UNIVERSIDAD DE EL SALVADOR

FACULTAD DE INGENIERIA Y ARQUITECTURA

DEPARTAMENTO DE MATEMATICA

SUPERFICIES DE RESPUESTA

Trabajo de Graduación presentado por

Carlos Antonio Contreras Renderos

Previo a la opción del Título de

LICENCIADO EN MATEMATICA

Mayo 1984

San Salvador El Salvador Centro América

UNIVERSIDAD DE EL SALVADOR

RECTOR: DR. MIGUEL ANGEL PARADA

SECRETARIO GENERAL: DRA. ANA GLORIA CASTANEDA DE MONTOYA

FACULTAD DE INGENIERIA Y ARQUITECTURA

DECANO: ING. MANUEL ANTONIO CAÑAS LAZO

SECRETARIO: ING. RENE MAURICIO MEJIA MENDEZ

DEPARTAMENTO DE MATEMATICA

JEFE DEL DEPARTAMENTO: LIC. JOSE JAVIER RIVERA LAZO

ASESOR: Lic. Daniel Flores De Paz

TRABAJO PRESENTADO POR:

CARLOS ANTONIO CONTRERAS RENDEROS

PARA OPTAR AL TÍTULO DE:

LICENCIADO EN MATEMÁTICA

DEDICO CARIÑOSAMENTE ESTE TRABAJO A:

MIS PADRES: Carlos Renderos
Ana Contreras

MI ESPOSA: Marta Gloria Galdámez de Conteras

MIS HIJOS: Karla Georgina Contreras Galdámez
Carlos Antonio Contreras Galdámez

MIS HERMANOS: Mario Ernesto Contreras Renderos
Marta Gladis Contreras Renderos
Ana Leticia Contreras Renderos.

I N T R O D U C C I Ó N

En el presente trabajo, se cumple una de mi más grandes aspiraciones de estudiante, la cual era trabajar en la parte aplicada de la Matemática, habiendo escogido por tal motivo la rama de - Estadística, específicamente el tema de Análisis de Superficies de Respuesta, que es de gran utilidad en el campo de la investigación, industria, ciencias sociales etc.

Es de hacer notar que en análisis de regresión el problema fundamental es la solución de un sistema de ecuaciones, las cuales son llamadas ecuaciones normales. Las respuestas obtenidas podrán variar (por errores de redondeo), dependiendo del método - utilizado para resolverlas o del equipo de computación utilizado.

En el capítulo I, se estudia el ajuste de una recta por mínimos - cuadrados a un conjunto de datos, calculando los estimadores de la ecuación verdadera. Examinando la precisión de la ecuación - en cuadros de análisis de varianza (ANAVA), definiendo pruebas para significancia de los estimadores encontrados, y errores - cuando hay datos repetidos.

En el capítulo II, se estudia la utilización de matrices para encontrar los estimadores, como su utilización en tablas de análisis de varianza, ensayos de hipótesis, asignar pesos y se encuentran sesgos en los estimadores.

En el capítulo III, estudio de la normalidad o anormalidad, de los residuos, que son errores de las diferencias entre los valores observados y ajustados, en sucesión de tiempo en forma gráfica

y una medida de esas anormalidades.

En el capítulo IV, se estudia la regresión múltiple con dos variables, primero una ecuación con una variable, luego ingresando la segunda, es decir, como una sucesión de líneas, dando procedimientos para examinar la contribución de cada una de ellas en la ecuación.

En el capítulo V, se estudia la regresión con modelos exponenciales, polinomiales, potenciales y también se hace un pequeño estudio de la utilización de variables falsas.

En el capítulo VI, se estudia la parte medular de este trabajo, encontrar la mejor ecuación de regresión, describiendo varios procedimientos siendo: 1) todas las regresiones; 2) procedimiento de eliminación hacia atrás; 3) procedimiento de selección hacia adelante; 4) procedimiento de regresión paso a paso. Se analiza la conveniencia de unos y de otros.

Finalmente quiero dejar constancia de la colaboración a las personas que ayudaron a la elaboración de este trabajo: especialmente al Lic. Daniel Flores De Paz, asesor; Sra. Miriam de Yánez, parte mecanográfica; Sr. Mauricio García, gráficos y a la Universidad Politécnica por permitirme el uso de su computador HP 3000, para obtener las ecuaciones que van en el apéndice de este trabajo.

I N D I C E

PAGINA

C A P Í T U L O I

AJUSTANDO UNA RECTA DE MÍNIMOS CUADRADOS

1.0 INTRODUCCION

En la industria de hoy se encuentran procesos grandes o peque-- ños, en los cuales se pueden hacer mediciones de sus diferentes estados (etapas) y se recopilan esos datos en tablas, no se esta interesado en como recolectar esos datos, sino las relacio-- nes que se pueden dar en las variables de la tabla de datos, y poder observar que un cambio en una de esas variables, puede - afectar a las otras, por lo que es necesario examinar sus efectos, de alli que pueden hacer relaciones simples entre los estados (etapas) fisicas consideradas. Pudiendo a menudo existir - una relación funcional entre las variables, la cual puede ser - complicada determinar o descubrir en términos simples, pudiendo se aproximar, por ejemplo con un polinomio, el cual puede contener la variable apropiada, y puede ajustarse o graduarse a la - función verdadera, para luego poder apreciar la separación y - efectos producidos en los cambios hechos en las variables al - examinar la función aproximada con la función verdadera.

Un ejemplo simple de este proceso implica la construcción de una linea recta con parámetros desconocidos que se obtiene, a par-- tir del conjunto de observaciones $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, donde su construcción se verá más adelante. O un modelo más complicado, es cuando se tiene una variable Y y dos variables x_1, x_2 que nos da una ecuación de un plano. Dentro de los conocimientos básicos de analisis de regresión que utilizaremos estan alge-

bra de matrices, ideas de parámetros, estimadores, distribuciones especialmente la normal, media y varianza de un rango variable, covarianza entre dos variables, hipótesis simples, preguntas de uno o dos lados de prueba t ó prueba F.

Lo que se pretende es resolver problemas prácticos de regresión. Así aquí vemos observaciones tomadas con intervalos de una planta de vapor de una industria. Donde se han obtenido los datos de:

1. Depósito de vapor usado mensualmente
2. Depósito de glicerina cruda hecha
3. Días de calendario por mes
4. Días operados por mes
5. Días debajo de 32°F
6. Promedio de temperatura atmósferica (°F).

Podemos distinguir dos clases de variables en una tabla de datos. Las variables serán llamadas independientes y dependientes, que son valores observados, a diferencia de que cuando se hacen cambios en la variable independiente, producen cambios en la variable dependiente (o de respuesta). En general estamos interesados en encontrar como los cambios en las variables independientes afecta los valores en la variable de respuesta. Si podemos descubrir una relación simple entre las variables.

1.1 DOS ILUSTRACIONES

Así en muchos trabajos se desea investigar como el cambio en una variable afecta a otra variable, algunas veces dos variables son alcanzadas por una relación lineal exacta. Por ejemplo, si la resistencia R de un circuito simple es considerada constante, la variable I varía directamente con el voltaje, por aplicación

de la ley de Ohm; $I = \frac{V}{R}$, si no estamos seguros de la ley de Ohm; podriamos obtener la relación empirica, haciendo cambios en v y observando I, al mismo tiempo que R es fijo y entonces se tendría el diagrama de dispersión de I contra v, que es más o menos una linea recta cerca del origen.

Decimos "más o menos", porque aunque si bien la relación lineal no es exacta, nuestras medidas pueden estar sujetas a errores insignificantes y entonces el diagrama de los puntos no caería exactamente en la linea verdadera y deberían variar alrededor de ella.

