8
2ej.

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

SISTEMA PARA DETECCION DE OUTLIERS PUNTOS DE ALTA PALANCA Y OBSERVACIONES INFLUYENTES EN EL MODELO DE REGRESION LINEAL MULTIPLE

T E S I S

QUE PARA OBTENER EL TITULO DE

M A T E M A T I C O

P R E S E N T A

JAIME GONZALEZ MARTINEZ

MEXICO, D. F. 1994

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

VNIVERSIDAD NACIONAL
AVFNºMA DE
MEXICO

M. EN C. VIRGINIA ABRIN BATULE
Jefe de la División de Estudios Profesionales
Facultad de Ciencias
P r e s e n t e

Los abajo firmantes, comunicamos a Usted, que habiendo revisado el trabajo de Tesis que realiz(ó)ron el pasante(s) Jaime González Martínez

con número de cuenta 8410575-6 con el Título:

Sistema para detección de outliers, puntos de alta palanca y observaciones influyentes en el modelo de regresión lineal multiple

Otorgamos nuestro Voto Aprobatorio y consideramos que a la brevedad deberá presentar su Examen Profesional para obtener el título de Matemático

GRADO	NOMBRE(S) APELLIDOS COMPLETOS	FIRMA
M. en C.	José Gabriel Huerta Gómez	
Director de Tesis		
M. en C.	José Antonio Flores Díaz	
M. en C.	Amparo López Gaona	
Act.	Claudia Lara Pérez Soto	
Suplente		
Fis.	Juan Jesús Gutiérrez García	
Suplente		

mayo/1994

A mi padre Julian González Villagomez
a mi madre Alicia Martínez Zamudio.

Agradesco también de manera muy especial a José Gabriel Huerta Gómez por su presión y apoyo, a Cristina Martínez Gallegos por su ayuda, a Alejandra Cortez quien me facilito los datos de su estudio de la cuenca de México y a todos los que de una forma u otra me han apoyado para realizar este trabajo.

Jaime González Martínez

SISTEMA PARA DETECCIÓN DE OUTLIERS, PUNTOS DE ALTA PALANCA Y OBSERVACIONES INFLUYENTES EN EL MODELO DE REGRESIÓN LINEAL MULTIPLE.

INDICE.

PRÓLOGO.

La intención de este trabajo fue crear una herramienta práctica sencilla y confiable que pueda ser utilizada por estudiantes y profesores para una área especifica de la estadistica, "La detección de puntos influyentes en modelos de regresión lineal". Pues este tipo de medidas o no se encuentran, o sólo las tienen ciertos programas muy especializados y de no fácil acceso.
Aunque aquí no están todas las medidas creadas para detectar influencia, si se encuentran algunas de las más representativas, y la bibliografía necesaria para recurrir a otras.
Por si alguna parte del programa de detección se llegara a necesitar en otro tipo de implementaciones, se adicionan los programas fuente.
El programa fue hecho en turbo C versión 2; utilizando una computadora personal con procesador 286.

INTRODUCCIÓN

Cuando se ajusta una línea recta a un conjunto de puntos es común ver que algunos de ellos se encuentran más lejanos, en comparación a la mayoría, de la recta ajustada; si estos puntos se omiten y se vuelve a calcular la recta, ésta puede cambiar notablemente o quedar prácticamente igual.

Es frecuente que cuando aparecen estos puntos lejanos a la mayoría de los demás, simplemente se les omita sin argumento alguno, sin embargo estos pueden describir una parte poco conocida del fenómeno en estudio. Por lo que es necesario determinar a través de algún criterio si su omisión repercute o no en el modelo ajustado.

Tener un sistema basado en varios criterios para poder determinar los puntos que influyen más fuertemente en el modelo, es el objetivo de la tesis. De tal manera que si alguno o algunos de los puntos lejanos son detectados como influyentes es una señal de su importancia. Por lo que ya no resulta tan conveniente quitarlos.

Aquí se trabaja con el modelo de regresión lineal múltiple, es decir si $\mathcal{Y}$ es una matriz de nx1, $\mathcal{X}$ de nxp con n>p, se busca una matriz $\hat{\beta}$ de px1, tal que la suma de sus diferencias al cuadrado de las entradas de $\mathcal{Y}$ con $\mathcal{X}\hat{\beta}$ sea mínima, de esta manera se tiene un programa que dadas las matrices $\mathcal{Y}$ y $\mathcal{X}$ calcula $\hat{\beta}$ y obtiene de acuerdo a los criterios DFBETAS, DFFIT y razón de verosimilitudes a los puntos que individualmente son influyentes y con la lamda de Wilk a aquellos que en grupo o individualmente también lo son.

El programa, que es parte de la tesis, fue hecho debido principalmente al poco software desarrollado al respecto y en apoyo a la investigación en este sentido.

En el primer capítulo se da un esbozo general de regresión lineal múltiple, con sus principales resultados y orientandolo hacia la

detección de observaciones influyentes.

En el segundo capítulo se introduce ya más en el tema y trata de las medidas utilizadas para la detección de esos puntos, y es útil en cuanto a su fundamentación teórica.

En el tercer y último capítulo, las primeras secciones muestran los algoritmos fundamentales para el cálculo de las medidas de influencia, sus puntos de corte y las limitantes del programa y en las secciones siguientes el uso del programa con un ejemplo de un estudio real de lluvia.

CAPITULO 1

REGRESION LINEAL MULTIPLE.

1.1 Relacion entre variables.

En este primer capitulo se da un bosquejo breve acerca de la regresión lineal múltiple, los problemas que ataca, soluciones, ejemplos y un caso particular de estudio el cual es el motivo de la tesis.
A grandes rasgos se puede decir que el objetivo de las técnicas de regresión es analizar la relación entre variables y para ilustrar esto se comenzará con un ejemplo.

1.1.1 Crecimiento de una planta.

La altura final que alcanza una planta en su desarrollo puede depender de la cantidad de agua en cada riego y de la materia orgánica que contenga la tierra.
En este ejemplo es de interés encontrar la relación de estos dos elementos con la altura final de la planta.
Supóngase entonces que

y .- La altura final de la planta.

x_1.- Cantidad de agua en cada riego.

x_2.- Proporción de materia orgánica contenida en la tierra.

Entonces se busca $f(\cdot)$ tal que

$$y=f(x_1,x_2).$$

Si además se tienen registros de la temperatura a la que se encuentra, la cantidad de agua en la atmósfera, salinidad de la tierra, etc. y se denota a estas variables por $x_3,x_4,\ldots,x_{p-1}$ entonces

$$y=f(x_1,x_2,\ldots,x_{p-1}) .$$

Conocer las variables restantes $(x_p,x_{p+1},\ldots)$ que influyen sobre y generalmente no es posible, y en ocasiones no es necesario, pues sólo se requiere estudiar la dependencia de y con las variables más importantes.

Al hecho de tener un conjunto de variables no especificadas dentro de $f(\cdot)$ o a la variabilidad que estas crean, es posible pensarla como un error aditivo al modelo y esta representarla con una variable aleatoria ε, de esta manera se tendrá que

$$y=f(x_1,\ldots,x_{p-1})+\varepsilon \ .$$

De acuerdo a lo anterior y denota a la variable dependiente o de respuesta; $x_1,x_2,\ldots,x_{p-1}$, a las variables independientes o regresores.

Por supuesto, es útil plantear este modelo en situaciones diferentes a las del ejemplo considerado. A partir de aquí y a menos que se indique lo contrario $y,x_1,x_2,\ldots,x_{p-1}$, son cualesquiera variables numéricas de tipo continuo.

1.2 Regresión lineal multiple.

1.2.1 Modelo notación y supuestos.

Se habla de análisis de regresión lineal múltiple cuando $f(\cdot)$ es de la forma

$$\beta_0+\beta_1x_1+\ \ldots\ +\beta_{p-1}x_{p-1}\ +\varepsilon$$

y $\beta_0,\beta_1,\ \ldots\ \beta_{p-1}\in\mathbb{R}$ son los parámetros desconocidos por estimar. Para estudiar este modelo se supondrá que se dispone de n observaciones, donde cada observación tendrá la forma:

$$(y_i,x_{i1},x_{i2},\ \ldots\ ,x_{ip-1})$$

con $i\in\{1,\ldots,n\}$.

Así la colección completa de observaciones es

$$\begin{aligned}&(y_1,x_{11},x_{12},\ \ldots\ ,x_{1p-1})\\&(y_2,x_{21},x_{22},\ \ldots\ ,x_{2p-1})\\&\qquad\vdots\\&(y_n,x_{n1},x_{n2},\ \ldots\ ,x_{np-1})\ .\end{aligned}$$

Para un manejo sencillo se requiere expresar matricialmente al modelo de regresión por lo que se considerará como

(1.1) $$y=X\beta+\varepsilon$$

Donde y será un vector de nx1, X una matriz de nxp, β un vector de px1, ε vector de nx1 y p es igual al número de parámetros

desconocidos a tratar en el modelo.

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p-1} \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np-1} \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{p-1} \end{bmatrix}, \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}.$$

El espacio de respuesta será el conjunto de valores que pueden tomar cada una de las entradas del vector y y el espacio de regresores será el conjunto de valores que pueden tomar cada una de las entradas de la matriz X.

El método usado para aproximar β, será el de mínimos cuadrados, es decir el estimador de β tal que $\sum(y_i-\hat{y}_i)^2$ sea mínimo, donde

(1.2) $$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots+\hat{\beta}_{p-1}x_{ip-1}.$$

y por supuesto

$$\hat{\beta}=\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_{p-1} \end{bmatrix}$$

es el estimador de β.

De este modo se puede probar (Draper y Smith, 1981)

(1.3) $$\hat{\beta}=(X^tX)^{-1}X^ty$$

Siendo X^t la matriz transpuesta de X y $(X^tX)^{-1}$ la matriz inversa de (X^tX). De manera general se usará M^{-t}, M^t, M^{-1}, para hablar de la inversa, la transpuesta y la inversa transpuesta de la matriz M respectivamente.

1.2.2 Estimación de intervalos de confianza.

Si además se supone que $E[\varepsilon_i]=0$, $var(\varepsilon_i)=\sigma^2$ (σ^2 constante, pero desconocida), $Corr(\varepsilon_i,\varepsilon_j)=0$ con $i\neq j$ y ε_i tiene una distribución Normal con media cero y varianza σ^2 para cada $i,j\in(1,\ldots,n)$ se obtiene (ver Montgomery y Peck, 1982)

(1.4) $$E[\hat{\beta}]=\beta$$

y

(1.5) $$Cov(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$$

que es una matriz de pxp cuyo j-ésimo elemento diagonal representa la varianza de β_j y las entradas fuera de la diagonal representan a la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$.

Un estimador insesgado de σ^2 es:

(1.6) $$\hat{\sigma}^2 = \frac{SCE}{n-p}$$

donde

(1.7) $$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e^t e \quad y \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

Con e vector de residuos. La matriz de varianzas de $\hat{y}$, vector de nx1 con i-ésima entrada $\hat{y}_i$, resulta ser

(1.8) $$Cov(\hat{y}) = Cov(X\hat{\beta}) = X(X^t X)^{-1} X^t \sigma^2$$

También se puede probar que (Montgomery y Peck, 1982)

(1.9) $$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \, k_{jj}}} \sim T(n-p) \ ,$$

Con k_{jj}, el j-ésimo elemento diagonal de la matriz $(X^t X)^{-1}$ y $T(n-p)$ la distribución t de Student con n-p grados de libertad. Por lo que un intervalo de confianza al $(1-\alpha)$ porciento para β_j $j \in \{0,1,\ldots,p-1\}$ es

(1.10) $$\hat{\beta}_j - T_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \, k_{jj}} \leq \beta_j \leq \hat{\beta}_j + T_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \, k_{jj}}$$

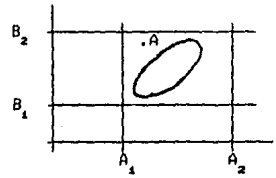
Con $T_{1-\alpha/2,n-p}$ el cuantil de la distribución t que acumula $1-\alpha/2$ de probabilidad.

Ocurre sin embargo que el producto cartesiano de intervalos para cada β_i no necesariamente representa el $(1-\alpha)$ porciento de confianza para β, en cambio un conjunto cuya confianza si es del $(1-\alpha)$ porciento, esta determinado por las entradas de β que caen dentro del elipse definido por

(1.11)
$$\frac{(\beta-\hat{\beta})^t (X^t X) (\beta-\hat{\beta})}{p\hat{\sigma}^2} \leq F_{\alpha,p,n-p}$$

donde $F_{\alpha,p,n-p}$ es el cuantil $(1-\alpha)$ de la distribución F con p y n-p grados de libertad respectivamente. Como ilustración, para el caso de dos dimensiones puede verse lo siguiente:

FIGURA 1



Elipse de confianza para las entradas de β. Aquí se muestra gráficamente como existen puntos que caen dentro de ambos intervalos y sin embargo su confianza conjunta no es del $1-\alpha$ porciento.

A_1,B_1, son los límites inferiores de los intervalos de confianza de β_0,β_1 respectivamente y A_2,B_2 los superiores. El punto A cae dentro de ambos intervalos de confianza sin embargo no cae dentro del elipse de confianza.

También es necesario presentar un intervalo para $\hat{y}_i$ que permita predecir sus valores con cierta probabilidad.

Un estimador es $\hat{y}_i=x_i\hat{\beta}$ siendo $x_i=(1,x_{i1},x_{i2},\ldots,x_{ip-1})$ igual al i-ésimo renglón de X. No esta por demás hacer notar que $\hat{y}_i$ es un estimador insesgado, pues $E[\hat{y}_i]=x_i\beta=y_i$.

