{i} + O(\varepsilon^2)$$

$$= \frac{(i-1)M_{i-1}+ X_i}{i} + \frac{(i-1)}{i}[M_{i-1}\cdot(\alpha_i + \beta_i+\gamma_i) + \Delta M_{i-1}] + \frac{\beta_i+\gamma_i}{i}x_i + O(\varepsilon^2)$$

De aquí que

$$i\ \Delta M_i = (i-1)\Delta M_{i-1} + (i-1)M_{i-1}\cdot(\alpha_i + \beta_i+\gamma_i) + (\beta_i+\gamma_i)X_i+O(\varepsilon^2),$$

A partir de esto, por inducción

$$m\ \Delta M_m = \sum_{j=2}^{m} [(j-1)M_{j-1}(\alpha_1 + \beta_1+\gamma_1) + (\beta_1 + \gamma_1)x_j] + O(\varepsilon^2),$$

así

$$|m\ \Delta M_m| \leq \sum_{j=2}^{m} \sqrt{j-1} \cdot 3\varepsilon \cdot \| x \| + 2m \| x \| e + O(\varepsilon^2)$$

ya que $|M_j| \leq \frac{1}{\sqrt{j}} \|X\|$. Y como se tiene que

$$\frac{2}{3}(m^{3/2} - 1) = \int_1^m x^{1/2}\, dx \geq \sum_{j=1}^{m-1} \sqrt{j}$$

se sigue que

$$|m\ \Delta M_m| \leq (2m^{3/2} + 2m) \| x \| e.$$

Finalmente, se tiene que

$$\boxed{| \Delta M_m | \leq (2m^{1/2} + 2) \| x \| e.} \qquad (C.3.1)$$

C.5 Análisis de Error para la Varianza según Hanson-West

Análisis para el algoritmo de actualización de West, que se basa en el desarrollado por Hanson para la media aritmética.

El algoritmo es:

$$M_1 = x_1$$
$$T_1 = 0.$$

Para $i = 2,3,4,\ldots,m$

$$M_i = M_{i-1} + \frac{x_i - M_{i-1}}{i}$$

$$T_i = T_{i-1} + (i-1) \cdot (x_i - M_{i-1}) \cdot \left(\frac{x_i - M_{i-1}}{i} \right)$$

Finalmente, $s^2 = T_m/(m-1)$

El uso del término entre paréntesis $(x_i - M_{i-1})/i$ permite actualizar tanto a M como a T.

Se tiene que si

$$\hat{r} = fl(x_i - \hat{M}_{i-1}) = (x_i - \hat{M}_{i-1}) \cdot (1+\alpha_1),$$

$$\hat{e} = fl(\hat{r} / i) = \left(\frac{x_i - \hat{M}_{i-1}}{i} \right) (1 + \alpha_1 + \beta_1) + O(\varepsilon^2)$$

Entonces

$$\hat{a} = fl((i-1) * \hat{r}) = (i-1)(x_i - \hat{M}_{i-1})(1 + \alpha_1 + \gamma_1) + O(\varepsilon^2)$$

$$\hat{b} = \hat{a} \cdot \hat{e} = \frac{(i-1)}{i} (x_i - \hat{M}_{i-1})^2 (1 + 2\alpha_1 + \beta_1 + \gamma_1) + O(\varepsilon^2$$

Finalmente,

$$\hat{T}_i = fl(\hat{T}_{i-1} + \hat{b}) = (\hat{T}_{i-1} + \hat{b}) (1 + \delta_1) + O(\varepsilon^2)$$

$$= \left[\hat{T}_{i-1} + \frac{(i-1)}{i} (x_i - \hat{M}_{i-1})^2 \right] (1 + \alpha_1)$$

$$+ \frac{(i-1)}{i} (x_i - \hat{M}_{i-1})^2 (2\alpha_1 + \beta_1 + \gamma_1) + O(\varepsilon^2)$$

Sustituyendo $T_{i-1} + \Delta T_{i-1}$ por $\hat{T}_{i-1}$ y

$M_{i-1} + \Delta M_{i-1}$ por $\hat{M}_{i-1}$, se encuentra que

$$T_i = T_i + \hat{\Delta T_i}$$

$$= \left[T_{i-1} + \frac{(i-1)}{i} (x_i - M_{i-1})^2 \right] (1 + \delta_1) + 2 \frac{(i-1)}{i} \Delta M_{i-1} \cdot$$

$$(x_i - M_{i-1}) + \frac{(i-1)}{i}(x_i - M_{i-1})^2 (2\alpha_i + \beta_i + \gamma_i) + O(\varepsilon^2).$$

Luego,

$$\Delta T_m = \sum_{j=2}^{m} \delta_j T_j + \sum_{i=2}^{m} 2 \frac{(j-1)}{j} \Delta M_{j-1} \cdot (x_i - M_{i-1})$$

$$+ \sum_{j=2}^{m} (2\alpha_j + \beta_j + \gamma_j)(x_j - M_{j-1})^2 \frac{(j-1)}{j} .$$