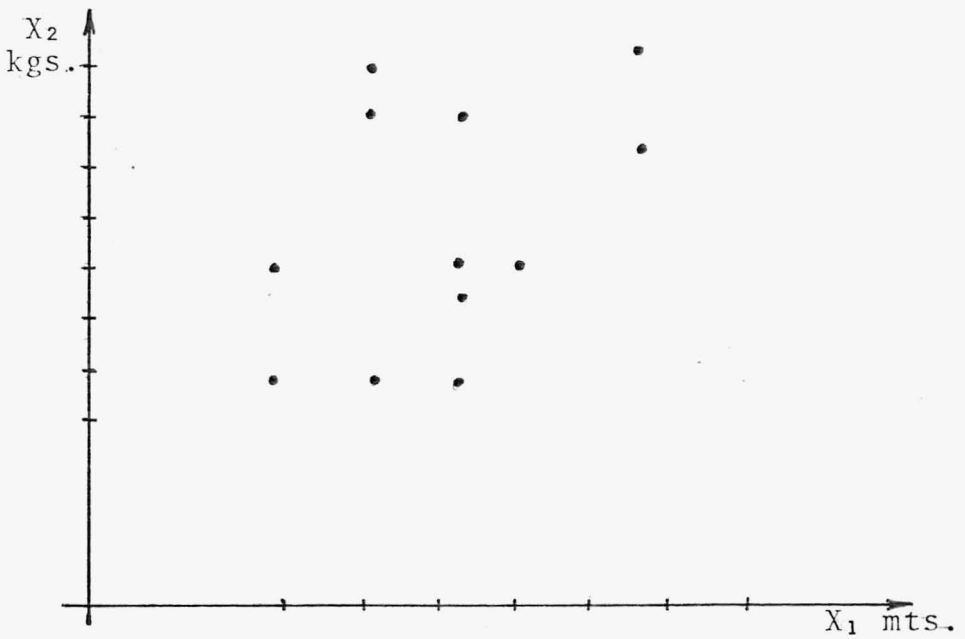
Para obtener un resultado de la predicción de I para un valor particular de v, usariamos la linea recta que pasa por el origen. Algunas veces la linea recta no es exacta (debido a los errores), pero todavia no podemos significarla bastante.

Otro ejemplo cuando queremos considerar peso y altura de varones adultos en una población dada. Si ploteamos el par (x_1,x_2) = (alturas, pesos). El diagrama viene dado en la figura 1.1

Note que para alguna altura dada hay un rango de pesos y viceversa . Esta variación en los rangos, será parcialmente debido a los errores en las medidas. Pero podemos notar que el promedio observado de peso para una altura dada incrementa cuando su altura incrementa. Este valor promedio observado de peso para un valor observado de altura (cuando la altura varía) es llamada la curva de regresión de peso en altura la que se denota por $x_2 = f(x_1)$. Pudiendose definir tambien una curva de regresión por $x_1 = g(x_2)$. Asumiendo que en general son las mismas, además que ambas son lineas rectas. La utilidad de una ecuación es por

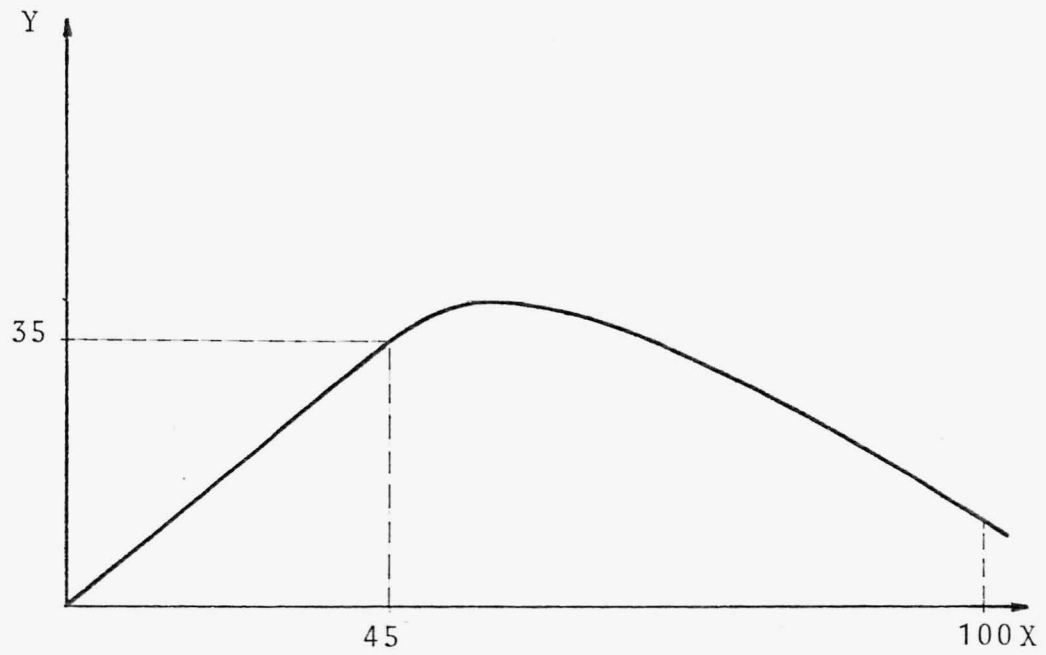
ejemplo, si desconocemos peso, lo podemos estimar a partir de - la linea de regresión de peso en altura. Cuando una variable x se relaciona con una variable y es usualmente llamada una ecuación de regresión, la cual es estimada cuando la relación entre ellas es desconocida.

FIGURA 1.1



Consideremos el gráfico en la figura 1.2, es obviamente no li-- neal en el rango $0 \leq x \leq 100$, sin embargo si se esta interesado en el rango de $0 \leq x \leq 45$, la relación es una linea recta, en - ese rango de observaciones, pero no podrian estimarse valores - mayores de 45, aunque es posible ajustar una linea recta en ese intervalo.

FIGURA 1.2



Similares observaciones pueden hacerse cuando tenemos más de una variable independiente. Supongamos que queremos examinar una - respuesta Y que depende de las variables $x_1, x_2, \ldots, x_k$.

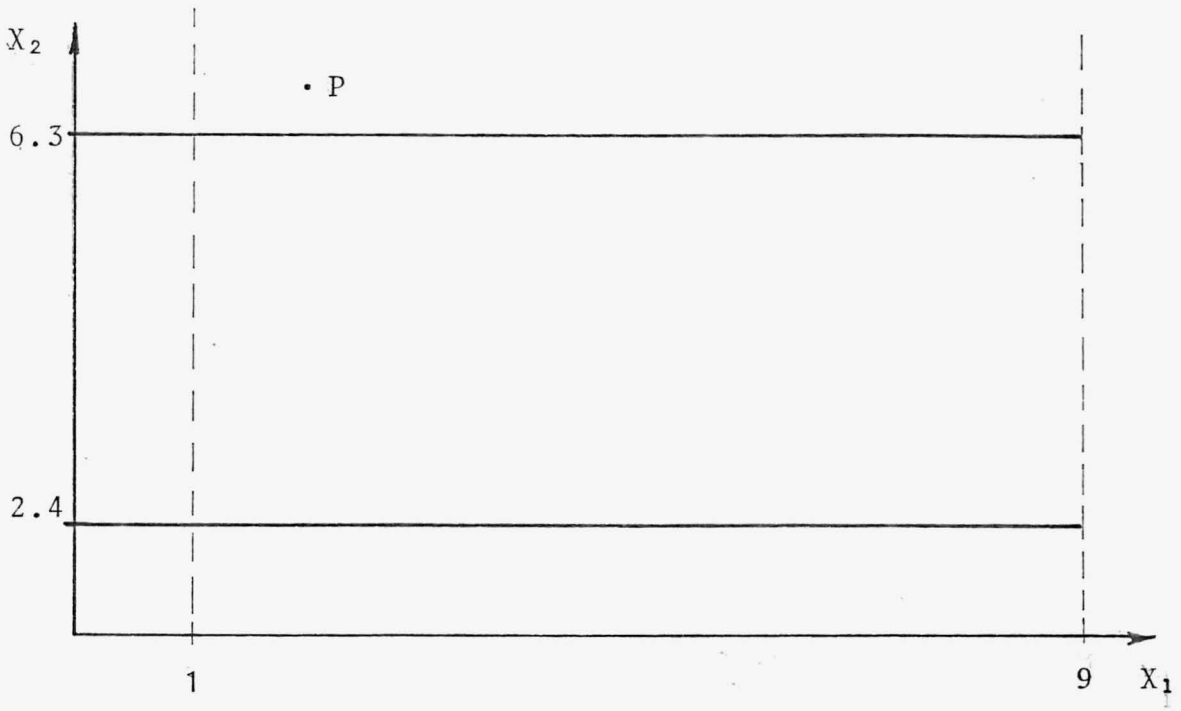
Determinamos una ecuación de regresión con los datos los cuales cubren una región del espacio y vamos a suponer que:

$x_{\hat{o}} = (x_{1o}, x_{2o}, \ldots, x_{ko})$ cae fuera de la región cubierta por los datos originales, entonces podemos obtener un valor estimado de $\hat{x}$ $(x_{\hat{o}})$ [representado, asi el valor estimado de x_o] para la respuesta de x_o, pero alguna vez puede suceder que el punto caiga más alla de la región, como se puede ver en la figura 1.3, donde hay puntos de la región para los cuales $1 \leq x_1 \leq 9$ y para - los cuales $2.4 \leq x_2 \leq 6.3$.