El intervalo de predicción al $(1-\alpha)\times100\%$ es (1.12)

$$\hat{y}_i - T_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_i (X^t X)^{-1} x_i^t} \leq y_i \leq \hat{y}_i + T_{1-\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_i (X^t X)^{-1} x_i^t}.$$

Por último, conviene señalar que $(SCE/\sigma^2)\sim\chi^2(n-p)$.

1.2.3 Coeficiente de determinación.

Es muy común medir lo adecuado de un modelo de regresión a través del coeficiente de determinación múltiple definido como

$$(1.13) \qquad R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

con

$$(1.14) \qquad \bar{y} = \sum_{i=1}^{n} y_i / n \ .$$

El coeficiente suele interpretarse como una medida de la reducción de la variabilidad de la respuesta obtenida usando los regresores $x_1, x_2, \ldots, x_{p-1}$. $R = \sqrt{R^2}$ se conoce como el coeficiente de correlación múltiple entre $x_1, x_2, \ldots, x_{p-1}$ y la y.
Es necesario mencionar que $0 \leq R^2 \leq 1$ y que un R^2 cercano a uno no necesariamente implica que el modelo de regresión sea bueno ya que a medida que se añaden regresores, R^2 crece. La mayor utilidad de R^2 es para medir el peso de cada regresor al modelo en un proceso de selección de variables.

1.2.4 Pruebas de hipótesis.

Hasta el momento sólo se ha presentado la forma de estimar a las entradas del vector β, pero puede resultar que el modelo no represente al fenómeno en estudio. Entonces lo que se hace es proponer una hipótesis y con base a esta construir una zona de probabilidad, aunque para datos fijos se hable de confianza; las hipótesis tratarán de resolver algunos cuestionamientos como los siguientes:

1) ¿El modelo lineal describe al fenómeno en estudio?

2) ¿Existen ciertas variables cuya contribución al modelo es nula?

3) ¿Existen ciertos parámetros iguales ?.

Traduciendo esto se tiene

1) H_o: $\beta=0$ contra: H_a: Existe al menos una $\beta_i \neq 0$.

2) $H_o:\beta^* = \bar{0}^*$ contra $H_a:\beta^* \neq \bar{0}^*$ Con β^* un vector formado con algunos de los parámetros del vector β.

3) H_o: $\beta' = \beta''$ contra $H_a:\beta' \neq \beta''$ Con β' y β'' vectores de la misma dimensión y β' formado con algunas entradas de β.

Para poder seguir adelante es necesario tomar en cuenta que si el modelo reducido o el modelo bajo la hipótesis nula es válido

$$F = \frac{[\,SCE(MR) - SCE(MC)\,]\,/\,k}{SCE(MC)/(n-p)} \qquad (1.15)$$

entonces F tiene una distribucion F con p-k y n-p grados de libertad siendo $SCE(MC)=\sum(y_i-\hat{y}_i)^2$, $\hat{y}_i$ el estimador de y_i con el modelo completo (MC) y $SCE(MR)=\sum(y_i-\hat{y}_i^*)^2$ con $\hat{y}^*$ el estimador de y_i con el modelo reducido (MR), y k es el número de entradas de β que se consideran igual a cero. Debido al comportamiento de la F y suponiendo que $H_o:\beta=0$ es cierta, entonces se tiene que

$$F_1 = \frac{\left[\sum(y_i-\bar{y}_i)^2 - \sum(y_i-\hat{y}_i)^2\right]/(k)}{\sum(y_i-\hat{y}_i)^2/(n-p)} \qquad \text{entonces } F_1 \sim F(k,n-p)\ .$$

Si se quiere probar $H_o:\beta=\bar{0}$ contra $H_1:\beta\neq\bar{0}$ primero se selecciona a_1 tal que

$$P\{F_1 \leq a_1\} = 1-\alpha \qquad (1.16)$$

y dada la muestra se acepta H_o si $F_1 \leq a_1$ y se rechaza en caso contrario todo con un nivel de confianza de tamaño α.

El principio del cociente de verosimilitudes justifica esta regla de decisión para probar la hipótesis de interés. Ver (Chatterjee y Price 1977)

1.2.5 Análisis de residuos.

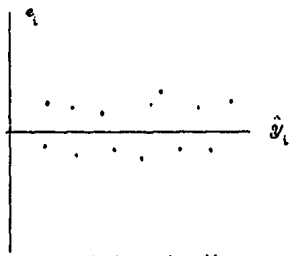
Revisar el comportamiento de los residuos es importante para corroborar o descartar las suposiciones acerca del modelo. Este

análisis se basa principalmente en la revisión de algunas gráficas como:

1) La prueba del papel normal en la cual las entradas del vector e se ordenan de menor a mayor y se grafican contra $(i-1/2)/n$ $i=1,2,\ldots,n$. Si lo que resulta es más o menos una línea recta, entonces es admisible suponer que la distribución de las entradas del vector *e* es normal, de no ser así indica carencia de normalidad.
2) Residuos contra el tiempo (si es conocido). Aquí se ordenan los residuos cronológicamente, y se grafican contra el tiempo, esto sirve para detectar si existe alguna dependencia entre los errores.
3) Residuos contra las entradas del vector $\hat{y}$. Se grafican los residuos contra los valores estimados de $\hat{y}$ y no contra las entradas de y porque los residuos y las entradas de y están correlacionadas. A grandes rasgos existen cuatro tipos de gráficas (Ver fig 2)

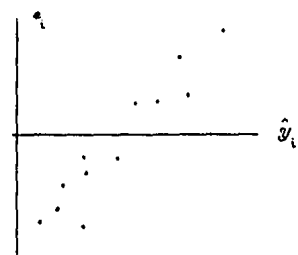
FIGURA 2

2.a



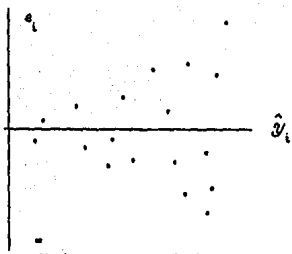
Comportamiento deseable.

2.b



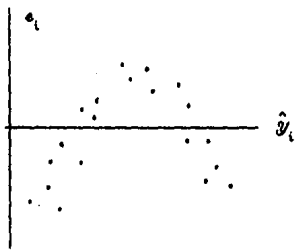
Error en el análisis.

2.c



Varianza no constante.

2.d



Indica la posibilidad de que el modelo necesite alguna transformación.

De este modo:

2.a Indica una distribución que sería lo ideal esperar si los supuestos son correctos.

2.b Indica error en el análisis, puede deberse a la omisión de β_0.

2.c Es una indicación de que σ^2 no es constante y que se va incrementando. y

2.d Es posible que el modelo necesite alguna transformación o falta agregar alguna otra variable.

4) Residuos contra cada una de los regresores. Aquí también lo deseable es la banda 2.a (de la figura 2), 2.c denota nuevamente que σ^2 no es constante, mientras que 2.b representa la necesidad de adicionar alguna variable más y 2.d indica la necesidad de hacer una transformación en el modelo.

El siguiente ejemplo se hizo tomando como referencia algunos puntos del plano z=-x+2y+1 y aumentole los errores e_i (obtenidos de un generador de números pseudo-aleatorios) y esta hecho sólo para mostrar los numéricamente los resultados anteriores.

Ejemplo 1.1

$$\mathcal{Y} = \begin{bmatrix} 5.01000 \\ 7.23000 \\ 4.77000 \\ 0.43000 \\ 8.85000 \\ 5.08000 \\ 1.66000 \\ 0.97000 \\ 2.30000 \\ 7.92000 \\ 2.70000 \\ 8.15000 \\ 2.83000 \end{bmatrix}, \qquad \mathcal{X} = \begin{bmatrix} 1.0000 & 2.000 & 3.0000 \\ 1.0000 & 4.500 & 5.2000 \\ 1.0000 & 6.000 & 5.0000 \\ 1.0000 & 6.200 & 2.8000 \\ 1.0000 & 2.000 & 5.0000 \\ 1.0000 & 4.000 & 4.0000 \\ 1.0000 & 3.000 & 2.0000 \\ 1.0000 & 4.000 & 2.0000 \\ 1.0000 & 7.000 & 4.0000 \\ 1.0000 & 1.600 & 4.3000 \\ 1.0000 & 4.000 & 3.0000 \\ 1.0000 & 3.000 & 5.0000 \\ 1.0000 & 6.000 & 4.0000 \end{bmatrix}$$

con estos datos se obtiene

$$\frac{\mathrm{Cov}(\hat{\beta})}{\sigma} = (\mathcal{X}^t\mathcal{X})^{-1} = \begin{bmatrix} 1.42921 & -0.10542 & -0.24262 \\ -0.10542 & 0.02652 & -0.00088 \\ -0.24262 & -0.00088 & 0.06493 \end{bmatrix}$$

y

$$\hat{\beta} = \begin{bmatrix} 0.60488 \\ -0.97353 \\ 2.06745 \end{bmatrix}$$

De esta forma: $\hat{y} = (0.60488) + (-0.97353)x_1 + (y2.06745)x_2$ como $k_{11}=1.4292, k_{22}=0.02652, k_{33}=0.06493$ y $\hat{\sigma^2}=SCE/(n-p)=.4353/10=0.04353$ tomando $\alpha=0.05$, $t(.975,10)=2.228$ y entonces

$$\beta_0 \in [\ 0.0485,\ 1.1594]$$
$$\beta_1 \in [-1.0479, -0.8980]$$
$$\beta_2 \in [\ 1.9490,\ 2.1857]$$

son los intervalos de confianza al 95%.

Haciendo la prueba de hipótesis $H_0: \beta = \bar{0}$ contra H_1: existe al menos

una $\beta_i \neq 0$

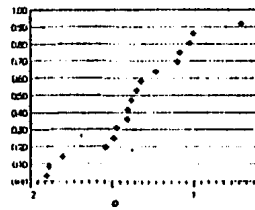
$$F = \frac{(99.7963 - 0.39034)/2}{(0.39034)/10} = \frac{49.80298}{0.039034} = 1275.88$$

así tomando $\alpha=0.05$, el cuantil para la $F(10,2)$ es $a_1=5.46$ y $a_1 \leq F$ con lo cual se rechaza la hipótesis nula.

También se obtiene $R^2=0.99565$, y las gráficas para residuos aparecen en la figura 3.

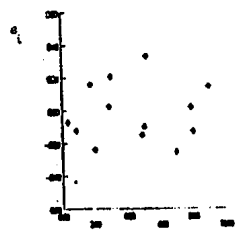
FIGURA 3

3.a



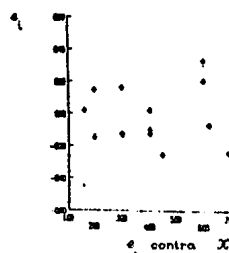
Papel normal

3.b



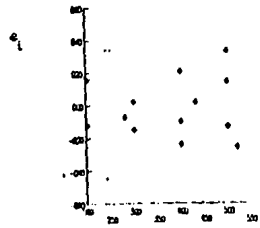
e_i contra $\hat{y}$

3.c



e_i contra x_i

3.d



e_i contra X_2

Revisando lo anterior, por la prueba de hipótesis, existe una variable (X_1 ó X_2) que es significativa bajo el modelo lineal.
R^2 también refleja una buena relación entre los regresores y la matriz $\hat{y}$.
No se rechaza el supuesto de normalidad, después de observar la gráfica de 3.a y al observar las demás figuras 3.b 3.c y 3.d. No hay evidencia de violaciones en los supuestos.
Existen también otras pruebas para detectar correlación en los residuos como son la de Durbin Watson y Rachas (Draper and Smith 1981)

13 IRREGULARIDADES.

En ocasiones al ajustar el modelo de regresión, no se rechaza $H_0:\beta=\bar{0}$ o la varianza es muy grande o varia de acuerdo a cierto patrón de comportamiento desconocido e inestable, pero entre todas las irregularidades que existen, hay algunas a las que aquí se les dará mayor importancia.
Hay puntos (uno o varios) para los cuales su residuo correspondiente en valor absoluto es muy grande en comparación con el resto. A estos puntos de aquí en adelante se les llamará outliers. Existen otros que estan ubicados fuera del sitio donde se encuentran los demás puntos X's. A estos puntos se les llamará "Puntos de Alta Palanca".
Ahora este tipo de puntos pueden o no influir fuertemente en el modelo, en el sentido de que si se omiten, el resultado del modelo

cambia notablemente (en los valores del vector β) si esto sucede, se les llamará "Puntos influyentes".

En este pequeño ejemplo se muestra la diferencia entre un modelo ajustado en base a una muestra de datos, y el modelo ajustado omitiendo una observación. Estos datos son puramente de ejemplo.

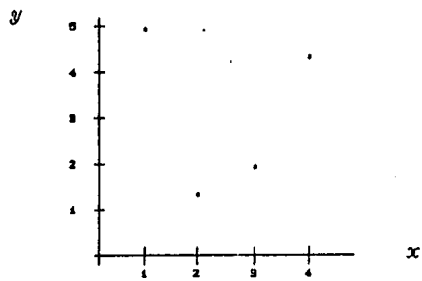
Ejemplo 1.2
Si

$$y = \begin{bmatrix} 5.00 \\ 2.50 \\ 2.80 \\ 4.30 \end{bmatrix} \qquad x = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

el modelo resultante es $\hat{y} = (3.95)+(-0.15)x_1$. Sin embargo si se omite la observación número 1 (1,5.00) de la muestra, (vease figura 4) y se vuelve a realizar la regresión el modelo queda como: $y=(-0.05)+(1.05)x_1$.