Pero

$$\left| \sum_{j=2}^{m} \delta_j T_j \right| \le (m-1)\, \varepsilon T_m ;$$

$$\left| \sum_{j=2}^{m} (2\delta_j + \beta_j + \gamma_j)(x_j - M_{j-1})^2 \frac{(j-1)}{j} \right| \le 4\varepsilon T_m ;$$

y

$$\left| \sum_{j=2}^{m} 2 \frac{(j-1)}{j} \Delta M_{j-1} \cdot (x_j - M_{j-1}) \right|$$

$$\le 2 \sqrt{T_m} \left(\sum_{j=2}^{m} \frac{(j-1)}{j} |\Delta M_{j-1}|^2 \right)^{1/2} = 2 \sqrt{T_m} \cdot B_m$$

donde

$$B_n = \left(\sum_{j=2}^{m} \frac{(j-1)}{j} |\Delta M_{j-1}|^2 \right)^{1/2}$$

Ahora, como

$$\hat{s}^2 = (\hat{T}_m / (m-1))(1 + \eta)$$

$$= \frac{T_m}{m-1} + \frac{T_m}{m-1}\eta + \frac{\Delta T_m}{m-1} + O(\varepsilon^2),$$

se tiene que

$$|\Delta s^2| \le \varepsilon s^2 + (m-1)\,\varepsilon\, s^2 + 4\varepsilon s^2 + 2\, s \cdot \frac{B_m}{\sqrt{m-1}}$$

$$= (m+4)\,\varepsilon s^2 + 2\, s^2 \cdot \frac{B_m}{|x|} \cdot k$$

Y por tanto,

$$\frac{|\Delta s^2|}{s^2} \leq (m + 4)\,\varepsilon + 2 \cdot \frac{B_m}{|x|} \cdot k \qquad (C.5.1)$$

Así, aplicando (C.4.1) , se tiene que

$$B_m < \left[\sum_{j=1}^{m-1} |\Delta M_j|^2\right]^{1/2}$$

$$= 2\left[\sum_{j=1}^{m-1} (j^{1/2}+1)^2\right]^{1/2} |x|\varepsilon$$

$$= 2\left[\sum_{j=1}^{m-1} j + 2\left[\sum_{j=1}^{m-1} j^{1/2} + (m-1)\right]^{1/2} |x|\,\varepsilon\right.$$

$$< 2\left[\frac{m(m-1)}{2} + 2\,\frac{2}{3}\,(m^{1/2} - 1) + (m-1)\right]^{1/2} |x|\,\varepsilon$$

ya que

$$\sum_{j=1}^{m-1} j^{1/2} < \int_1^m x^{1/2}\,dx = \frac{2}{3}\,(m^{3/2} - 1).$$

Luego, después de completar cuadrados, se obtiene que

$$B_m < 2\left\{\left[\frac{m}{2} + \frac{2}{3}\sqrt{2m}\right]^2 - \frac{7}{18}\,m\right\}^{1/2} |x|\,\varepsilon$$

Y así que,

$$B_m < \left(\sqrt{2}\,m + \frac{4}{3}\sqrt{2m}\right)|x|\,\varepsilon \qquad (C.5.2)$$

así, finalmente de (C.5.1) y (C.5.2) se concluye que

$$\boxed{\frac{|\Delta s^2|}{s^2} \leq \left(2\sqrt{2}\,m + \frac{8}{3}\sqrt{2m}\right) k\varepsilon + (m + 4)\,\varepsilon.}$$

lo cual se quería establecer.

BIBLIOGRAFIA :

[1] Chan T. F. y J. G. Lewis, "*ROUNDING ERROR ANALISYS OF ALGORITHMS FOR COMPUTING MEANS AND STANDARD DESVIATIONS*", Johns Hopkins University Reporte Técnico No. 289 Abril, 1978.

[2] Chattierjee S., "*Regression Analysis by Sample*", John Wiley & Sons (1977).

[3] Dongarra J.J., Moler C.B., Bunch J.R. y Stewart G.W. "*LINPACK Users' Guide*", Siam (1979).

[4] Kahaner D. , Moler C. y Nash S. , "*Numerical Methods and Software*", Prentice-Hall, Inc. (1989).

[5] Montgomery D. C. y Peck E. A. , "*Introduction to Linear Regression Analysis*", John Wyley & Sons (1982).

[6] Mood A. M. , Graybill F. A. y Boes D. C. , "*Introduction to the theory of Statistics*" third edition, McGraw-Hill.

[7] Reichel L. y Gragg W.B., "*ACM Transc. Math. Softw.*", Vol. 16, No. 4, pags.366-377, December 1990.

[8] Searle S. R., "*Linear Models*", John Wiley & Sons (1971).

[9] Seber G.A.F., "*Linear Regression Analysis*", John Wiley & Sons (1977).

[10] Thisted Ronald A., "*Elements of Statistical Computing*", Chapman and Hall (1988).

[11] Weisberg Sanford, "*Applied Linear Regression*", John Wiley & Sons (1985).