Para un punto p, esta fuera de la región, y se complica más cuan

do aumenta el número de dimensiones; lo que haría más dificulto so hacer predicciones.

FIGURA 1.3



1.2 REGRESION LINEAL: AJUSTANDO UNA LINEA RECTA

Para ajustar una ecuación de una linea recta, podrá ser obtenida por el método de mínimos cuadrados cuando los datos son precisos.

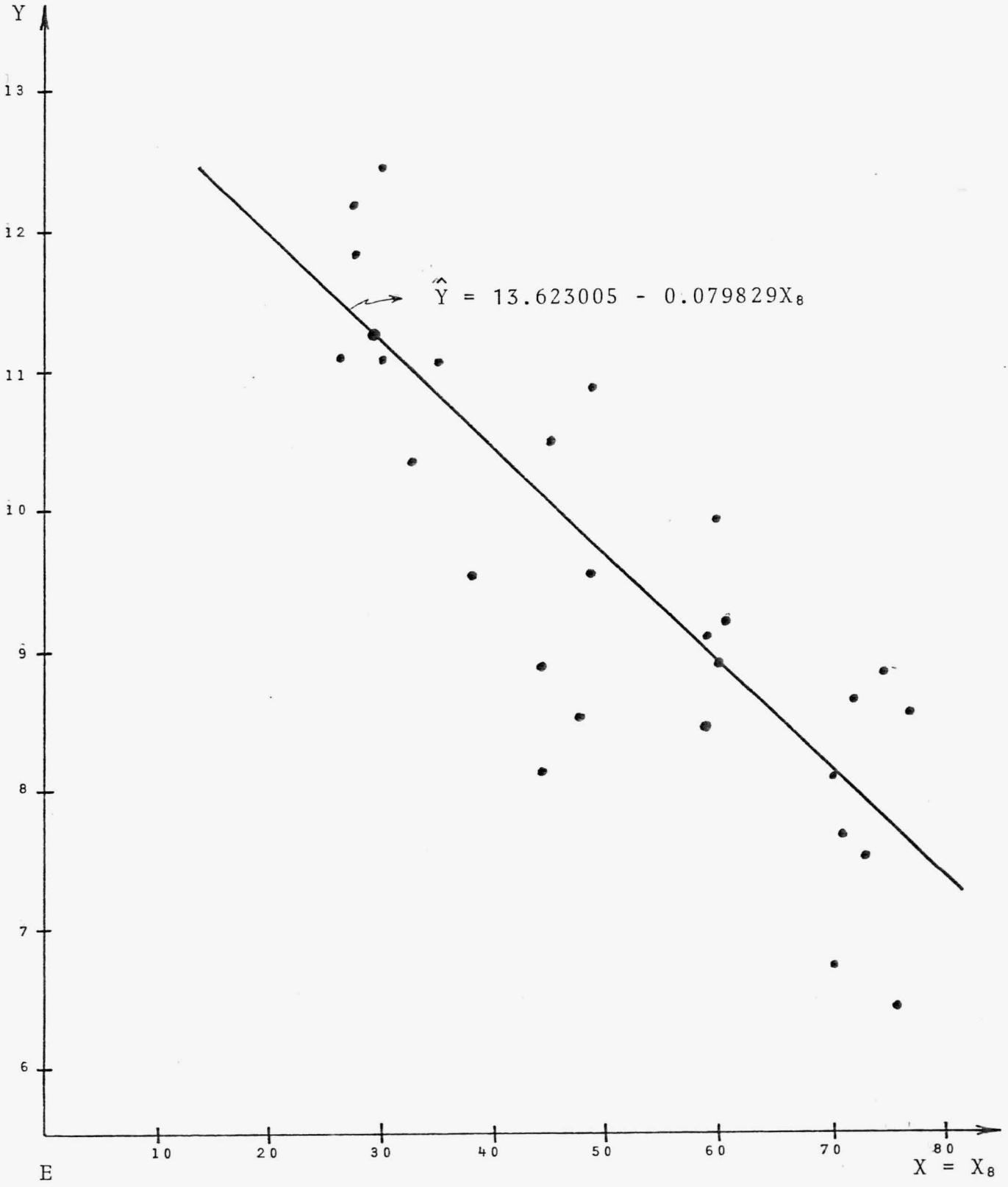
Consideremos las siguientes 25 observaciones de datos Y (vapor encerrado usado por mes) y X_8= (temperatura atmosférica en °F).

TABLA 1.1

NUMERO DE OBSERVACIONES	Y	X_8
1	10.98	35.3
2	11.13	29.7
3	12.51	30.8
4	8.40	58.8
5	9.27	61.4
6	8.73	71.3
7	6.36	74.4
8	8.50	76.7
9	7.82	70.7
10	9.14	57.5
11	8.24	46.4
12	12.19	28.9
13	11.88	28.1
14	9.57	39.1
15	10.94	46.8
16	9.58	48.5
17	10.09	59.3
18	8.11	70.0
19	6.83	70.0
20	8.88	74.5
21	7.68	72.1
22	8.47	58.1
23	8.86	44.6
24	10.36	33.4
25	11.08	28.6

Ver gráfico en figura 1.4

FIGURA 1.4


Y
13
12
11
10
9
8
7
6
$\hat{Y} = 13.623005 - 0.079829X_8$
10
20
30
40
50
60
70
80
E
$X = X_8$

De donde asumiremos que quien mejor la representa una linea recta de regresión de Y en la variable X, que tiene la forma $\beta_o + \beta_1 X$, la cual podemos escribir como

$$Y = \beta_o + \beta_1 X + \varepsilon \qquad 1.2.1$$

que esta ecuación de primer orden(1). Donde X es la variable independiente que le corresponde un valor Y de la forma $\beta_o + \beta_1 X$, más un valor ε, que seria la diferencia entre cada observación de Y y el valor de Y de la linea verdadera o la linea media, -- pudiendose llamarsele desviación.

Esta ecuación será nuestro modelo matemático, en donde β_o y β_1 son parámetros del modelo.

(1) Nota. Cuando decimos que un modelo es lineal o no lineal, - nos estaremos refiriendo a la linealidad o no linealidad en los parámetros. El valor más alto del exponente de una variable independiente en el modelo es llamado orden del modelo Ej.

$$Y = \beta_o + \beta_1 X + \beta_2 X^2 + \varepsilon$$

es de segundo orden en X, y es lineal en los betas.

De la ecuación 1.2.1, β_o, β_1 y ε son desconocidos, y el más - dificultoso de obtener es ε ya que tendría que encontrarse para cada una de las observaciones Y, mientras que β_o y β_1 quedan fijos, aunque no es posible encontrarlos exactamente, podemos utilizar la información de la tabla para encontrar b_o y b_1 estimadores de β_o y β_1, respectivamente, entonces podemos escribir la ecuación

$$\hat{Y} = b_o + b_1 X \qquad 1.2.2$$

donde $\hat{Y}$ es el valor estimado y denota la predicción del valor de Y para un X dado, cuando los b_o y b_1 son determinados. La ecuación 1.2.2 es usada para predecir.

El procedimiento utilizado será por minimos cuadrados para escoger el método de estimación de los parámetros.

Supongamos que tenemos n pares de observaciones (X_1,Y_1), (X_2,Y_2),. ..., (X_n,Y_n), entonces para la ecuación 1.2.1, podemos escribir

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i, \qquad 1.2.3$$

tq. la suma de los cuadrados de las desviaciones de la linea verdadera es

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2 \qquad 1.2.4$$

Como ya se mencionó antes no es posible calcular exactamente a β_o y β_1, por lo que se necesitarán estimadores b_o y b_i respectivamente, que se sustituirán en 1.2.4, de tal manera que produzcan el menor valor posible de S.