El cambio en los estimadores es notable.
A este tipo de puntos son a los que se va a dedicar mayor atención en lo subsecuente.

FIGURA 4



El punto (1,5.00) causa desviación en el ajuste.

1.4 Referencia.

- CHATTERJEE, S. and PRICE B. (1977). Regression Analysis by example, Wiley New York.

- DRAPER, N.R. and SMITH, H. (1981). Applied Regression Analysis 2nd Ed. Wiley New York. p 141-187

- MONTGOMERY, DOUGLAS C. and PECK ELIZABETH. (1982) Introduction to Linear Regression Analysis, Capitulos 1,2,3,4. Wiley New York.

CAPITULO 2

MEDIDAS DE INFLUENCIA.

2.1 Observaciones influyentes.

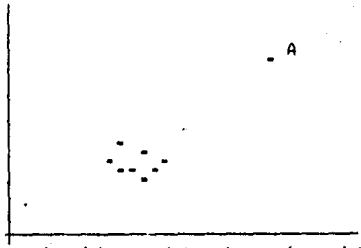
En el capítulo anterior se dió un breve esbozo de lo que es un punto influyente, sin embargo para los fines que aquí se persiguen, es necesario estudiar más a fondo este concepto.
Intuitivamente un punto influyente es aquel que afecta al modelo en una forma más notable que el resto. Es decir, los resultados de la regresión son visiblemente diferentes cuando éste se omite.
Una definición informal de punto influyente dada por Belsley, Kuh, Welsch (1980), es la siguiente:

> Observación influyente es aquella que sola o junto con otras, tiene un gran impacto demostrable sobre los valores calculados de varios de los aspectos de interés del modelo (coeficientes, errores estandarizados, valores de t, etc), en comparación con las demás observaciones

De esta definición cabe resaltar que la influencia la puede ejercer una sóla observación o un grupo y se puede definir ya sea sobre las entradas del vector $\hat{\beta}$, o del vector $\hat{y}$, sobre $\hat{\sigma}^2$ o algún otro aspecto de interés.

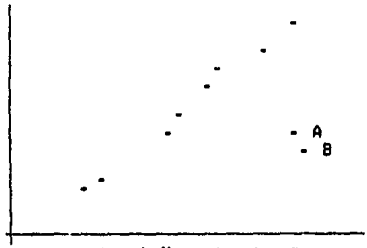
Para dar una idea gráfica en dos dimensiones, de las situaciones que se pueden presentar, se tienen las siguientes figuras

FIGURA 2.1



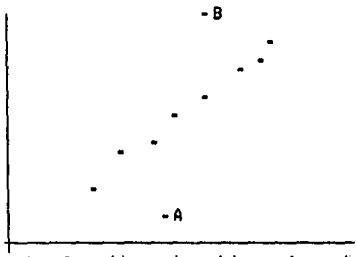
El modelo se determina casi completamete por el punto A.

FIGURA 2.2



Dos puntos influyentes A y B

FIGURA 2.3



A y B cambian al modelo en forma distinta.

En la figura 2.1 si no se toma al punto A, la recta ajustada, por mínimos cuadrados, será muy diferente a que si se toma. Este es un claro ejemplo de un conjunto de puntos con uno influyente.
En 2.2 los puntos A y B son los que ejercen el palanqueo suficiente como para cambiar el valor de la pendiente de la recta ajustada.

Este ejemplo también resulta interesante, por que si se borra el punto A o el punto B y se vuelve a realizar la regresión, estos individualmente no resultan ser influyentes, pues su influencia es en conjunto y uno cubre el efecto del otro. En 2.3 se muestra un ejemplo de influencia en grupos, pero con la diferencia respecto de 2.2 de que uno solo A o B o ambos son puntos influyentes.

2.2 OUTLIERS.

Los puntos lejanos a la nube principal de datos ya sea en el espacio de regresores, y/o en el espacio de respuesta, pueden ser influyentes.

Los puntos alejados en el espacio de respuesta, que en lo sucesivo se les llamará **outliers**, se pueden detectar al estudiar los residuos del modelo. Los puntos distantes en el espacio de regresores, que en lo sucesivo se les llamará **puntos de alta palanca**, se detectan al revisar las entradas de la diagonal de la matriz $\mathcal{P}$, donde

$$\mathcal{P}=X(X^tX)^{-1}X^t \tag{2.1}$$

Para poder detectar la presencia de outliers es necesario considerar algunos aspectos adicionales sobre los residuos. El residuo

$$e_i=y_i-\hat{y}_i \tag{2.2}$$

se puede entender como el estimador del i-ésimo error, ademas

$$\begin{aligned} e&=y-\hat{y}\\ &=y-X\hat{\beta}\\ &=y-X[(X^tX)^{-1}X^ty]\\ &=(I-\mathcal{P})y \end{aligned} \tag{2.3}$$

Como $E[y]=X\beta$, entonces

$$\begin{aligned} e-E[e]&=(I-\mathcal{P})y-E[y-\mathcal{P}y]\\ &=(I-\mathcal{P})(y-X\beta)\\ &=(I-\mathcal{P})\varepsilon \end{aligned} \tag{2.4}$$

De aqui se obtiene que la matriz de varianzas y covarianzas de e es

$$V(e)=E\{[e-E(e)][e-E[e]]^t\}=(I-\mathcal{P})E(\varepsilon\varepsilon^t)(I-\mathcal{P}) \tag{2.5}$$

entonces como

$$V(\varepsilon)=I\sigma^2,$$

se tiene

$$V(e) = (I-\mathcal{P})\,I\sigma^2\,(I-\mathcal{P})$$

(2.6) $$= (I-\mathcal{P})\sigma^2 \quad ;$$

de donde se concluye que

(2.7) $$V(e_i) = (1-\mathcal{P}_{ii})\sigma^2 \quad ,$$

Si $i \neq j$

(2.8) $$Cov(e_i, e_j) = -\mathcal{P}_{ij}\sigma^2 \quad ;$$

y

(2.9) $$Corr(e_i, e_j) = \frac{Cov(e_i, e_j)}{\sqrt{V(e_i)V(e_i)}} \quad .$$

Si a los residuos no se les estandariza su valor variará en gran medida cuando se trate de un experimento en el que los residuos esten en unidades distintas.

De tal manera que para eliminar el efecto de las unidades al observar los residuos, se toma generalmente

(2.10) $$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-\mathcal{P}_{ii}}}$$

$\hat{\sigma}$ dada en (1.6) Aunque para los casos se discuten a continuación se usará

(2.11) $$t_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-\mathcal{P}_{ii}}}$$

donde $\hat{\sigma}_{(i)}$ es el estimador de σ sin tomar en cuenta la i-ésima observación, asimismo $\hat{\beta}_{(i)}$ y $\hat{\mathcal{Y}}_{(i)}$ serán los estimadores de β y de $\mathcal{Y}$ respectivamente sin considerar la i-ésima observación y de manera análoga $\mathcal{X}_{(i)}, \mathcal{Y}_{(i)}$ serán la matriz $\mathcal{X}$ y la matriz $\mathcal{Y}$ respectivamente sin la i-ésima observación.

A t_i^* también se le conoce como el residuo estudentizado debido a su comportamiento similar a una t de Student con n-p-1 grados de libertad, (ver Belsley, Kuh, Welsch 1980 pág 20) y esto a su vez da el criterio para decidir cuando un residuo es grande, lo cual no es posible evaluar directamente con e_i.

Los residuos que pueden corresponder a un outlier serán aquellos para los cuales $|t_i^*|$ es mayor a dos, pues de acuerdo a la distribución T, los $|t_i^*|>2$ ocurren con un 5% de probabilidad.
Al valor 2 se le puede tomar como una cota en general, pero aquí para mayor precisión se recurrirá al cuantil especifico de la distribución T.

Es importante tener en cuenta que un outlier no necesariamente es un punto de influencia y viceversa. Por ejemplo, si en una regresión lineal simple, la mayoria de las observaciones determinan una recta, y además existe una observación que por su cercanía a $\bar{x}$ no afecta tanto a β, dicha observación puede ser un outlier, pero no influyente sobre β_1. (Ver figura 2.4)

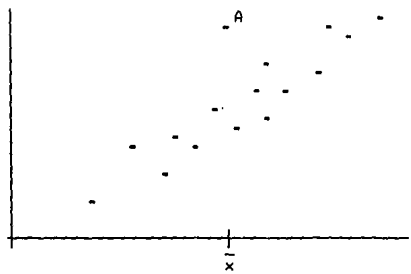


FIGURA 2.4

Un ejemplo de un punto influyente que tiene un residuo pequeño aparece en el ejemplo dos del capítulo uno.

2.3 PUNTOS DE ALTA PALANCA.

En el caso de que los regresores sean independientemente distribuidos y bajo los supuestos de Normalidad, se llega a un resultado obtenido por Besley, Kuh, Welsch (1980) en donde

demuestran que

$$(2.12) \qquad \frac{(n-p)(\mathcal{P}_{ii}-1/n)}{(p-1)(1-\mathcal{P}_{ii})} \sim F_{(p-1,n-p)}$$

De esta manera si $F^{\alpha}_{(p-1,n-p)}$ es el cuantil de una F con p-1 y n-p grados de libertad respectivamente, para la cual se acumula $(1-\alpha)$ de probabilidad entonces

$$(2.13) \qquad \mathcal{P}_{ii} < \frac{\left(\frac{p-1}{n-p}\right)F^{\alpha}_{(p-1,n-p)} + \frac{1}{n}}{1 + \left(\frac{p-1}{n-p}\right)F^{\alpha}_{(p-1,n-p)}}$$

Las observaciones para las cuales su $\mathcal{P}_{ii}$ correspondiente no cumpla con (2.13) para una α dada, serán consideradas como de alta palanca. Para p>10 y n-p>50 con α=0.05 se puede tomar a $F^{\alpha}_{(p-1,n-p)}$=2, pues ya se acumula un 95% de probabilidad, así sustituyendo y simplificando (2.13) se obtiene que

$$(2.14) \qquad \mathcal{P}_{ii} < \frac{2np-n-p}{n(n+p-2)}$$

Una cota ligeramente mayor y buena para cuando n es grande se obtiene de (2.14) tomando a n^2 en vez de n(n+p-2) en el denominador y esta resulta ser 2p/n; entonces los $\mathcal{P}_{ii}$>2p/n serán considerados como puntos de alta palanca. Así se maneja por algunos autores, (Belsey Kuh y Welsh, 1980; Chaterjee y Hadi 1982) pero aquí se optó por usar (2.13)

Cuando no hay puntos de alta palanca los valores de la diagonal de la matriz $\mathcal{P}$ son similares a p/n pues

$$(2.15) \qquad \sum_{i=1}^{n} \mathcal{P}_{ii} = p$$

Ver Besley, Kuh, Welsch págs 66 y 67, y este resultado se ocupará más adelante.

Es posible encontrar un punto de alta palanca que no necesariamente

sea influyente. Para ver esto considere el siguiente ejemplo

EJEMPLO 2.1

sea

$$\mathcal{X} = \begin{bmatrix} 1.0 & 5.0 & 3.0 \\ 1.0 & 5.0 & 4.0 \\ 1.0 & 6.0 & 2.3 \\ 1.0 & 5.5 & 4.1 \\ 1.0 & 2.0 & 7.0 \\ 1.0 & 6.4 & 3.0 \end{bmatrix} \qquad \mathcal{Y} = \begin{bmatrix} 6.02 \\ 0.98 \\ 13.48 \\ 2.49 \\ -26.00 \\ 11.53 \end{bmatrix}.$$

si se ajusta el modelo $\mathcal{Y}=\mathcal{X}\beta+\varepsilon$ por mínimos cuadrados

$$\hat{\beta} = \begin{bmatrix} 1.26 \\ 3.96 \\ -5.02 \end{bmatrix};$$

$\hat{\sigma}^2=0.000625$, $\hat{\sigma}=0.025$, con

$$\mathcal{P} = \begin{bmatrix} 0.5719 & 0.1133 & 0.4412 & -0.1584 & 0.0925 & -0.0604 \\ 0.1133 & 0.1737 & 0.1320 & 0.2103 & 0.1718 & 0.1988 \\ 0.4412 & 0.1320 & 0.4355 & -0.0099 & -0.1290 & 0.1302 \\ -0.1584 & 0.2103 & -0.0099 & 0.4505 & 0.0968 & 0.4108 \\ 0.0924 & 0.1718 & -0.1290 & 0.0968 & 0.9078 & -0.1399 \\ -0.0604 & 0.1988 & 0.1302 & 0.4108 & -0.1399 & 0.4605 \end{bmatrix}.$$

Como p=3,n=6, si $\alpha=0.1$ $F^{\alpha}(2,3)=9.16$ la cota que se obtiene de acuerdo a (2.13) es 0.8827 y los elementos de la diagonal $\mathcal{P}$ mayores a este número estan asociados a puntos de alta palanca. La observación número cinco resulta ser de alta palanca y si se vuelve a realizar la regresión sin este dato se obtiene que

$$\hat{\beta}_{(5)} = \begin{bmatrix} 1.13 \\ 3.98 \\ -5.01 \end{bmatrix},$$

muy parecido a $\hat{\beta}$. En conclusión la observación 5 no es influyente en $\hat{\beta}$, pero si es de alta palanca.

También sucede que un punto influyente no implica necesariamente que este sea de alta palanca.

2.4 Medidas de influencia para una sóla observación.

Las medidas que enseguida se van, a ver revisan diferentes aspectos de la regresión, para así poder hacer un análisis más completo. Fueron seleccionadas debido a ser computarizables, de sencillo entendimiento, escala-invariante (es decir que sin importar el tipo de unidades con que se este midiendo, los resultados son los mismos) y por lo regular tratan a los puntos tanto individualmente como en grupo.

2.4.1. DFBETAS

Una forma natural de revisar como cambia el modelo cuando se incluye a una observación y cuando no, es comparando $\hat{\beta}$ y $\hat{\beta}_{(i)}$ o de otra manera revisando el vector

$$\text{DFBETA}_i = \hat{\beta} - \hat{\beta}_{(i)} \tag{2.16}$$

$$= (x^t x)^{-1} x^t y - (x^t_{(i)} x_{(i)})^{-1} x^t_{(i)} y_{(i)}. \tag{2.17}$$

Haciendo uso de la igualdad Sherman-Morrison-Woodbury [Rao (1973) p.33], si $\mathcal{A}$ es una matriz no singular y u y v son dos vectores columna entonces

$$(\mathcal{A} - uv^t)^{-1} = \mathcal{A}^{-1} + \frac{\mathcal{A}^{-1} u v^t \mathcal{A}^{-1}}{1 - v^t \mathcal{A}^{-1} u}, \tag{2.18}$$

cuando $\mathcal{A} = x^t x$ y $u = v = x_i^t$, donde x_i es el i-ésimo renglón de la matriz x, es decir, el vector cuyas entradas corresponden a la i-ésima observación.
Utilizando el hecho de que

$$x^t x = \sum_{i=1}^{n} x_i^t x_i \tag{2.19}$$

se obtiene

$$(x^t_{(i)} x_{(i)})^{-1} = (x^t x)^{-1} + \frac{(x^t x)^{-1} x_i^t x_i (x^t x)^{-1}}{1 - p_{ii}}, \tag{2.20}$$

Con esto se llega a que vease apéndice A.1

(2.21) $$DFBETA_i = \frac{(x^t x)^{-1} x_i^t e_i}{1 - \mathcal{P}_{ii}} .$$

Si se denota con $\hat{\beta}_j$ la j-ésima entrada del vector $\hat{\beta}$, y $\hat{\beta}_{(i)j}$ la j-ésima entrada del vector $\hat{\beta}_{(i)}$ con $\hat{\beta}_{(i)}$ el estimador de β cuando la i-ésima observación ha sido omitida, y $\mathcal{C}=(x^t x)^{-1}x^t$, entonces una forma de ver el cambio en una sóla entrada del vector $\hat{\beta}$, es viendo el valor

(2.22) $$\hat{\beta}_j - \hat{\beta}_{(i)j} = \frac{\mathcal{C}_{ji} e_i}{1 - \mathcal{P}_{ii}}$$

Sin embargo es necesario que este cambio se vea independientemente de las unidades con que se este trabajando es por ello que aquí se usará

(2.23) $$| DFBETAS_{(i)j} | = \left| \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{\hat{\sigma}_{(i)} \sqrt{(x^t x)^{-1}_{jj}}} \right|$$

(2.24) $$= \left| \frac{\mathcal{C}_{ji}}{\sqrt{(x^t x)^{-1}_{jj}}} \right| \left| \frac{e_i}{\hat{\sigma}_{(i)}(1-\mathcal{P}_{ii})} \right| .$$

como

(2.25) $$\sum_{i=1}^{n} (x^t x)^{-1} x_i^t x_i (x^t x)^{-1} = (x^t x)^{-1}$$

entonces

(2.26) $$| DFBETAS_{(i)j} | = \frac{| e_i |}{\hat{\sigma}_{(i)} \sqrt{1-\mathcal{P}_{ii}}} \frac{| \mathcal{C}_{ji} |}{\sqrt{\sum_{k=1}^{n} \mathcal{C}^2_{jk}} \sqrt{1-\mathcal{P}_{ii}}}$$

(2.27) $$= |t_i^*| \frac{| \mathcal{C}_{ji} |}{\sqrt{\sum_{k=1}^{n} \mathcal{C}^2_{jk}} \sqrt{1-\mathcal{P}_{ii}}}$$

(2.28)
$$\leq \frac{2\,|v_{ji}|}{\sqrt{\sum_{k=1}^{n} v_{jk}^2}\ \sqrt{1-p_{ii}}}\ .$$

Si cada v_{jk} es similar una con otra, es decir, si para cada $k\in(1,\ldots,n)$ $v_{jk}\approx v$. Como $\hat{\beta}_j=\sum_{k=1}^{n} v_{jk}y_k$, esto quiere decir que cada una de las v_{jk}, $k\in(1,\ldots,n)$ contribuyen de más o menos con el mismo peso para determinar el valor de $\hat{\beta}_j$, entonces

(2.29)
$$\frac{v_{ji}}{\sqrt{\sum_{k=1}^{n} v_{jk}^2}} \approx \frac{C}{\sqrt{n\,C^2}} = \frac{1}{\sqrt{n}}\ .$$

De este modo, si $p_{ii}\approx p/n$

(2.30)
$$\frac{2\,|v_{ji}|}{\sqrt{\sum_{k=1}^{n} v_{jk}^2}\ \sqrt{1-p_{ii}}} \approx \frac{2}{\sqrt{n}\ \sqrt{\frac{n-p}{n}}} = \frac{2}{\sqrt{n-p}}\ .$$

La desigualdad (2.28) es valida en un 95% de los casos para una n grande. Es por esto que las $DFBETAS_{(i)\,j}$ que rebasen a $2/\sqrt{n-p}$ deberán de tomarse en cuenta.

2.4.2 DFFITS.

Ahora se discute como detectar la influencia de la i-ésima observación, sobre el j-ésimo pronóstico. Por el momento se analiza el caso particular de como afecta la omisión de la i-ésima observación sobre el i-ésimo pronóstico.
La diferencia de los pronósticos estimados es

(2.31)
$$\hat{y}_i-\hat{y}_{(i)i} = x_i(\hat{\beta}-\hat{\beta}_{(i)})$$

Con $\hat{y}_{(i)i}$ el pronóstico de la i-ésima observación cuando la i-ésima observación ha sido omitida.

Además

$$\hat{y}_i - \hat{y}_{(i)i} = x_i \left[\frac{(x^t x)^{-1} x_i^t e_i}{1 - p_{ii}} \right] \tag{2.32}$$

$$= \frac{p_{ii} e_i}{1 - p_{ii}} \quad . \tag{2.33}$$

para anular el efecto de unidades distintas se divide (2.33) por la varianza $\hat{\sigma}_{(i)}\sqrt{p_{ii}}$ y se obtiene

$$DFFITS_i = \left[\frac{p_{ii}}{1-p_{ii}} \right]^{\frac{1}{2}} \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}} \tag{2.34}$$

Para el caso en que $p_{ii} \approx p/n$ para cada $i \in (1,\ldots,n)$

$$|DFFITS_i| \approx \left| \left[\frac{p/n}{1-p/n} \right]^{\frac{1}{2}} t_i^* \right| \leq 2 \left[\frac{p}{n-p} \right]^{\frac{1}{2}} . \tag{2.35}$$

y para valores grandes de n las observaciones para las cuales $|DFFITS_i| > 2\sqrt{p/(n-p)}$ se les puede considerar como influyentes.

La influencia de la i-ésima observación sobre el j-ésimo pronóstico esta acotada superiormente por |DFFITS|. Vease Belsley, Kuh, Welsch (1980) pág 15.

2.4.3. Lamda de Wilk.

Comparar el cambio de medias es otra forma de ver como afecta una o un grupo de observaciones al modelo obtenido.
Si

$$\mathcal{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p-1} & y_1 \\ \vdots & & & \\ x_{n1} & \cdots & x_{np-1} & y_n \end{bmatrix}$$

$$\bar{\mathcal{X}} = \begin{bmatrix} \bar{x}_1 & \cdots & \bar{x}_{p-1} & \bar{y} \end{bmatrix}$$

donde $\bar{x}_1 = \sum_{k=1}^{n} x_{k1}/n \quad \ldots \quad \bar{x}_{p-1} = \sum_{k=1}^{n} x_{kp-1}/n \quad \bar{y} = \sum_{k=1}^{n} y_k/n \quad \tilde{\mathcal{X}} = \mathcal{X} - \bar{\mathcal{X}}$

y $\tilde{\mathcal{Z}}_{(i)}$ la matriz $\tilde{\mathcal{Z}}$ sin el i-ésimo renglón

$$\Lambda_{(\tilde{\mathcal{Z}}_i)} = \frac{\det(\tilde{\mathcal{Z}}'_{(i)}\tilde{\mathcal{Z}}_{(i)} - (n-1)\bar{\tilde{\mathcal{Z}}}'_{(i)}\bar{\tilde{\mathcal{Z}}}_{(i)} - \tilde{\mathcal{Z}}'_i\tilde{\mathcal{Z}}_i)}{\det(\tilde{\mathcal{Z}}'\tilde{\mathcal{Z}})} . \tag{2.36}$$

$\bar{\tilde{\mathcal{Z}}}_{(i)}$ el vector de medias de $\tilde{\mathcal{Z}}_{(i)}$. Para el caso de $\tilde{\mathcal{X}}$, ver apéndice A.6.

$$\Lambda_{(\tilde{\mathcal{X}}_i)} = \left(\frac{n}{n-1}\right)(1-\mathcal{P}_{ii}) \tag{2.37}$$

Generalizando este resultado para $\mathcal{Z}$ da

$$\Lambda_{(\tilde{\mathcal{Z}}_i)} = \left(\frac{n}{n-1}\right)(1-\mathcal{P}_{ii})\left[\ 1 + \frac{t_i^{*2}}{n-p-1}\ \right]^{-1} \tag{2.38}$$

y $\Lambda_{(\tilde{\mathcal{Z}}_i)}$ se relaciona con la distribución F por

$$\left[\frac{n-p-1}{p}\right]\frac{1 - \Lambda_{(\tilde{\mathcal{Z}}_i)}}{\Lambda_{(\tilde{\mathcal{Z}}_i)}} \sim F_{(p,n-p-1)} . \tag{2.39}$$

Por medio de esta distribución es posible encontrar un cuantil determinado para el cual las observaciones puedan ser consideradas como influyentes.

Si $\mathcal{X}$ es una matriz con una columna de unos, supongase $\mathcal{X}_0=1$ como $\tilde{\mathcal{Z}}=\mathcal{Z}-\bar{\mathcal{Z}}$, $\tilde{\mathcal{Z}}$ tendrá una columna de ceros y esto implica que $\tilde{\mathcal{Z}}'\tilde{\mathcal{Z}}$ tenga una columna y un renglón de ceros, por lo que esto no hace correcto el análisis para cuando se tiene una matriz $\mathcal{X}$ con la columna de unos adjunta.

Esta es la razón por la cual en el siguiente capítulo en la sección de observaciones individuales no se considera esta medida.

2.4.4. Razón de varianzas (COVRATIC)

Esta medida trata sobre el cambio en la matriz de varianzas y covarianzas con, y sin una observación y esta dada por

$$(2.40) \qquad COVRATIO_i = \det\left[\frac{\hat{\sigma}^2_{(i)}\,(X^t_{(i)}X_{(i)})^{-1}}{\hat{\sigma}^2\,(X^tX)^{-1}}\right]$$

$$(2.41) \qquad = \left[\frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2}\right]^p \det\left[\frac{(X^t_{(i)}X_{(i)})^{-1}}{(X^tX)^{-1}}\right] .$$

Utilizando el hecho de que

$$(2.42) \qquad \det\left[\frac{(X^t_{(i)}X_{(i)})^{-1}}{(X^tX)^{-1}}\right] = \frac{1}{1-\mathcal{P}_{ii}}$$

y de que (Belsley, Kuh, Welsch, 1980), pág 14)

$$(2.43) \qquad (n-p-1)\hat{\sigma}^2_{(i)} = (n-p)\hat{\sigma}^2 - \frac{e_i^2}{1-\mathcal{P}_{ii}}$$

se obtiene

$$(2.44) \qquad COVRATIO_i = \frac{1}{\left[\frac{n-n-1}{n-p} + \frac{t_i^{*2}}{n-p}\right]^p (1-\mathcal{P}_{ii})}$$

Dos casos particulares son de tomarse en cuenta, de acuerdo a lo visto en la sección anterior

1) $|t_i^*| \geq 2$ y $\mathcal{P}_{ii}$ cercano a p/n, n grande

$$(2.45) \qquad COVRATIO_i = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{t_i^{*2}}{n-p}\right]^p (1-\mathcal{P}_{ii})}$$

$$(2.46) \qquad \approx \frac{1}{\left[1 + \frac{t_i^{*2}-1}{n-p}\right]^p}$$

(2.47) $$\leq \frac{1}{\left[1 + \frac{3}{n-p}\right]^p}$$

(2.48) $$\leq \frac{1}{\left[1 + \frac{3}{n}\right]^p}$$

(2.49) $$\approx \frac{1}{1 + 3p/n} \approx 1-3p/n$$

La aproximación de (2.48) a (2.49) se obtiene primero elevando (1+3/n) a la p y eliminando los demás terminos; y la siguiente es una aplicación del teorema de Taylor.

2) Para el caso $\mathcal{P}_{ii} \geq 2p/n$ y $t_i^* \cong 0$

(2.50) $$COVRATIO_i \cong \frac{1}{\left[\frac{n-p-1}{n-p}\right]^{p(1-\mathcal{P}_{ii})}}$$

(2.51) $$\geq \frac{1}{\left[1 - \frac{1}{n-p}\right]^{p(1-2p/n)}}$$

(2.52) $$\approx (1-3p/n)^{-1}$$

(2.53) $$\approx 1+3p/n .$$

De esta manera son de tomarse en cuenta aquellas observaciones para las cuales $COVRATIO_i$ sea menor a 1-3p/n o mayor a 1+3p/n.