Note que los valores X_i y Y_i son valores fijos en la tabla podemos determinar b_o y b_1 diferenciando parcialmente primero con respecto a b_o y luego respecto a b_1 e igualandolos a cero.

$$\frac{\partial S}{\partial b_o} = -2\sum_{i=1}^{n}(Y_i - b_o - b_i X_i) \quad ; \quad \frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n}(Y_i - b_o - b_1 X_i)\, X_i \qquad 1.2.5$$

$$\sum_{i=1}^{n}(Y_i - b_o - b_1 X_i) = 0 \quad ; \quad \sum_{i=1}^{n}(Y_i - b_o - b_1 X_i)\, X_i = 0 \qquad 1.2.6$$

Aplicando sumatoria, tenemos.

$$\sum_{i=1}^{n} Y_i - nb_o - b_1 \sum_{i=1}^{n} X_i = 0 \quad ; \quad \sum_{i=1}^{n} X_i Y_i - b_o \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i^2 = 0 \qquad 1.2.7$$

ó

$$\sum_{i=1}^{n} Y_i = nb_o + b\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_iY_i = b_o\sum_{i=1}^{n} X_i + b_1\sum_{i=1}^{n} X_i^2 \qquad 1.2.8$$

Estas son llamadas ecuaciones normales, que al resolverlas para b_1 y b_o, encontramos el valor de b_1, que es

$$b_1 = \frac{\Sigma X_i \Sigma Y_i - n\Sigma X_iY_i}{(\Sigma X_i)^2 - n\Sigma X_1^2} = \frac{\sum X_iY_i - \frac{\sum X_i \sum Y_i}{n}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad 1.2.9$$

donde las sumatorias son desde $i = 1$ hasta n.

Observese que hay dos expresiones ligeramente diferente de b_1, por lo que definiendo. Para su comprobación, tenemos:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n} \; ; \; \bar{X} = \frac{\sum X_i}{n} \Rightarrow \sum X_i = n\bar{X} \qquad 1.2.10$$

$$\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y}{n} = \frac{\sum Y_i}{n} \; ; \; \bar{Y} = \frac{\sum Y_i}{n} \Rightarrow \sum Y_i = n\bar{Y} \qquad 1.2.11$$

tenemos que:

a) $\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y})$

$= \sum X_iY_i - \bar{Y}\sum X_i - \bar{X}\sum Y_i + \sum \bar{X}\bar{Y}$

por 1.2.10 y 1.2.11 tenemos

$= \sum X_iY_i - n\bar{Y}\bar{X} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y}$

$= \sum X_iY_i - n\bar{Y}\bar{X}$

$= \sum X_iY_i - n\frac{\sum Y_i}{n}\frac{\sum X_i}{n}$

$= \sum X_iY_i - \frac{\sum X_i Y_i}{n}$

de donde

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

b) $\sum(X_i - \bar{X})^2 = \sum(X_i^2 - 2X_i\bar{X} - (\bar{X})^2]$

$$= \sum X_i^2 - 2\bar{X}\sum X_i - (\bar{X})^2$$

por 1.2.10 tenemos

$$= \sum X_i^2 - 2n(\bar{X})^2 - n(\bar{X})^2$$

$$= \sum X_i^2 - n(\bar{X})^2$$

$$= \sum X_i^2 - n(\frac{\sum X_i}{n})^2$$

$$= \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

de donde

$\sum(X_i - \bar{X})^2 = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$. Así de a y b queda comprobado.

La primera forma de la ecuación 1.2.9 es normalmente usada cuando computamos el valor de b_1, la solución para b_o es obtenida de la primera ecuación normal de 1.2.8 que es:

$$nb_o + b_1\sum X_i = \sum Y_i$$

$$nb_o = \sum Y_i - b_1\sum X_i \quad , \text{ dividiendo por n}$$

$$b_o = \frac{\sum Y_i}{n} - b_1\frac{\sum X_i}{n}$$

$$b_o = \bar{Y} - b_1\bar{X} \qquad 1.2.12$$

La cantidad $\sum X_i^2$ es llamada suma no corregida de los cuadrados de las equis y $\frac{(\sum X_i)^2}{n}$ es la corrección para la media de las equis. Su diferencia es llamada la suma corregida de los cuadra

dos de las equis.

Similarmente $\sum X_i Y_i$ es llamada suma no corregida de productos y $\frac{\sum X_i \sum Y_i}{n}$ es llamada la suma corregida para la media.

La diferencia es llamada la suma corregida del producto XY.

Ahora sustituyendo 1.2.12 en la ecuación 1.2.2 tenemos

$$\hat{Y} = \bar{Y} - b_1\bar{X} + b_1X$$

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}) \qquad 1.2.13$$

donde b_1 esta dado por la ecuación 1.2.9

Ahora se ejecutarán cálculos para determinar la ecuación de predicción de la tabla 1.1

$\sum Y_i = 235.60$ $\quad \sum X_i^2 = 76323.42$ $\quad \bar{X} = \frac{1315}{25} = 52.60$

$\sum X_i = 1315.0$ $\quad \bar{Y} = \frac{235.60}{25} = 9.424$

$\sum X_i Y_i = 11821.432$

cuando b_1

$$b = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{11811.432 - \frac{(235.60)(1315)}{25}}{76323.42 - \frac{(1315)^2}{25}} = \frac{-571.128}{7154.42} = -0.079829$$

La ecuación ajustada será entonces

$$\hat{Y} = \bar{Y} + b_1(X_8 - \bar{X})$$

$$\hat{Y} = 9.424 + (-0.079829)(X_8 - 52.60)$$

$$\hat{Y} = 13.623005 - 0.079829\,X_8 \qquad 1.2.14$$

La ecuación ajustada esta ploteada en la figura 1.4

Ahora se va a tabular, para cada valor de los X_i, dados en el orden de la tabla anterior, los valores de $\hat{Y}_i$ y ademas los Y_i -

para encontrar $Y_i - \hat{Y}_i$, que les llamaremos residuos.

TABLA 1.2 valores ajustados, observaciones y residuos, en el orden de la tabla 1.1

NUMERO DE OBSERVACIONES	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	10.98	10.81	0.17
2	11.13	11.25	-0.12
3	12.51	11.17	1.34
4	8.40	8.93	-0'53
5	9.27	8.72	0.55
6	8.73	7.93	0.80
7	6.36	7.68	-1.32
8	8.50	7.50	1.00
9	7.82	7.98	-0.16
10	9.14	9.03	0.11
11	8.24	9.92	-1.68
12	12.19	11.32	0.87
13	11.88	11.38	0.50
14	9.57	10.50	-0.93
15	10.94	9.89	1.05
16	9.58	9.75	-0.17
17	10.09	8.89	1.20
18	8.11	8.04	0.07
19	6.83	8.04	-1.21
20	8.88	7.68	1.20
21	7.68	7.87	-0.19
22	8.47	8.98	-0.51
23	8.86	10.06	-1.20
24	10.36	10.96	-0.60
25	11.08	11.34	-0.26

NOTA: los $\hat{Y}_i$ son valores ajustados aproximados.

Note que si

$$\hat{Y}_i = \overline{Y} + b_1(X_i - \overline{X}),$$ y utilizando su opuesto

$-\hat{Y}_i = -Y - b_1(X_i - \bar{X})$, sumandole Y , tenemos

$$Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - b_1(X_i - \bar{X})$$

Aplicando sumatoria de $i = 1$ a n. tenemos

$$\sum(Y_i - \hat{Y}_i) = \sum[(Y_i - \bar{Y}) - b_1(X_i - \bar{X})]$$

$$= \sum(Y_i - \bar{Y}) - b_1\sum(X_i - \bar{X})$$

$$= \sum X_i - n\bar{Y} - b_1(\sum X_i - n\bar{X})$$

Aplicando 1.2.10 y 1.2.11

$$= n\bar{Y} - n\bar{Y} - b_1(n\bar{X} - n\bar{X})$$

$$\sum(Y_i - \hat{Y}) = 0$$

Entonces la suma de residuos debería ser cero.