Por supuesto este criterio tendrá mayor relevancia cuando haya más observaciones.

2.4.5 Razón de verosimilitudes.

La razón de verosimilitudes o distancia de verosimilitudes como también se le conoce, (ver Chaterjee & Hadi, 1982 pág 382) detecta la influencia sobre todo β y σ^2. Para éste caso

$$(2.54) \qquad \hat{\sigma}'^2 = \sum_{j=1}^{n} \frac{(y_j - x_j\hat{\beta})^2}{n} = \left(\frac{n-p}{n}\right)\hat{\sigma}^2$$

$$(2.55) \qquad \hat{\sigma}'^2_{(i)} = \sum_{j\neq i} \frac{(y_j - x_j\hat{\beta}_{(i)})^2}{n-1} = \left(\frac{n-p-1}{n-1}\right)\hat{\sigma}^2_{(i)}$$

y

$$L(\hat{\beta},\hat{\sigma}'^2) = \log\left[\prod_{j=1}^{n} (2\pi\hat{\sigma}'^2)e^{-\frac{(y_j - x_j\hat{\beta}_{(i)})^2}{2\hat{\sigma}'^2}}\right]$$

$$(2.56) \qquad = -\frac{n}{2}\log(2\pi\hat{\sigma}'^2) - \frac{n}{2}$$

$$L(\hat{\beta}_{(i)},\hat{\sigma}'^2_{(i)}) = \log\left[\prod_{j=1}^{n} (2\pi\hat{\sigma}'^2_{(i)})e^{-\frac{(y_j - x_j\hat{\beta}_{(i)})^2}{2\hat{\sigma}'^2_{(i)}}}\right]$$

$$(2.57) \qquad = -\frac{n}{2}\log(2\pi\hat{\sigma}'^2_{(i)}) - \frac{n-1}{2} - \frac{(y_i - x_i\hat{\beta}_{(i)})^2}{2\hat{\sigma}'^2_{(i)}}$$

por lo que (2.58)

$$RV_i = 2\left[L(\hat{\beta},\hat{\sigma}'^2) - L(\hat{\beta},\hat{\sigma}'^2_{(i)})\right] = n\log\left[\frac{\hat{\sigma}'^2_{(i)}}{\hat{\sigma}'^2}\right] - \frac{(y_i - x_i\hat{\beta}_{(i)})^2}{2\hat{\sigma}'^2_{(i)}} - 1$$

como t_i^* esta dada en función de $\hat{\sigma}^2_{(i)}$ sustituyendo y simplificando se obtiene

$$(2.59) \qquad = n\log\left[\left(\frac{n}{n-1}\right)\frac{n-p-1}{t_i^{*2}+n-p-1}\right] + \frac{t_i^{*2}(n-1)}{(1-\mathcal{P}_{ii})(n-p-1)} - 1$$

y está asintóticamente relacionado a una distribución χ^2 con p+1 grados de libertad. Por lo que las observaciones para las cuales RV_i sean mayores a un cuantil de una χ^2 con p+1 grados de libertad con un α dado, serán consideradas como influyentes.

2.5 Generalización a grupos de observaciones.

En esta sección se trabaja con medidas para la detección de puntos de influencia en grupo. D_m representa a un subconjunto de observaciones de tamaño m.
El tamaño de m debe ser menor a n-p, y a n/2.

2.5.1. Lamda de Wilk.

La Λ de Wilk se generaliza de la siguiente manera. Si ι es un vector de nx1 de unos y ι_1 es un vector de nx1 de unos en los renglones que corresponden a las observaciones contenidas en D_m y cero en otra parte $\iota_2=\iota-\iota_1$

$$\Lambda_{(D_m)} = \frac{\det[\tilde{\mathcal{Z}}'\tilde{\mathcal{Z}}-(1/m)\tilde{\mathcal{Z}}'\iota_1\iota_1'\tilde{\mathcal{Z}} - (1/(n-m))\tilde{\mathcal{Z}}'\iota_2\iota_2'\tilde{\mathcal{Z}}\ ]}{|\tilde{\mathcal{Z}}'\tilde{\mathcal{Z}}|} \tag{2.60}$$

que se simplifica a

$$\Lambda_{(D_m)} = 1 - \frac{n}{m(n-m)}\,(\iota_1'\tilde{\mathcal{P}}\,\iota_1) \tag{2.61}$$

donde $\tilde{\mathcal{P}}=\tilde{\mathcal{Z}}(\tilde{\mathcal{Z}}'\tilde{\mathcal{Z}})\tilde{\mathcal{Z}}'$. Si aquí se toma el supuesto de que los renglones de $\tilde{\mathcal{Z}}$ provienen de una muestra independiente que se distribuye de manera normal (Beleley, Kuh, Welech, 1982 pág. 37)

$$\left[\frac{n-p-1}{p}\right]\left[\frac{1-\Lambda_{(D_m)}}{\Lambda_{(D_m)}}\right] \sim F_{\alpha(p,n-p-1)} \tag{2.62}$$

esta medida generaliza de una manera sencilla la detección de puntos influyentes de una sóla observación a un grupo, y nuevamente por su distribución es posible encontrar un criterio para la detección de grupos influyentes con un α dado.
Previo a esta medida se experimentó con la razón de verosimilitudes generalizada, y con la T que surge de aumentar variables dummy a

las observaciones que se consideraron como influyentes; los resultados con ambas no fueron satisfactorios en el sentido de que estas medidas son poco sensibles. En cambio, ésta última resultó ser más sensible y es la que se utilizará más adelante.

2.5.3. RESUMEN.

Las medidas que aquí se trataron miden diferentes aspectos de influencia. Su eficiencia depende de los supuestos de que se parte; por lo cual sería incorrecto utilizar estas medidas sin antes constatar que los supuestos realmente se cumplan.
Aunque los outliers y los puntos de alta palanca no son medidas de influencia, resulta conveniente tenerlos identificados.

Outliers.

$$t_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-\mathcal{P}_{ii}}}$$

Punto de Corte $|t_i^*| > t_\alpha(n-p-1)$

Supuestos de Normalidad.

Puntos de Alta Palanca

$\mathcal{P}_{ii} = x_i (x^t x)^{-1} x_i^t$

Punto de corte

$$\mathcal{P}_{ii} < \frac{\left(\frac{p-1}{n-p}\right) F^{\alpha}_{(p-1,n-p)} + \frac{1}{n}}{1 + \left(\frac{p-1}{n-p}\right) F^{\alpha}_{(p-1,n-p)}}$$

Supuestos de Normalidad sobre las Observaciones $\mathcal{X}_i$

DFBETAS

$$|DFBETAS_{(i)}| = \frac{|\hat{\beta}_j - \hat{\beta}_{(i)j}|}{\hat{\sigma}_{(i)}\sqrt{(x^t x)^{-1}_{jj}}}$$

Punto de corte

$$\frac{2}{\sqrt{n-p}}$$

Supuestos de Normalidad en los errores.

DFFITS

$$|DFFITS_i| = \left|\frac{\hat{y}_i - \hat{y}_{(i)i}}{\hat{\sigma}_{(i)}\sqrt{p_{ii}}}\right|$$

Punto de corte

$$|DFFITS_i| < 2\sqrt{\frac{p}{n-p}}$$

Supuestos de Normalidad.

Lamda de Wilk

$$\Lambda_{(\tilde{z}_i)} = \left[\frac{n}{n-1}\right](1-p_{ii})\left[1 + \frac{t_i^{*2}}{n-p-1}\right]^{-1}$$

o para un grupo

$$\Lambda_{(D_m)} = 1 - \frac{n}{m(n-m)}(\iota_i^t \tilde{p}\, \iota_i)$$

Punto de corte, para cualquiera de las dos Λ

$$F^{\alpha}_{(p,n-p-1)} \quad \text{para} \quad \left[\frac{n-p-1}{p}\right]\frac{1-\Lambda}{\Lambda}$$

Supuestos de Normalidad sobre las Observaciones z_i o z_{D_m}.

COVRATIO

$$COVRATIO_i = \frac{1}{\left[\frac{n-p-1}{n-p} + \frac{t_i^{*2}}{n-p}\right]^p (1-\mathcal{P}_{ii})}$$

Los puntos de corte $COVRATIO_i \in [1-3p/n, 1+3p/n]$

Supuestos de Normalidad.

Razón de Verosimilitudes

$$RV_i = 2[L(\hat{\beta},\hat{\sigma}) - L(\hat{\beta}_{(i)},\hat{\sigma}_{(i)})]$$

Punto de corte, para cualquiera de las RV $\geq \chi^2_{\alpha,p+1}$

Supuestos de Normalidad.

2.6 Referencia.

- Belsley, D.A. Kuh, E. and Welsch, R.E. (1980) Regression Diagnostics identifying Influential Data and sources of collinearity. Wiley, New York.

- Mood Alexander M., Graybill Franklin A., Boes Duane C. (1974) Introduction to the theory of statistics third Edition, McGraw-Hill International Editions.

- Rao, C.R. (1973), Linear statical inference and its Applications. 2nd ed. John Wiley and Sons, New York.

- Saprit Chaterjee and Ali S. Hadi. (1982) Influential Observations High Leverage Points, and Outliers in Linear Regression. Statical Sience Vol 1, No 3, p 379-416.

- Dennie R. Cook and Sanford Weisberg. (1982) Residuals and Influence in Regression first Edition, Chapman and Hall. New York p 182-186

CAPITULO 3

PROGRAMACION DE MEDIDAS DE INFLUENCIA.

3.1 Introducción al capítulo.

Hasta la sección 3.4 se muestran las funciones utilizadas para el cálculo de las medidas de influencia y de la 3.6 en adelante el uso del programa.
Los fundamentos de los algoritmos respectivos no se muestran por que no es el objetivo principal del trabajo, sin embargo se hace referencia a la bibliografía utilizada.
Se eligió desarrollar el programa en el lenguaje "C" por generar código muy eficiente y existir compiladores de éste lenguaje en casi todos los sistemas operativos, además de la facilidad que éste dá para el manejo de archivos.
Todas las matrices que aquí se usan estan definidas como arreglos lineales para darle flexibilidad a los valores de "n" y de "p" (n es el número de observaciones y "p" el número de regresores incluyendo al 1 que se le adjunta). De tal manera que si se quiere hacer referencia a la entrada x_{ij} de la matriz X que tiene tamaño de nxp, se hará referencia como X[i*p+j] o por el manejo de apuntadores que aquí se hace *(X+i*p+j), es decir recorre i*p+j elementos para ubicarse en la posición correcta.
A continuación se muestran las funciones básicas utilizadas para las operaciones matriciales. En el disco que se anexa se encuentran los programas fuente para mayor detalle.

```
#define NO  0
#define SI  1
/* Obtiene la matriz transpuesta de ap de n*m y la deja en tr*/
void transp(double *ap,int n,int m,double *tr)
    { int i,j,k=n*m;
      for(j=0;j<k;j++)
       *(tr+j)=0.0;

       for(j=0;j<m;j++)
        for(i=0;i<n;i++)
         *(tr+j*n+i)=*(ap+i*m+j);
    };

/* Producto de matrices C=A*B A es de nxk y B de kxm  */
void producto(double *A,double *B,int n, int k, int m,double *C)
  { int i,j,l,k;

    for(i=0;i<k;i++)
      *(C+i) = 0.0;
```

```
    for(i=0;i<n;i++)
     for(j=0;j<m;j++)
      for(l=0;l<k;l++)
       *(C+i*m+j) += (*(A+i*k+l))*(*(B+l*m+j));

  };

/* Busca en los renglones mayores o iguales a I de la matriz A     */
/* el I-ésimo diferente de cero y regresa su indice                */
/* si no lo encuentra regresa -1                                   */
int no_cero(double *A,int N, int I)
  {
    int j,FIN=-1;
      for(j=I;(j<N) && (FIN==-1);j++)
  if (fabs(*(A+j*N+I)) > CERO) FIN=j;
      return FIN;
   };

/* Intercambia renglones  */
void intercambia(double *A,int N,int reng_a,int reng_b)
  { int i;
    double B;

     for(i=0;i<N;i++)
      { B = *(A+reng_a*N+i);
        *(A+reng_a*N+i) = *(A+reng_b*N+i);
        *(A+reng_b*N+i) = B;

      };

  };

/* Regresa a la matriz identidad de tamaño N*N */
void ident(int N,double *identidad)
 { int i,j;
   for (i=0;i<N;i++)
    for(j=0;j<N;j++)
     *(identidad+i*N+j)=((i==j)?(1.0):(0.0));
 };

/* Regresa la matriz inversa de A */
int inv(double *A,int N,double *inversa)
  { int I=0,j,k,l,bien=1;
    double IND;
    ident(N,inversa);

    while ((I<N) && (bien !=-1))
     {if ((bien=(j=no_cero(A,N,I)))!=-1)
 {
   if (j!=I)
     {  intercambia(A,N,I,j);
        intercambia(inversa,N,I,j);
      };

  IND =*(A+I*N+I);

 for(k=0; k<N; k++)
  { *(A+I*N+k) /= IND;
    *(inversa+I*N+k) /= IND;
```

```
 );

for (k=0; k<N; k++)
 if ((k!=I) &&   ((IND = *(A+k*N+I))  != 0.0))
  for(l=0;l<N; l++)
   { *(A+k*N+l)  -= IND*(*(A+I*N+l));
     *(inversa+k*N+l) -= IND*(*(inversa+I*N+l));
   };
};
      I += 1;
     };
     return bien;
    };

/* Hace el producto de la matriz A nxp por B pxn, pero obtiene
   solamente la diagonal del resultado y lo pone en C nx1    */
   void prod_diag(double *A,double *B,int n,int p,double *C)
{ int i,l;
    for(i=0;i<n;i++)
       *(C+i)=0.0;
    for(i=0;i<n;i++)
     for(l=0;l<p;l++)
      *(C+i) += (*(A+i*p+l))*(*(B+l*n+i));
 };
```

3.2 CÁLCULO NUMÉRICO DE $\hat{\beta}$.

El siguiente algoritmo para determinar las entradas del vector $\hat{\beta}$ se basa en el método de la factorización de Cholesky (Golub 1969, pág 142)

1) Encontrar a $\mathcal{G}$ (de pxp) tal que $\mathcal{X}^t\mathcal{X} = \mathcal{G}\mathcal{G}^t$ ($\mathcal{G}$ se obtiene por medio del método de la factorización de Cholesky)

2) Hallar $\mathcal{D}$ (de px1) tal que $\mathcal{G}\mathcal{D} = \mathcal{X}^t\mathcal{Y}$.

3) Hallar $\hat{\beta}$ (de px1) tal que $\mathcal{G}^t\hat{\beta}=\mathcal{D}$

Las funciones que se utilizan aparte de las vistas en 3.1 se muestran a continuación.

```
/* Sistema triangular inferior para hallar a B */
 void despinf(double *A,double *C, int n, double *B)
  { double suma;
    int i,j;
   *B= *C/(*A);
    for (i=1;i<n; i++)
     { suma = 0.0;
   for(j=0; j<i ; j++)
     suma += (*(A+i*n+j))*(*(B+j));
   *(B+i) = (*(C+i)-suma)/(*(A+i*n+i));
```

```
        );
    );

   /* Sistema triangular superior para hallar a B */
   void despsup(double *A,double *C, int n, double *B)
    { double suma;
      int i,j;
     *(B+n-1)= *(C+n-1)/(*(A+n*(n-1)+n-1));
      for (i=n-2;i>=0; i--)
       { suma = 0.0;
     for(j=i+1; j<n ; j++)
       suma += (*(A+i*n+j))*(*(B+j));

      *(B+i) = (*(C+i)-suma)/(*(A+i*n+i));
       };
    };

  /* Pone en G la factorización de Cholesky (Hacer antes  G = XtX) */
    void  Cholesky(double  *G, int p)
     { double suma;
       int i,j,k;

       for(k=0; k<p; k++)
{ i=0, suma=0.0;
  while (i<k)
   { suma += (*(G+k*p+i))*(*(G+k*p+i));
       i += 1;
   }
  *(G+k*p+k) -= suma;
  *(G+k*p+k)  = sqrt(*(G+k*p+k));

  for(i=k+1; i<p ; i++)
   { j=0,suma=0.0;
      while (j<k)
       { suma += (*(G+i*p+j))*(*(G+k*p+j));
     j += 1;
       };

      *(G+i*p+k) -= suma;
      *(G+i*p+k) /= *(G+k*p+k);
   };

      );

      for(i=0; i<p-1; i++)
for(j=i+1; j<p ; j++)
 *(G+i*p+j)= 0.0;

    );

/* Obten a BETA por el método de Cholesky */
void obten_beta(double *Y,double *X,int n,int p,double *BETA)
 { double     G[MAX_P*MAX_P],Gt[MAX_P*MAX_P],
          XtX[MAX_P*MAX_P],D[MAX_P],R[MAX_P],Xt[MAX_P*MAX_N];
   int i,j;

   transp(X,n,p,Xt);
   producto(Xt,X,p,n,p,XtX);
```

```
    for(i=0;i<n;i++)
     for(j=0;j<p;j++)
      *(G+i*p+j)=*(XtX+i*p+j);

   Cholesky(G,p);
   tranep(G,p,p,Gt);
   producto(Xt,Y,p,n,1,D);
   despinf(G,D,p,R);
   despsup(Gt,R,p,BETA);
);
```

Para asegurar que $\hat{\beta}$ exieta se calcula antes la matriz inversa de $x^t x$ y en caso de no existir manda un mensaje de error.
Además, $(x^t x)^{-1}$ se utiliza para el cálculo de las entradas de la matriz diagonal $\mathcal{P}$.

3.3 Medidas de probabilidad.

Los algoritmos para el cálculo de medidas de probabilidad, que a continuación se muestran se basan primero en el cálculo de sus funciones de distribución respectivas y después en obtener los puntos en los cuales acumulan la confianza requerida.

```
/* Obtiene la distribución Normal */
double Normal(double x)
 {
   int i,n,f=4,sig=0,m=10;
   double S=1,b=fabs(x),com;

   if (x>0) sig=1;
   if (x<0) sig=-1;

    com = 480.0*sqrt(2*PI);
    com = pow(b,5)*pow(10,(double) m)/com;
    com = sqrt(com);
    com = sqrt(com);

    n = 1 + floor(com);

    if (n<4) n=4;

    for(i=1;i<=2*n;i++)
     { com = (i*b)/n;
       com = 0.0 - (pow(com,2)/8);
       com = exp(com);

S += f*com;
```

```
      if (i==(2*n-1)) f=1;
       else if (f==4) f=2;
      else       f=4;
      }

        com = 6*n*sqrt(2*PI);
        com = sig*8*b/com;

        return (0.5+com);
    };

/* Regresa a la  distribución F8nedekor */
double F8nedekor(int n,int m,double x)
 { int a,b,i,j;
   double w,y,z,d,p;

     a = 2*(m/2)-m+2;
     b = 2*(n/2)-n+2;
     w = x*(((double) m)/((double) n));
     z = 1/(1+w);

     if (a==1)
      { if (b==1)
{ p = sqrt(w);
  y = 1/PI;
  d = y*(z/p);
  p = 2*y*atan(p);
 }
else
{ p = sqrt(w*z);
  d = 0.5*p*'z/w);
};
      }
    else
     if (b == 1)
      { p  = sqrt(z);
d  = 0.5*z*p;
p -= 1;
      }
     else
      { d = z*z;
p = w*z;
      };

     y = 2*(w/z);

     for(j=b+2;j<=n;)
      { d  = (1+((double)a)/((double) (j-2)))*d*z;
p  = ((a==1)?(p+d*(y/((double) (j-1)))):((p+w)*z));
j += 2;
      };

      y = w*z;
      z = 2/z;
      b = n-2;

      for(i=a+2;i<=m;)
       { j  = i+b;
```

```
d  = y*d*((double) j)/((double) (i-2));
p -= z*d/((double) j);
i += 2;
     );

   return p;

 );

/* Distribución T de Student */
double Tstudent(double n,double t)
 ( double a,b,y,z=1.0,tstudent=-1;
   int j;
   t *= t; y = t/n; b =1.0+y;
   if ((n>=20) && (t<n) | (n>200))      /* Suma anidada de    */
     ( if (y>CERO) y = log(b);          /* series de coseno */
       a = n-0.5,b=48.0*(a*a),y=a*y;
       y=(((((-0.4*y-3.3)*y-24.0)*y-85.5)/
(0.8*y*y+100.0+b)+y+3.0)/b+1.0)*sqrt(y);
       tstudent=2.0*(1.0-Normal(y));
     }
   else
   if ((n<20) && (t<4.0))
   ( a=(y=sqrt(y)); if (n==1) a=0.0;
     loop:  n -= 2;
     if (n>1)
       ( a=((n-1)/(b*n))*a+y; goto loop; );
     a = (   (n==0)?(a/sqrt(b)):((atan(y)+a/b)*(2/PI))    );
     tstudent = z-a;
   }
   else ( a = sqrt(b); y = a*n; j=0;
   for(j=j+2;a != z;)
    ( z=a; y *= (j-1)/(b*j); a += y/(n+j); );
      n += 2; z = (y=0.0); a = 0-a; goto  loop;
    }
  return (1.0 -tstudent/2.0);
 );

/* Distribución Chi Cuadrada */
double Chicuadrada(int n,double x)
 ( double a=0.5*x,y,s,chi=0.0,s,c,z;
       int par,grande;

       if (n%2==0)     par = SI; else     par = NO;
       if (x>10)    grande = SI; else grande = NO;

    if ((par==SI) || (n>2) && (grande == NO)) y=exp(-a);
    s=((par==SI)?(y):(2.0*(1.0-Normal(sqrt(x)))));

    if (n>2)
    ( x=0.5*((double)n-1.0);
      z=((par==SI)?(1.0):(0.5));
      if (grande==SI)
       ( s=((par==SI)?(0.0):(log(sqrt(PI)))));
  c = log(a);
  for(z=z;z<=x;z++)
   ( s = log(z)+s;
     s = exp(c*z-a-s)+s;
```

```
  };
chi = s;
}
      else
{ s=((par==SI)?(1.0):(1.0/(sqrt(a)*sqrt(PI))));
  c=0;
  for(z=z;z<=x;z++)
  { s = s*a/z;
    c = c+s;
  };
 chi=c*y+s;
};
   }
   else chi=s;
  return  (1-chi);
 };

/* Todas las distribuciones. Aquí n y m son los grados de
   libertad, x es la variable de la función y funcion es la
   Función que va a utilizar.                                  */
double distribucion(int n,int m,double x,int funcion)
 { double dist = -1;
   switch(funcion){
       case 0: dist = Normal(x);
        break;
       case 1: dist = FSnedekor(n,m,x);
        break;
       case 2: dist = Tstudent(n,x);
        break;
       case 3: dist = Chicuadrada(n,x);
        break;
    };
    return dist;
  };

/* Halla cuantiles por medio del método de secantes */
double cuantil(int n,int m, double alfa, int funcion)
 { double x1=0.05,x2=0.50,y1,y2,z;

  if ((funcion==1) || (funcion==3) || (n>10)) x2=2.0;

   y1 = distribucion(n,m,x1,funcion)-alfa;
   y2 = distribucion(n,m,x2,funcion)-alfa;

  while (fabs(x1-x2)>CERO)
   {  z = x2-((x1-x2)/(y1-y2))*y2;
     x1 = x2, y1 = y2;

     if (funcion==3)
if (z<0) z=min(fabs(x2-0.05),0.05);
else if (z>54.0) z=54.0;

     y2 = distribucion(n,m,x2=z,funcion)-alfa;
   };
  return x1;
 };
```

Para mayor información al respecto vease Selected Algorithms from

the ACC. vol I y II. (1980).

3.4 Manejo de memoria.

Para el manejo de datos todo se lee de un archivo, lo guarda en memoria, y debido a que para hacer esto es necesario tener a las variables y a los arreglos matriciales previamente definidos, la memoria de la computadora reserva un espacio para ellos, de modo que si se utilizan pocos o muchos datos, el espacio reservado es el mismo.

A causa del tamaño ocupado por cada arreglo en cada función del programa, éste no permite tamaños de arreglos muy grandes, por lo que se tienen las limitantes de que el valor máximo que permite para estos es de 1000 (todos los arreglos de números y las variables estan declaradas como dobles, ocupando cada uno de estos 8 bytes)

Si n es el número de observaciones y p es el número de regresores, las únicas limitantes deseadas son que si k es el tamaño máximo del arreglo

$$nxp \leq k \qquad \text{(pues } \mathcal{X} \text{ es de nxp)}$$

(3.1) $$p \leq n \qquad nxp > 0$$

entonces $$p^2 \leq k$$

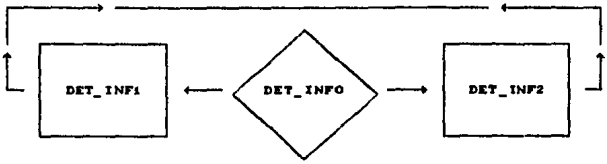
(3.2) y $$p \leq \sqrt{k}$$

Por lo que el valor máximo de p es el máximo entero menor a $\sqrt{k}$ y el mínimo será 2 (pues siempre se considerará el modelo con pendiente al origen β_0), entonces si p vale 2 y $2n \leq k$ el valor máximo de n será de k/2.

Entonces los arreglos que tengan relación o dependencia con p tomarán el supuesto del valor máximo que en este caso es 31 y para n el de 500.

Debido también al tamaño y al espacio que el programa ocupa en memoria fue necesario dividirlo en tres partes DET_INF0, DET_INF1, y DET_INF2. DET_INF0 se encarga del manejo de menúe y manipulación de archivos; DET_INF1 de la detección individual y DET_INF2 de la detección en grupos, en sus dos formas, y hay dependencia entre los tres.