Al sumarlos en la tabla 1.2 nos da -0.02 es un error de redon-- deo. La suma de residuos en un problema de regresión es siempre cero.

Cuando hay un término β_o en el modelo es consecuencia de la primera ecuación normal. Puede tenerse su omisión en el modelo - es decir, hemos perdido un parámetro, pero hay una correspondiente pérdida en los datos puesto que las cantidades $Y_i - \bar{Y}$, $i=1,\ldots,n$, representa solamente n-1 datos separados de información, debido al hecho que su suma es cero, mientras que $Y_1, Y_2, \ldots, Y_n$ representa n datos de información. Efectivamente el dato de información ha sido usado,para permitir el propio ajuste a ser hecho - al modelo talque el intercepto puede ser removido.

1.3 LA PRECISION DE LA REGRESION ESTIMADA

La pregunta es ¿De que medida de precisión, puede ser unida a - nuestra estimación de la linea de regresión?.

Consideremos la siguiente identidad.

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - (\hat{Y} - \bar{Y}) \qquad 1.3.1$$

elevando al cuadrado y luego aplicando sumatoria de $i = 1$ a n

$$\Sigma(Y_i-\hat{Y}_i)^2 = \Sigma\left[(Y_i-\bar{Y}) - (\hat{Y}-\bar{Y})\right]^2$$

$$= \Sigma\left[(Y_i-\bar{Y})^2 - 2(Y_i-\bar{Y})(\hat{Y}-\bar{Y}) + (\hat{Y}-\bar{Y})^2\right]$$

$$\Sigma(Y_i-\hat{Y}_i)^2 = \Sigma(Y_i-\bar{Y})^2 + \Sigma(\hat{Y}-\bar{Y})^2 - 2\Sigma(Y_i-\bar{Y})(\hat{Y}-\bar{Y}) \qquad 1.3.2$$

el tercer término se puede reescribir

$$-2\Sigma(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}) = -2\Sigma(Y_i-\bar{Y}).b_1(X_i-\bar{X}) \quad \text{por 1.2.13}$$

$$= -2b_1\Sigma(Y_i-\bar{Y})(X_i-\bar{X})$$

de 1.2.9 tenemos

$$b_1 = \frac{\Sigma(Y_i-\bar{Y})(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2} \Rightarrow \Sigma(Y_i-\bar{Y})(X_i-\bar{X}) = b_1\Sigma(X_i-\bar{X})^2$$

Así

$$-2\,\Sigma(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}) = -2b_1^2\Sigma(X_i-\bar{X})^2$$

pero por 1.2.13

$$\hat{Y} - \bar{Y} = b_1(X_i-\bar{X})$$

$$\Sigma(\hat{Y}-\bar{Y})^2 = \Sigma b_1^2(X_i-\bar{X})^2 = b_1^2\Sigma(X_i-\bar{X})^2 \qquad 1.3.3$$

por lo que continuando, se tenía

$$= -2b_1^2\Sigma(X_i-\bar{X})^2$$

$$= -2\Sigma(\hat{Y}-\bar{Y})^2 \quad \text{al sustituir 1.3.3}$$

tenemos:

$$-2\Sigma(Y_i-\bar{Y})(\hat{Y}_i-\bar{Y}) = -2\Sigma(\hat{Y}_i-\bar{Y})^2$$

Así al sustituir en 1.3.2 esta última expresión tenemos:

$$\Sigma(Y_i-\hat{Y})^2 = \Sigma(Y_i-\bar{Y})^2 + \Sigma(\hat{Y}_i-\bar{Y})^2 - 2\Sigma(\hat{Y}_i-\bar{Y})^2$$

$$\Sigma(Y_i-\hat{Y})^2 = \Sigma(Y_i-\bar{Y})^2 - \Sigma(\hat{Y}-\bar{Y})^2 \qquad 1.3.4$$

que se puede escribir asi

$$\Sigma(Y_i-\bar{Y})^2 = \Sigma(Y_i-\hat{Y}_i)^2 + \Sigma(\hat{Y}_i-\bar{Y})^2 \qquad 1.3.5$$

Ahora $Y_i-\bar{Y}$, es la i-esima desviación respecto a la media, del - lado izquierdo de 1.3.3, es la suma de los cuadrados de las desviaciones respecto a la media y que representaremos por sc y es tambien la suma corregida de cuadrados de las y es

Entonces $Y_i-\hat{Y}$, es el i-ésimo residuo (es la desviación de la - i-ésima observación con respecto al valor ajustado), y $\hat{Y}_i-\bar{Y}$, es la i-ésima observacion del valor ajustado respecto a la media.

La ecuación 1.3.5 en palabras es como sigue.

SUMA DE CUADRADOS RESPECTO A LA MEDIA = SUMA DE CUADRADOS ACERCA DE LA REGRESION + SUMA DE CUADRADOS DEBIDO A LA REGRESION

Esto demuestra que la variación de los Y's alrededor de la me-- dia, puede ser debido a la regresión y a algún, $\Sigma(Y_i-\hat{Y}_i)^2$, al - hecho que las actuales observaciones no todas se mantuvieran.

CUADRO DE ANALISIS DE VARIANZA

FUENTE	SUMA DE CUADRADOS (sc)	gl	CUADRADOS MEDIOS (cm)
DEBIDO A LA REGRESION	$b_1\left\{\Sigma X\ Y - \frac{\Sigma X\ \Sigma Y}{n}\right\}$	1	MS_R
ACERCA DE LA REGRESIOM (RESIDUOS)	Por diferencia	n_1-2	$S = \frac{sc}{n-2}$
ACERCA DE LA MEDIA (TOTAL, CORREGIDO PARA LA MEDIA)	$\Sigma Y_i - \frac{(\Sigma Y_i)}{n}$	n-1	

La columna de "cuadrados medios" es obtenida dividiendo cada suma de cuadrados por el correspondiente número de grados de libertad.

Una forma más general de la tabla de análisis de varianza, la cual no utilizaremos aquí, es obtenida por incorporación del factor de corrección para la media de los yes, el cual denotaremos por $sc(b_o)$. La tabla toma la forma.

CUADRO DE ANALISIS DE VARIANZA

FUENTE	SUMA DE CUADRADOS	gl	CUADRADOS MEDIOS
REGRESION (b_o)	$sc(b_o) = \frac{(\Sigma Y_i)^2}{n}$	1	
DEBIDO A LA REGRESION (b_1/b_o)	$sc(b_1/b_o) = b_1\left\{\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}\right\}$		MS_R
RESIDUOS	por sustracción	n-2	$S^2 = \frac{sc}{n-2}$
TOTAL, NO CORREGIDO PARA LA MEDIA.	ΣY_i^2	n	

La notación $sc(b_1/b_o)$ es leido " la suma de cuadrados para b_1 despues de la asignación hecha para b_o ". La media de cuadrados acerca de la regresión, s^2, es un estimador en la linea de regresión. (Si todos los puntos permanecieran en ella, la suma de cuadrados acerca de la regresión sería cero).

De este procedimiento podemos ver que un camino de indicar la utilidad de la linea de regresión como predictor es ver como la sc acerca de la media se relaciona con la sc acerca de la regresión. Estaremos complacidos si la suma de cuadrados debido a la regresión es mucho más grande que la sc acerca de la regre

sión o que el valor R^2 = (sc debido a la regresión)/(sc acerca de la media) no es demasiado lejos de la unidad.

Alguna suma de cuadrados tiene asociada un número llamado grados de libertad. Este número indica cuantos datos son independientes, al implicar los números Y_1, $Y_2,\ldots,Y_n$., son necesitados para obtener la suma de cuadrados.

Por ejemplo, la sc acerca de la media necesita (n-1) datos independientes (para los números $Y_1-\bar{Y},Y_2,-\bar{Y},\ldots,Y_n-\bar{Y}$, solamente --- (n-1) son independientes puesto que los n números sumados dan - cero, por propiedad de la media). Podemos calcular sc debido a la regresión de una función sencilla de $Y_1,Y_2,\ldots,Y_n$ nombrandola b_1(puesto que $\Sigma(\hat{Y}_1-\bar{Y}) = b_1^2\Sigma(X-\bar{X})^2$ por 1.3.3) y esta suma -- tiene un grado de libertad, ya que la única cantidad independiente es b_1.

Por diferencia la sc acerca de la regresión tiene n-2 grados de libertad. Esto se explica así

$$(n - 1) = (n - 2) + 1 \qquad 1.3.6$$

Usando las ecuaciones 1.3.5 y 1.3.6 empleando una forma alternativa para la expresión 1.3.5 podemos construir una tabla de analisis de varianza en la siguiente forma basado en n-2 grados de libertad de la varianza sobre la regresión que denotaremos por $\sigma^2_{Y\hat{Y}}$, que nos representaría una medida del error con el cual un valor observado de Y, debería ser predecido de un valor de X de una ecuación determinada.

Haremos ahora en nuestro ejemplo que traemos y discutir un número de caminos la ecuación de regresión puede ser examinada.

La sc debido a la regresión es

$$b_1 \left\{ \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n} \right\} = \frac{\{\sum X_i Y_i - \sum X_i \sum Y_i / n\}^2}{\{\sum X_i^2 - (\sum X_i)^2/n\}}$$

$$= \frac{(-571.128)^2}{7154.42} = 45.59$$

$$\text{La sc del total (corregido)} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

$$= 2284,1102 - (235.6)^2/25$$

$$= 63.82$$

TABLA 1.3 TABLA DE ANALISIS DE VARIANZA PARA EL EJEMPLO.

FUENTE	gl	sc	cm	F valor calculado
total (corregido	24	63.82		
Regresión (b_1)	1	45.59	45.59	$\frac{45.59}{0.7920} = 57.52$
Residuos	23	18.23	s^2=0.7926	

Note que las entradas en esta tabla, no estan en el mismo orden como la correspondiente tabla teórica, lo que no hace una diferencia. En muchos casos el orden depende del camino en el cual el programa es escrito. Una inspección cuidadosa de analisis de varianza debería siempre ser hecho y no se podría asumir que algún orden particular es el mismo. Nuestro estimador de σ^2_{YX} es $s^2 = 0.7926$ basada en 23 grados de libertad.

1.4 EXAMINANDO LA ECUACION DE REGRESION

Haremos las suposiciones básicas que en el modelo

$$Y = \beta_o + \beta_1 X_i + \varepsilon_i, \quad i = 1,2,\ldots,n.$$

(1). ε_i es una variable aleatoria con media cero y varianza

σ^2 (desconocida)

(2). ε_i y ε_j son no correlacionados, $i \neq j$, se tiene $\text{co}V(\varepsilon_i,\varepsilon_j)=0$

entonces $E(Y_i) = \beta_o + \beta_1 X_i$, $\quad V(Y_i) = \sigma^2$

y Y_i y Y_j, $i \neq j$, son no correlacionados. Una suposición adicional, la cual no es inmediatamente necesaria y será llamada cuando sea usada, es la siguiente.

(3). ε_i es una variable aleatoria, normalmente distribuida, con media cero y varianza σ^2 por (1), esto, es

$$\varepsilon_i \sim N(0,\sigma^2)$$

bajo esta condición adicional, ε_i y ε_j no son únicamente no correlacionados, pero necesariamente independientes.

OBSERVACIONES

(1) Anteriormente hemos encontrado $\sigma^2 YX$ en el cual, si $\sigma^2=\sigma^2_{YX}$ tenemos que el modelo encontrado es el verdadero, si no lo es, entonces $\sigma^2 < \sigma^2_{YX}$

(2) Cuando se hacen mediciones, cálculos, se tienen errores ε_i, tq. una suma de errores que tenderá a ser normal cuando el número de componentes aumenta más y más, por el teorema del límite central.

Ahora usaremos esas suposiciones en examinar la ecuación de regresión.

ERROR ESTANDAR DEL SESGO b_1: INTERVALO DE CONFIANZA DE β_1

sabemos que $b_1 = \dfrac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2} =$

$$= \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}$$

$$b_1 = \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}$$

(puesto que el segundo término del numerador es $\Sigma(X_i-\bar{X})\bar{Y} = \bar{Y}\Sigma(X_i-\bar{X}) = 0$). esto es porque $\Sigma(X_i-\bar{X}) = 0$)

$$b_1 = \frac{(X_1-\bar{X})Y_1 + (X_2-\bar{X})Y_2 + \ldots + (X_n-\bar{X})Y_n}{\Sigma(X_i-\bar{X})^2}$$

Ahora la varianza de la función.

$$F = a_1Y_1+a_2Y_2+ \ldots + a_nY_n$$

donde cada a_i, son constantes y para cada par de Y_i son no correlacionados, además $V(Y_i) = \sigma^2$, $\forall_i$.

Entonces su varianza sería.

$$V(F) = V(a_1Y_1+a_2Y_2+ \ldots + a_n Y_n)$$

$$= V(a_1Y_1)+V(a_2Y_2)+ \ldots + V(a_nY_n)$$

$$= a_1^2V(Y_1) + a_2^2V(Y_2) + \ldots + a_n^2V(Y_n)$$

$$= a_1^2 \sigma^2+ a_2^2 \sigma^2+ \ldots + a_n^2 \sigma^2$$

$$V(F) = (a_1^2 + a_2^2 + \ldots + a_n^2) \sigma^2$$

Ahora encontraremos la varianza de b_1, si

$$b_1 = \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X -\bar{X})^2}, \quad \text{con } a_i = \frac{(X_i-\bar{X})}{\Sigma(X_i-\bar{X})^2}$$

donde cada X_i puede ser considerado constante

$$V(b_1) = V\left\{\frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}\right\} = V[a_iY_i] \quad , i = 1 \text{ a } n$$

asi tenemos:

$$V(b_1) = \left\{\frac{(X_1-\bar{X})^2}{[\Sigma(X_i-\bar{X})]^2} + \frac{(X_2-\bar{X})^2}{[\Sigma(X_i-\bar{X})^2]^2} + \frac{(X_n-\bar{X})^2}{[\Sigma(X_i-\bar{X})^2]^2}\right\} \sigma^2$$

$$= \left\{ \frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \ldots + (X_n-\bar{X})^2}{[\Sigma(X_i-\bar{X})^2]^2} \right\} \sigma^2$$

$$= \frac{\Sigma(X_i-\bar{X})^2}{[\Sigma(X_i-\bar{X})^2]^2} \cdot \sigma^2 = \frac{\sigma^2}{\Sigma(X_i-\bar{X})^2}$$

de donde $V(b_1) = \dfrac{\sigma^2}{\Sigma(X-\bar{X})^2}$ 1.4.1

La desviación típica para b_1, es la raiz cuadrada de la varianza.

$$d\cdot t\cdot(b_1) = \frac{\sigma}{\{\Sigma(X-\bar{X})^2\}^{1/2}}$$

y si σ es desconocida, entonces toma s como su estimador, quedando la desviación tipica asi:

$$d\cdot t\cdot e\cdot(b_1) = \frac{s}{\{\Sigma(X-\bar{X})^2\}^{1/2}} \qquad 1.4.2$$

Ahora se encontrará un intervalo de confianza para β_1, asumiendo que los ε_i, forman un solo conjunto con distribución normal $N(o,\sigma^2)$, pero es necesario definir una variable que contenga al estimador, definiendola asi:

$$t = \frac{(b_1-\beta_{1o})}{\{d\cdot t\cdot e\cdot(b_1)\}} = \frac{(b_1-\beta_{1o})\{\Sigma(X_i-\bar{X})^2\}^{1/2}}{s} \qquad 1.4.3$$

donde β_{1o} es un valor específico, que podría ser β_1 ó cero, t se usará con (n-2) grados de libertad que corresponden al estimador s^2.

Para encontrar los límites de confianza, se asigna un $100(1-\alpha)\%$ de nivel de confianza, donde los límites vienen dados por

$$b_1 \pm \frac{t(n-2,1-\frac{\alpha}{2}).s}{\{\Sigma(X_i-\bar{X}|^2\}^{1/2}} \qquad 1.4.4$$

donde $t(n-2,\ 1 - \frac{\alpha}{2})$, el valor de una distribución t con n-2 - grados de libertad y $1 - \frac{\alpha}{2}$ de nivel de confianza.