El diagrama de flujo entre ellos es el siguiente:



3.5 Estructura del archivo de datos.

Para poder leer un archivo de datos, éste necesita estar en el mismo directorio donde se encuentren los programas ejecutables (DET_INFO.EXE, DET_INF1.EXE, DET_INF2.EXE) y tener la extensión .DAT ya que sólo con esta podrá leer.
Debe estar en código ascii y de la siguiente forma

---

n p
$y_1\ x_{11}\ x_{12}\ \ldots x_{1p-1}$
$y_2\ x_{21}\ x_{22}\ \ldots x_{2p-1}$
.
.
.
$y_n\ x_{n1}\ x_{n2}\ \ldots x_{np-1}$

---

n es el número de observaciones y p es el número de variables independientes incluyendo a la columna de unos.
En el primer renglón deben estar los valores enteros n y p, pues a partir de la siguiente línea leerá a lo más n renglones y por cada renglón leerá a lo más p números.
Es importante que n y p estén en el primer renglón del archivo, pues de no ser así considerará a sus valores como cero y no tomará ningún dato.
Leyendo de esta manera pondrá los datos en una matriz y y en una matriz x de la siguiente forma:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad x = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots x_{1p-1} \\ 1 & x_{21} & x_{22} & \ldots x_{2p-1} \\ . & & & \\ . & & & \\ . & & & \\ 1 & x_{n1} & x_{n2} & \ldots x_{np-1} \end{bmatrix}$$

EJEMPLO 3.1

Supóngase al archivo ejemp.dat de la siguiente forma

```
7 3
1 4 6 7 8
3 4 6
2 43 5

2.3 2 1
3 4.5 -2.6
8 8.4 8.25 3.73
3 4 56 -56.67 67
```

lo convertira en

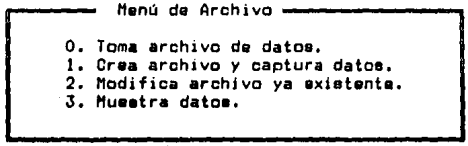
$$y = \begin{bmatrix} 1.0 \\ 3.0 \\ 2.0 \\ 0.0 \\ 2.3 \\ 3.0 \\ 8.0 \end{bmatrix} \qquad x = \begin{bmatrix} 1.0 & 4.00 & 6.00 \\ 1.0 & 4.00 & 6.00 \\ 1.0 & 43.00 & 5.00 \\ 1.0 & 0.00 & 0.00 \\ 1.0 & 2.00 & 1.00 \\ 1.0 & 4.50 & -2.60 \\ 1.0 & 8.40 & 8.25 \end{bmatrix}$$

Es importante hacer notar que en ausencia de números, cuando ya se acaba la linea, llena a los elementos restantes de la entrada con ceros. Observese la quinta linea del archivo con el cuarto renglón de la matriz y y de la matriz x. Aunque el archivo tenga más renglones sólo lee n.

Este tipo de archivo puede ser creado en cualquier editor de texto, siempre y cuando esté en código ascii y además de eso es posible crearlo, modificarlo y verlo en el mismo programa como se muestra adelante.

3.6 Menú de archivo.

En esta sección se pueden hacer cuatro cosas: Opción 1.- Seleccionar un archivo de datos ya creado previamente, para detectar la influencia sobre los datos que éste contenga. Opción 2.- Crear un archivo y capturar los datos manualmente. Opción 3.- Modificar un archivo que ya exista (ya sea que se haya creado en esta sección o que se tenga para su uso) tanto en los valores de n y p como de sus entradas y Opción 4 Mostrar los datos de un archivo sin poder modificarlo.
Este es el desplegado para esta sección en el programa



Para seleccionar la opción deseada sólo hay que ubicarse en ella iluminandola moviendose con las flechas arriba (↑) o abajo (↓) y oprimiendo enseguida intro (↵) u oprimiendo el primer número de la opción deseada; el valor por omisión es cero.

3.6.1 Toma archivo de datos.

Aquí sólo aparecen los archivos con extención .DAT (ejemplos MOOR.DAT, LLUVIA.DAT, etc.) siempre y cuando la cantidad que exista de estos no exceda a 96. En caso de rebasar esta cantidad es necesario mover o quitar del directorio a aquellos que por el momento no se vayan a ocupar.
De modo similar para seleccionar el archivo deseado hay que ubicarse en él, iluminándolo y luego oprimir intro (↵).
Oprimiendo la tecla de escape <esc> regresa al menú de archivo.
Después de esta selección se pasa al menú de influencia con el archivo seleccionado para su análisis.

3.6.2 CREA ARCHIVO Y CAPTURA DATOS.

Para identificar los archivos hechos en esta sección el nombre será siempre de tamaño 8 y terminará con 00.DAT (aquí sólo admite letras). Enseguida pide el valor de n y de p, para regresar en esta sección se con escape <esc>. Despuée de esto aparece la zona de captura (de acuerdo con n y p) para los números que se quieran usar.

Para moverse de una celda a otra es posible hacerlo con las flechas (← → ↑ ↓) con las teclas de inicio y fin, pág sig, pág ant, ctrl+tab y cambio+tab.

Para salir de esta sección se con escape <esc> enseguida preguntan si se guardan o no las modificaciones, en caso de que se le responda NO, no se graba ni un dato del archivo por lo que éste no se crea. Despuée de esto pide la confirmación de salida de esta sección, si se le dice que NO: regresa a la zona de captura y si la respuesta es afirmativa regresa al menú de influencia.

3.6.3 MODIFICA ARCHIVO YA CREADO.

Aquí toma el archivo de la misma manera que en la sección 3.6.1, pide que se le confirmen o modifiquen los valores de n y p y entra bajo las mismas condiciones de captura que en la sección 3.6.2. Aquí si no se le guardan las modificaciones, el archivo permanece igual que en su última versión.

Si se le modifican los valores de n y/o p retoma los datos anteriores hasta donde existan o hasta donde n y p lo permitan.

EJEMPLO 3.2

Si el archivo MUESTRA.DAT contiene:

```
5 2
1 3
4 6
7 2
5 1
3 6
```

Aquí n=5 y p=2.

Si ahora n=4 y p=3 el archivo quedará como:

```
4 3
1 3 0
4 6 0
7 2 0
5 1 0
```

3.6.4 MUESTRA DATOS.

Aquí toma el archivo al igual que en las secciones 3.6.1 y 3.6.3 y muestra los datos existentes (aquí no es posible modificarlo). Sin embargo es más fácil moverse para revisar los datos, las teclas que se pueden usar son las flechas (← → ↑ ↓) las teclas de inicio y fin, pág sig, pág ant, ctrl+tab y cambio+tab.

3.7 MENÚ DE INFLUENCIA.

El desplegado para esta sección es el siguiente

```
Menú de Influencia

0. Observaciones individuales.
1. Hasta un grupo de tamaño fijo.
2. Grupos especificos.
```

Para seleccionar cualquiera de estas opciones es igual que en la sección 3.6, con escape <esc> regresa al menú de archivo.

3.7.1 OBSERVACIONES INDIVIDUALES.

Aquí pide el nombre del archivo donde se pondrán los resultados. Si el archivo ya existe manda el mensaje Archivo ya existente ¿Continua S/N ? si se le contesta afirmativamente borrará completamente el archivo anterior sustituyendolo por los nuevos resultados del cálculo, en caso contrario volverá a pedir el nombre. En seguida pide el nivel de confianza va desde 1 hasta 99 porciento.

Después de darle estos datos procede a crear el archivo, de no ser posible manda un mensaje de error y de manera opcional regresa al menú de influencia.

A fín de constatar el cumplimiento de los supuestos para llevar acabo la factorización de Cholesky se calcula $(x^t x)^{-1}$ de no existir esta matriz da un mensaje de error y opcionalmente regresa al menú de influencia.

Al terminar pregunta si se desea realizar otra detección de influencia, si se le responde afirmativamente regresa al menú de influencia, de no ser así termina el programa.

De esta manera los resultados quedan grabados en el archivo de salida el cual siempre tendrá por extención .RES

Los resultados se mandan a un archivo y no a la pantalla por que la visualización de estos no es tan cómoda y/o tan agradable como el hecho de tenerlos por escrito, sobre todo cuando se trata de muchos datos.

La salida, al igual que el archivo de entrada está en código ascii.

En la salida aparece el archivo de donde fueron tomados los datos, el número de observaciones y el número de regresores. Los puntos de corte para los cuales se consideraron como observaciones influyentes al nivel de confianza dado, aunque hay puntos de corte que no dependen del nivel de confianza como $DFBETAS_{(i)j}$ $DFFITS_i$ $COVRATIO_i$ (Vease el capítulo anterior) por lo cual no deberá confundirse el hecho que con diferentes niveles de confianza estos puntos no cambien su valor.

En seguida se muestra el valor de $\hat{\sigma}^2$ (sigma^2) y las entradas del vector $\hat{\beta}$.

Sólo aquellas observaciones que se hayan detectado como influyentes por cualquiera de las medidas anteriores, será mostrado y marcado con un (*) indicando con esto cual medida lo detectó. Además también de mostrar a aquellas que detectó como outliers y puntos de alta palanca.

Al final pone el nombre del archivo de resultados la fecha y la hora según el calendario y reloj que tenga el sistema.

3.7.2 Hasta un grupo de tamaño fijo.

En esta opción como en la siguiente se consideró la condición de que para realizar las medidas de detección, el cuadrado del número de observaciones debe ser menor a 1000 (el tamaño máximo de un arreglo) por lo que mandará un mensaje de error si esta condición no se cumple. Esto se debe a que se necesita calcular una matriz de nxn.
Aquí pide el nombre del archivo, el nivel de confianza y el tamaño máximo del grupo. Así pues, si k es el tamaño máximo del grupo revisará la influencia de todas las combinaciones de grupos de tamaño 1,...,k tomando las n observaciones y dará como resultado sólo a aquellas que haya detectado como influyentes.
En el archivo de resultados escribe el nombre del archivo de datos, el número de observaciones y el número de regresores. Para cada grupo de regresores que haya detectado como influyente registra
$(\beta(X)_{(i_1,i_2,\ldots,i_f)}$ siendo $i_1,i_2,\ldots,i_f$ los índices del cual se compone el grupo, y f el tamaño de éste; pone además su medida de influencia y el punto de corte. Al final pone el nombre del archivo de resultados la fecha y la hora.

3.7.3 Grupos específicos.

Aquí pide el nombre del archivo de salida (con las mismas características que en las secciones anteriores), el nivel de confianza, el tamaño máximo del grupo y el número máximo de grupos que se va a revisar. Enseguida pasa a la sección de captura (con las mismas ventajas que en la sección 3.8, de modificar archivo) pero con la diferencia de que sólo captura enteros mayores o iguales a cero y menores o iguales a n.
Cuando se capture cero quiere decir que no existe índice alguno en ese lugar.
A diferencia de la sección anterior aquí recibe a un conjunto de grupos, con los cuales determinará si son o no influyentes, al

igual que en las secciones anteriores sólo pondrá a aquellos que haya detectado como influyentes.

EJEMPLO 3.3

Si n=12, el tamaño máximo del grupo es 4 y el número de grupos m=4, se tiene

---

1 3 7 5
6 12 4 0
8 6 6 5
3 2 0 0

Realizará el análisis de la influencia sobre los grupos
(1,3,7,5); (6,12,4); (8,6,5); (3,2)
y nuevamente mostrará sólo los que detecte como influyentes.

EJEMPLO 3.4

Los datos que a continuación se presentan son muestras tomadas en la cuenca de México, para probar la relación del deuterio con el oxígeno 18.

MUESTRAS CON O-18 Y DEUT. DENTRO DE LA CUENCA DE MEXICO DE CORTES Y MORALES, PERO QUITANDO MUESTRAS QUE NO TENGAN LOS DOS TIPOS O SEA, O-18 Y DEUT.

REF.	Lugar	Muestra	O-18	DEUT
1	E.Tlamacas	154	-11.4	-72
2	Ex-Convento	Ex-C	-11.1	-71
2	Caset-Derr.	C-D	-6.2	-31
2	Ex-Convento	Ex-C	-3.4	-9
2	Ex-Convento	Ex-C	-3.5	-15
2	Ex-Convento	Ex-C	-13	-96
2	Ex-Convento	Ex-C	-10	-74
2	Ex-Convento	Ex-C	-5.3	-38
2	Ex-Convento	Ex-C	-10.3	-77
2	Ex-Convento	Ex-C	-10.8	-82
2	Ex-Convento	Ex-C	-3.1	-9
2	Ex-Convento	Ex-C	-20.9	-153
2	Ex-Convento	Ex-C	-17.9	-114
2	Ex-Convento	Ex-C	-6.4	-40
2	Ex-Convento	Ex-C	-15.4	-116
2	Ex-Convento	Ex-C	-9	-63
2	Ex-Convento	Ex-C	-15.1	-109
2	Ex-Convento	Ex-C	-18.7	-137
2	Ex-Convento	Ex-C	-11	-75

2	Ex-Convento	Ex-C	-11	-75
2	Dee.d.Leon.	D.L.	-10.8	-74
2	Ex-Convento	Ex-C	-10.5	-76
2	Ex-Convento	Ex-C	-16.6	-121
2	Ex-Convento	Ex-C	-10.7	-84
2	Ex-Convento	Ex-C	-4.4	-19
3	Tlamacas	TLA	-12.1	-86
3	Monte Alegre	MAL	-15.6	-120
3	Col.Roma Sur	CCM	-8	-57
3	Tecómitl 19	19	-6.5	-38
3	Inst.Geofis.UNAM	IGF	-2.8	-11
3	Inst.Geofis.UNAM	IGF	-3.3	-17
3	Inst.Geofis.UNAM	IGF	-13.2	-97
3	Inst.Geofis.UNAM	IGF	-5.6	-14
3	Inst.Geofis.UNAM	IGF	-12.1	-81
3	Inst.Geofis.UNAM	IGF	-12.8	-84
3	Inst.Geofis.UNAM	IGF	-15.8	-122
3	Inst.Geofis.UNAM	IGF	-11.8	-80
3	Inst.Geofis.UNAM	IGF	-5.8	-34
13	Condesa	IF-1	-8.4	-55
13	Condesa	IF-2	-14.1	-101
13	Portales	IF-3	-6.8	-44
13	Portales	IF-4	-14.5	-105
13	Ramos Millán	IF-5	-7.3	-47
13	Ramos Millán	IF-6	-15	-109
13	Tetelpan	IF-7	-4.4	-23
13	Tetelpan	IF-8	-12.7	-90
13	Inst.de fisica	IF-9	-11.4	-79
13	Trea Marías	IF-10	-14.6	-102
14	Tecómitl 19	M1a	-6.7	-55
14	Tecómitl 19	M1b	-6.9	-43
14	Ojo De Agua	M-2a	-8	-52
14	Ojo De Agua	M-2b	-9.1	-63
14	Tlalpuente	M-3	-8.9	-58
14	Teotihuacan	M-4	-9.8	-64
14	IGF-UNAM	M-5a	-6.1	-46
14	IGF-UNAM	M-5b	-10.5	-75
14	IGF-UNAM	M-5c	-8.2	-51
14	Dee.d.l.Leon.	M-6a	-7.8	-63
14	Dee.d.l.Leon.	M-6b	-7.4	-44
14	Col.Roma Sur	M-7a	-6.9	-45
14	Col.Roma Sur	M-7b	-8	-51
14	Col.Roma Sur	M-7c	-10.1	-71
14	Tlamacas	M-8	-11.1	-71
14	Ramos Millán	M-9	-9.1	-81
14	Pachuca	M-10	-13.7	-97