Ejemplo. Calcule el intervalo de cofianza para β_1, con los datos del ejemplo continuado.

$s^2 = 0.7926$; $\Sigma(X_i - \bar{X})^2 = 7154.42$, $n = 25$, $b_1 = -0.0798$

$$V(b_1) = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$$

$$V.e.(b_1) = \frac{s^2}{\Sigma(X_i - \bar{X})^2} = \frac{0.7926}{7154.42} = 0.00011078$$

$$d.t.e.(b_1) = \sqrt{V.e.(b_1)} = \sqrt{0.00011078} = 0.0105$$

Supongamos ahora que tenemos un nivel de confianza del 95%, lo que da $\alpha = 0.05$, al aplicarlo en la distribución t, y por ta-- blas de $t(23,0.975) = 2.069$ como valor. Entonces los límites - de confianza al 95% son

$$b_1 \pm \frac{t(23,0.975).s}{\{\Sigma(X - \bar{X})^2\}^{1/2}} \Rightarrow b_1 \pm t(23,0.975).d.t.e(b_1)$$

sustituyendo los valores correspondientes tenemos

$$-0.0798 \pm (2.069)(0.0105)$$

entonces los límites para β_1 son:

$$-0.1015 \leq \beta_1 \leq -0.0581$$

lo que significa que el valor verdadero de β_1 estará entre esos dos valores, con una probabilidad de 0.95.

Igualmente la prueba de la hipotesis nula que el valor de β_1 es - cero o que no hay relación entre temperatura atmosférica y la - cantidad de vapor usado. Como notamos anteriormente, escribimos (usando $\beta_{10} = 0$)

$$H_o: \beta_1 = 0 \quad , \quad H_1: \beta_1 \neq 0$$

y evaluamos

$$t = \frac{b_1}{d.t.e(b_1)} = \frac{-0.079}{0.0105} = -7.60$$

puesto que $|t| = 7.60$ excede al valor crítico dado por $t(23,0.975) = 2.069$ $H_o: \beta_1 = 0$ es rechazada. (con este valor $|t|$ excede también a $t(23,0.995)$), elegimos un nivel de la prueba 95% a dos lados, como quiera que sea, de tal manera el intervalo de cofianza y la prueba t ambos harían uso del mismo nivel de probabilidad.

Al observar los datos puede causarnos la idea de rechazar que podría existir una relación lineal entre Y y X.

Si se ha llegado, al caso de que el valor observado $|t|$, es más pequeño que el valor crítico, tendríamos que haber dicho que no rechazaríamos la hipótesis.

Note que no usamos la palabra aceptar .
Puesto que normalmente no podemos aceptar una hipotesis, lo más que podemos decir, es que en base a ciertos datos observados no podemos rechazarla.

Una vez tengamos el intervalo de confianza para β_1, podemos no computar el valor de $|t|$ para una prueba t particular. Es símple examinar el intervalo de confianza para β_1 y ver si contiene el valor de β_{10}, si esta, entonces la hipotesis $\beta_1 = \beta_{10}$ no puede ser rechazada, sino está, la hipotesis es rechazada, en el nivel $1 - \alpha$ si

$|t| > t(n-2, 1- \frac{1}{2} \alpha)$ lo que implica que

$|b_1 - \beta_{10}| > t(n-2,1- \frac{1}{2} \alpha).s/\{\Sigma(X - \bar{X})^2\}^{1/2}$

esto es, que β_{10} esta fuera de los limites dados en la ecuación 1.4.3

En conclusión decimos que hay relación entre X e Y.

DESVIACION TIPICA DEL INTERCEPTO: INTERVALO DE CONFIANZA PARA β_o

De manera similar, que se encontró un intervalo de confianza para β_1, se demostrará más adelante que

$$d.t.(b_o) = \left\langle \frac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} \right\rangle^{1/2} \cdot \sigma$$

si σ es desconocida, entonces es sustituida por s, su estimador., asi.

$$d.t.e(b_o) = \left\langle \frac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} \right\rangle^{1/2} .s \qquad 1.4.5$$

Entonces los límites de confianza al 100 $(1 - \alpha)$% para β_o vienen dados por

$$b_o \pm t(n-2, 1 - \tfrac{1}{2}\alpha) \left\langle \frac{\Sigma X^2}{n\Sigma(X-\bar{X})^2} \right\rangle^{1/2} .s \qquad 1.4.6$$

Una pruebat para hipotesis nula H_o: $\beta_o = \beta_{oo}$ contra la alternativa H_1: $\beta_o \neq \beta_{oo}$, donde β_{oo} es un valor específico, será rechazado al nivel de $(1-\alpha)$ si β_{oo} cae fuera del intervalo de confianza, o no rechazado si β_{oo} cae dentro, o puede ser manejado separadamente al encontrar la cantidad

$$t = \frac{(b_o-\beta_{oo})}{\left\langle \frac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} \right\rangle^{1/2} .s} = (b_o-\beta_{oo}) \left\langle \frac{n\Sigma(X_i-\bar{X})^2}{\Sigma X_i^2} \right\rangle \frac{1}{s} \qquad 1.4.7$$

y comparandolos con el porcentaje de puntos $t(n-2, 1-\frac{1}{2}\alpha)$. Puesto que $(n-2)$ es el número de grados de libertad de s^2 estimador de σ^2, es basado.

Desviación tipica de $\hat{Y}$

Hemos demostrado que la ecuación de regresión es

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}).$$

donde ambos $\bar{Y}$ y b_1 son sujetos de error, los cuales influenciaran a $\hat{Y}$. Ahora si a_i y c_i son constantes y

$$a = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$$

$$c = c_1Y_1 + c_2Y_2 + \ldots + c_nY_n$$

se ha visto que Y_i y Y_j son no correlacionados, cuando $i \neq j$, y si $V(Y_i) = \sigma^2$, para todo i.

$$\text{cov}(a,c) = (a_1c_1 + \ldots + a_nc_n)\,\sigma^2 \qquad 1.4.8$$

se sigue que haciendo $a = \bar{Y}$, lo que implica que $a_i = \frac{1}{n}$ y haciendo $c = b_1$, implica que $c_i = \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2}$ tq

$$\text{cov}(\bar{Y},b_1) = \text{cov}\left[\frac{\Sigma Y_i}{n}\ ,\ \frac{\Sigma(X_i-\bar{X})Y_i}{\Sigma(X_i-\bar{X})^2}\right] \text{ aplicando} \qquad 1.4.8$$

$$\text{cov}(\bar{Y},b_1) = \left[\frac{1}{n} \times \frac{X_1-\bar{X}}{\Sigma(X_i-\bar{X})^2} + \frac{1}{n} \times \frac{X_2-\bar{X}}{\Sigma(X_i-\bar{X})^2} + \ldots + \frac{1}{n}\,\frac{X_n-\bar{X}}{\Sigma(X_i-\bar{X})^2}\right] \cdot \sigma^2$$

$$= \frac{(X_1-\bar{X}) + (X_2-\bar{X}) + \ldots + (X_n-\bar{X})}{n\,\Sigma(X_i-\bar{X})^2}\,\sigma^2$$

$$= \frac{\Sigma(X_i-\bar{X})}{n\Sigma(X_i-\bar{X})^2}\ \sigma^2 = \frac{0}{n\Sigma(X_i-\bar{X})^2}\,\sigma^2 = 0.\sigma^2 = 0$$

de donde $\text{cov}(\bar{Y},b_1) = 0$, lo que nos implica que $\bar{Y}$ y b_1, son variables aleatorias no correlacionadas. Encontraremos la varianza de $\hat{Y}$ (varianza del valor medio predictado Y).