Haciendo las modificaciones respectivas (que se muestran en el archivo lluvia.dat), y realizando el análisis, el archivo de resultados muestra lo siguiente:

Detección de Observaciones Influyentes (Tesis de Licenciatura)
Jaime González Martínez. Asesor: Gabriel Huerta Gomez
Facultad de Ciencias. Ciudad Universitaria. México 1994

Datos tomados del archivo: LLUVIA.DAT

Número de observaciones = 65
Número de regresores = 2

Puntos de corte para
outliers > 1.66864
Puntos de alta palanca > 0.07408
DFBETAS > 0.25198
DFFITS > 0.35635
Covarianza < 0.90759 o > 1.09231
Razón de Verosimilitudes > 7.81473 al 95% de confianza.

sigma = 37.185619
BETA[0] = 10.920140
BETA[1] = 7.954741

Obs	Outlier.	Alta Palanca.	Dffits.	CovRatio.	Raz de Veros.
4	1.2073	0.0563	0.2948	1.0444	0.0917
Dfbet	0.2908*	0.2513			
12	0.4070	0.1274*	0.1555	1.1770*	0.0303
Dfbet	-0.1149	-0.1458			
13	3.1865*	0.0743*	0.9025*	0.8235*	1.4717
Dfbet	-0.5909	-0.8036			
18	0.1419	0.0868*	0.0437	1.1299*	0.0094
Dfbet	-0.0299	-0.0397			
23	0.0215	0.0565	0.0053	1.0943*	0.0078
Dfbet	-0.0031	-0.0045			
30	0.0594	0.0641	0.0155	1.1031*	0.0080
Dfbet	0.0154	0.0135			
31	-0.2799	0.0575	0.0692	1.0928*	0.0115
Dfbet	0.0683	0.0592			
33	3.5652*	0.0335	0.6641*	0.7358*	1.4619
Dfbet	0.6216*	0.4885*			
49	-2.1570*	0.0256	0.3495	0.9168	0.2332
Dfbet	0.3062*	0.2206			
58	-2.0139*	0.0199	0.2870	0.9281	0.1607
Dfbet	0.2214	0.1367			
64	-3.5068*	0.0161	0.4490*	0.7308*	1.1488
Dfbet	0.2541*	0.0966			

Archivo de Resultados: LLUVIA__.res 21 de mayo de 1994 17:16

Si se omiten las observaciones 13, 33, y 64, pues más de dos medidas de influencia lo detectan, el nuevo modelo que se obtiene es BETA[0]=11.349, BETA[1]=8.026 por lo que resulta importante tomar en cuenta estas observaciones.

3.8 ERRORES CAUSAS Y SOLUCIONES.

Los errores que se señalan fueron vistos en las secciones anteriores, pero se pusieron aquí como una referencia rápida para el caso en que estos se puedan presentar.

- Archivo con mala estructura o tamaños de n y p incorrectos.
  Cuando aparece el mensaje:
  "Tamaño incorrecto de n o p o error en ARCHIVO.DAT <INTRO>=Cont"
  Esto se debe a que no pudo leer los valores de n o/y p, o que estos son o menores o iguales a cero; o nxp > 1000.
  Esto puede ocurrir si n o/y p no estan dados en el primer renglón del archivo.
  Solución: Revise que las condiciones anteriores se cumplan.

- Matriz no invertible.
  Cuando manda el siguiente mensaje en la opción 0 de influencia
  " Matriz XtX no invertible <INTRO>=Cont "
  o
  " Matriz ZtZ no invertible <INTRO>=Cont "
  en las opciones 1 y 2; es debido a que x^2x ó z^1z no tienen inversa y por lo tanto no es posible realizar su regresión y no se puede efectuar su análisis.
  Solución: Revise si los datos son correctos, de estar bien no se puede efectuar el análisis.

- Cuando en la selección de un archivo ésta sección esta llena, y no se encuentra el archivo buscado.
  Solución: Borrar o mover para otro lugar los archivos que en ese momento no se vayan a ocupar.

- Cuando al llamar a una función de influencia pierda el archivo seleccionado.
  Solución: Esto llega a ocurrir cuando existen programas residentes en memoria. Para corregir esto hay que descargarlos y volver a llamar al programa.

- Si estando en el Menú de influencia se le pide cualquier opción y no responde, es decir no se empieza a ejecutar la opción deseada, es por que o no estan todos los programas ejecutables en el mismo directorio o por que alguno de estos esta dañado.
  Solución: Revise que todos los programas ejecutables esten en el mismo directorio. Si estos estan entonces resuetituyalos pues estos pueden estar dañados.

- Terminación anormal: Matriz mal planteada.
  Esto llega a ocurrir muy raramente cuando uno de los parámetros estimados está enteramente determinado por una variable dependiente.
  Solución: Revise que los datos sean los correctos, de estar bien no es posible realizar el análisis.

3.9 REFERENCIA

- Gene H. Golub. Matrix Computations. 1983, The Johns Hopkins University press. Baltimore Maryland.

- Brian W. Kernighan, Dennis M Ritchie. El lenguaje de programación C, segunda edición, (con base en el ANSI C) 1991 Prentice Hall.

- Selected Algorithms from the ACC. A publication of the Association for comptuing Machinery Inc (Volume I & II 1980) (226 Normal Distribution Funcion; 229 Chi-Squared Integral; 322 F Distribution; 395 Students Distribution)

- R.E. Soraton. Métodos Numéricos Básicos, Introducción a las matemáticas numéricas con base en la microcomputadora, 1987, McGraw-Hill.

CONCLUSIONES.

Las medidas más sensibles con algunos de los ejemplos y datos mostrados fueron la Lamda de Wilk y el COVRATIO. La detección de outliers y puntos de alta palanca también resultó eficaz, no así la razón de las verosimilitudes que mostró ser poco sensible; las entradas de la matriz DFBETA también resultaron ser buenos detectores de influencia y útiles sobre todo para cuando se tiene un mayor interés sobre alguna entrada especifica del vector β y la DFFITS útil en cuanto a pronóstico.

El tener un programa de cómputo con diferentes criterios para detectar la influencia o bien la importancia de ciertos datos en el modelo de regresión lineal múltiple, fue cumplido, con las limitantes de tamaño anteriormente mencionadas, sin embargo los algoritmos quedan para su uso posterior para sistemas más poderosos.

APÉNDICE.

Aquí se revisan con un poco más de detalle algunos de los resultados antes mencionados.

$$\hat{\beta}-\hat{\beta}_{(i)} = (x^t x)^{-1}x^t y - (x^t_{(i)}x_{(i)})^{-1}x^t_{(i)}y_{(i)}$$

$$= (x^t x)^{-1}x^t y - \left[(x^t x)^{-1} + \frac{(x^t x)^{-1}x_i^t x_i (x^t x)^{-1}}{1-p_{ii}}\right]x^t_{(i)}y_{(i)}$$

$$= (x^t x)^{-1}x^t y - \left[(x^t x)^{-1} + \frac{(x^t x)^{-1}x_i^t x_i (x^t x)^{-1}}{1-p_{ii}}\right]\left(x^t y - x_i^t y_i\right)$$

como $p_{ii}=x_i(x^t x)^{-1}x_i^t$ $\quad \hat{y}_i=x_i(x^t x)^{-1}x^t y$ y $\quad e_i=y_i-\hat{y}_i$

se obtiene que

A.1 $$\hat{\beta}-\hat{\beta}_{(i)} = \frac{(x^t x)^{-1}x_i^t e_i}{1-p_{ii}}$$

si se multiplica x_i por la izquierda y se despeja $x_i\hat{\beta}_{(i)}$ se tiene

A.2 $$x_i\hat{\beta}_{(i)} = \frac{y_i-\hat{y}_i p_{ii}}{1-p_{ii}} .$$

Para probar la relación entre $\hat{\sigma}^2$ y $\hat{\sigma}^2_{(i)}$

$$(n-p-1)\hat{\sigma}^2_{(i)} = \sum_{j\neq i}(y_i-x_i\hat{\beta}_{(i)})^2$$

utilizando A.1

$$(n-p-1)\hat{\sigma}^2_{(i)} = \sum_{j=1}^{n}\left(e_j + \frac{p_{ij}e_i}{1-p_{ii}}\right)^2 - \frac{e_i^2}{(1-p_{ii})^2}$$

$$= (n-p)\hat{\sigma}^2 + \frac{2e_i}{1-p_{ii}}\sum_{j=1}^{n}e_j p_{ij} + \frac{e_i^2}{(1-p_{ii})^2}\sum_{j=1}^{n}p_{ij}^2 - \frac{e_i^2}{(1-p_{ii})^2}$$

$$= (n-p)\hat{\sigma}^2 - \frac{e_i^2}{(1-p_{ii})^2} .$$

aquí se utiliza el hecho de que p anula al vector de residuales.

Para probar $|X_{(i)}^t X_{(i)}| = (1-p_{ii})|X^t X|$ es necesario mostrar que

$$|I-uv^t| = 1 - uv^t$$

donde u y v son vectorse columna. Si Q es una matriz ortonormal tal que

$$Qu = \|u\| J_1$$

donde J_1 es el primer vector de la base estandar entonces

$$\begin{aligned} |I-uv^t| &= |Q(I-uv^t)Q^t| \\ &= |I-\|u\| J_1 v^t Q^t| \\ &= 1-v^t Q^t J_1 \|u\| \\ &= 1-uv^t. \end{aligned}$$

Ahora

$$\begin{aligned} |X_{(i)}^t X_{(i)}| &= |X^t X - X_i^t X_i| \\ &= |(I - X_i^t X_i (X^t X)^{-1}) X^t X| \end{aligned}$$

y si $u = X_i^t$ y $v^t = X_i (X^t X)^{-1}$ y como $p_{ii} = X_i (X^t X)^{-1} X_i^t$ se llega a

A.4 $$|X_{(i)}^t X_{(i)}| = (1-p_{ii})|X^t X|.$$

Por la forma en que esta construida la matriz $P = X(X^t X)^{-1} X^t$ $P^2 = P$ por lo que

$$p_{ii} = \sum_{j=1}^{n} p_{ij}^2 = p_{ii}^2 + \sum_{j \neq i} p_{ij}^2$$

entoncee $p_{ii} \geq 0$ y como

$$p_{ii} - p_{ii}^2 = \sum_{j \neq i} p_{ij}^2 \geq 0$$

entoncee $1 \geq p_{ii}$

Ahora si $\tilde{X}$ es la matriz X sin la columna de unos y restandole las medias obtenidas por columnas entonces

$$\hat{y} - \bar{y} = Py - \bar{y} = \tilde{P}y$$

por lo que

$$\tilde{p}_{ii} = p_{ii} - 1/n$$

se tiene que

$$1/n \leq p_{ii} \leq 1$$

y como los eigenvalores de una matriz de proyección o son cero o son 1 y el número de eigenvalores distintos de cero es igual al rango de la matriz y en este caso el rango de $\mathcal{P}$ es igual al rango de $\mathcal{X}$ y la traza de $\mathcal{P}$ es p, se obtiene

A.5 $$\sum_{i=1}^{n} \mathcal{P}_{ii} = p \ .$$

La comparación de medias de dos grupos en el cual uno consiste en un sólo punto es la estadistica Λ de Wilk y es

$$\Lambda(\tilde{x}_i) = \frac{| \tilde{x}^t\tilde{x} - (n-1)\bar{\tilde{x}}^t_{(i)}\bar{\tilde{x}}_{(i)} - \tilde{x}_i^t\tilde{x}_i |}{| \tilde{x}^t\tilde{x} |}$$

si el numerador se reescribe como

$$| \tilde{x}^t\tilde{x} - n^2/(n-1)\left(\bar{\tilde{x}} - \tilde{x}_i/n\right)^t\left(\bar{\tilde{x}} - \tilde{x}_i/n\right) - \tilde{x}_i^t\tilde{x}_i |$$

y usando el hecho de que $\tilde{x}$ es centrada lo anterior se reduce a

$$| \tilde{x}^t\tilde{x} - (n/(n-1))\tilde{x}_i^t\tilde{x}_i |$$

y utilizando A.4 para probar que el resultado anterior es

$$(1-n/(n-1)\tilde{x}_i(\tilde{x}^t\tilde{x})\tilde{x}_i^t)\,|\tilde{x}^t\tilde{x}|$$

entonces

$$\Lambda(\tilde{x}_i) = 1 - (n/(n-1))\mathcal{P}_{ii} = (n/(n-1))(1-\mathcal{P}_{ii})$$

De este modo si los renglones de $\tilde{x}$ son identicamente distribuidos con una distribución Gaussiana de dimención p-1 (Rao 1973 p.570) se llega a que

A.6 $$\left(\frac{n-p}{p-1}\right)\left(\frac{1-\Lambda(\tilde{x}_i)}{\Lambda(\tilde{x}_i)}\right) \sim F_{(p-1,n-p)} \ .$$