$\hat{Y}_k$ para un valor específico X_k, de X es

$$V(\hat{Y}_k) = V(\bar{Y} + (X_k - \bar{X})b_1)$$

$$= V(\bar{Y}) + V((X_k - \bar{X}).b_1)$$

$$= V(\bar{Y}) + (X_k - \bar{X})^2 V(b_1)$$

$$V(\hat{Y}_k) = \frac{\sigma^2}{n} + (X_k - \bar{X})^2 . \frac{\sigma^2}{\Sigma(X_i - \bar{X})} \qquad 1.4.9$$

$$\left[v(\bar{Y}) = V\left[\frac{\Sigma Y_i}{n}\right] = \frac{1}{n^2} V(\Sigma Y_i) = \frac{1}{n^2} . \Sigma V(Y_i) = \frac{1}{n^2} \Sigma \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \right]$$

de donde la desviación típica estimada es la raiz cuadrada de - $V(\hat{Y})$ y al sustituir σ^2 por s^2, su estimador, tenemos

$$d.t.e\ (\hat{Y}_k) = s \left\langle \left[\frac{1}{n} + \frac{(X_k - \bar{X})}{(X_i - \bar{X})^2} \right]^{1/2} \right\rangle$$

Esta última ecuación es un mínimo cuando $X_k = \bar{X}$, y aumenta su - valor al ir tomando valores diferentes alejados de él en ambas direcciones, aumentando por lo tanto el error, por lo que no se podrían hacer buenas predicciones para valores alejados de $\bar{X}$.

EJEMPLO. Calcular $d.t.e(\hat{Y}_k)$. con los datos del ejemplo que traemos

$n = 25$, $\Sigma(X - \bar{X})^2 = 7154.42$, $s^2 = 0.7926$, $\bar{X} = 52.60$

$V.e.(\hat{Y}_k)$ = varianza estimada de $\hat{Y}_k$.

$$V.e.(\hat{Y}_k) = s^2 \left[\frac{1}{n} + \frac{(X_k - \bar{X})}{(X - \bar{X})^2} \right]$$

$$= 0.7926 \left[\frac{1}{25} + \frac{(X_k - 52.60)}{7154.42} \right]$$

de qui se puede obtener el mínimo cuando $X_k = \bar{X} = 52.60$

$V.e.(\hat{Y}) = 0.7926 \left[\frac{1}{25} + 0 \right] = 0.031704.$, que al encontrar

$$d.t.e.(\hat{Y}_k) = \sqrt{V.e.(\hat{Y}_k)} = \sqrt{0.031704} = 0.1781$$

para un valor cualquiera $X_k = 28.6$ tenemos

$$V.e.(\hat{Y}_k) = 0.7926 \left[\frac{1}{25} + \frac{(28.6 - 52.6)^2}{7154.42}\right] = 0.095516.$$

siendo su $d.t.e.(\hat{Y}_k) = \sqrt{0.095516} = 0.3091$ que es un valor mucho más grande que el que corresponde a $\bar{X}$., resultando el mismo valor de 0.3091 cuando $X_k = 76.60$ ya que está a la misma - distancia que 28.60, respecto a $\bar{X} = 52.60$

Los límites de confianza al 95% para el valor medio de Y para - un X_k dado, son entonces dados por $\hat{Y}_k \pm (2.069) d.t.e.(\hat{Y})$.

Si solamente una predicción es hecha, $\hat{Y}_k$, digamos para $X = X_k$, entonces la probabilidad que el intervalo calculado contendrá - a este punto el valor medio de Y, es 0.95.

Ejemplos de intervalos al 95% para diferentes valores de X_k, con la respectiva longitud (l) de los intervalos:

a) $X_k = \bar{X} = 52.60$; $\hat{Y}_k = 9.424$, obtenido de

$\hat{Y}_k = 9.424 - 0.0798 (X_k - \bar{X})$; $d.t.e(\hat{Y}_k) = 0.1781$

Así el intervalo es $9.424 \pm (2.069).(d.t.e.(\hat{Y}_k))$

$9.424 \pm (2.069)(0.1781)$

estando Y entre

$9.055511 \leq Y \leq 9.7924889$, Y = Y medio

la longitud l = 0.7369779 0.74

b) Para $X_k = 76.60$ y $X_k = 28.60$ que tienen el mismo valor de - $d.t.e.(\hat{Y}_k) = 0.3091$

i) si $X_k = 76.60$, $\hat{Y}_k = 9.424 - 0.0798(76.60 - 52.60) = 7.5088$

donde $7.5088 \pm (2.069)(0.3091)$, asi Y está entre

$6.8685761 \leq Y \leq 8.1476319$

siendo $l = 1.279$

ii) si $X_k = 28.60$; $\hat{Y}_k = 9.424 - 0.0798(28.60-52.60) = 11.3392$

de donde $11.3292 \pm (2.069)(0.3091)$, así Y está entre

$10.699673 \leq Y \leq 11.978727$

así $l = 1.279$

Como puede apreciarse para este caso, para los valores simétricos, sobre la recta ajustada, la longitud es la misma, entre esos límites y mayor que el que corresponde a $\bar{X}$, lo que se puede observar en la figura 1.5, que se forman hiperbólas.

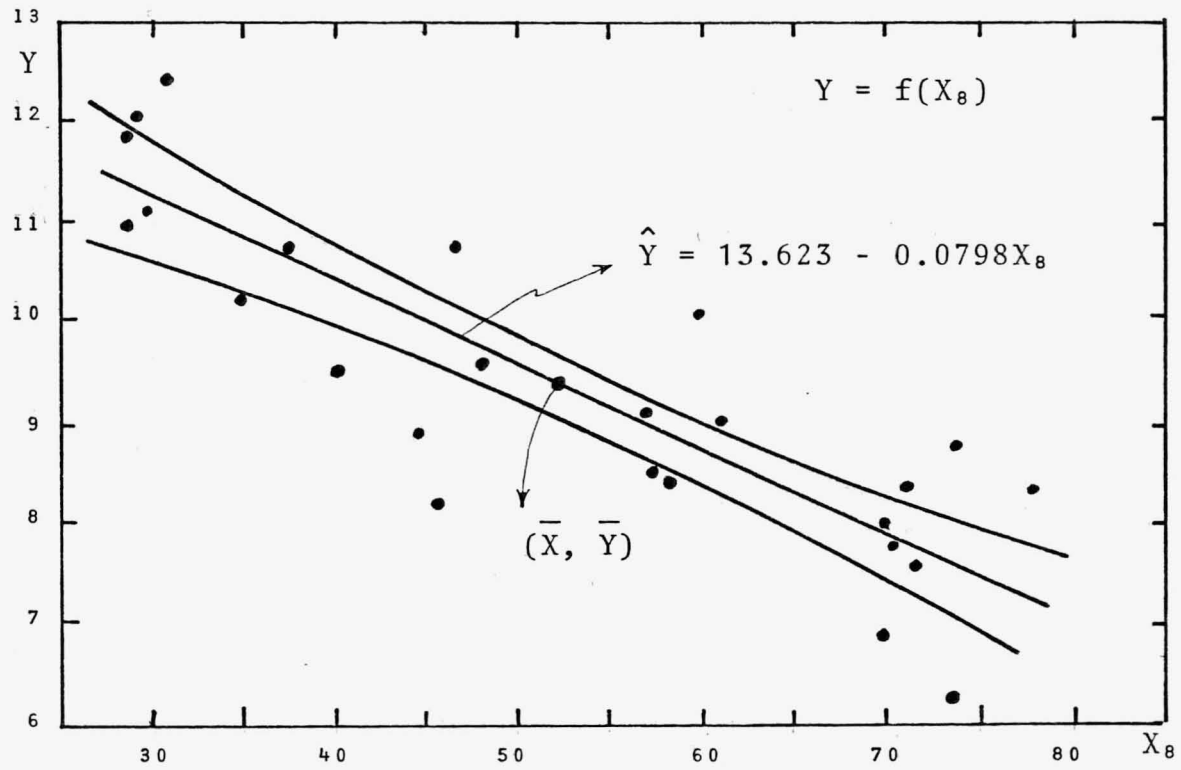


Figura 1.5

La varianza y error estandar pueden aplicarse para un valor medio proyectado de Y para un X_k dado. Puesto que el valor observado de Y varía acerca del valor medio verdadero con varianza -

σ^2 (que es independiente de $V(\hat{Y})$), así para un valor proyectado de una observación individual estará dada por $\hat{Y}$, pero tendrá varianza.

$$\sigma^2 + V(\hat{Y}) = \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right\} \qquad 1.4.10$$

si el valor de σ^2 es desconocido, se utiliza su estimador s^2, así

$$s^2 + V(\hat{Y}) = s^2 \left\{ 1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right\}$$

pudiéndose, ahora encontrar un intervalo de confianza al 95%, para una nueva observación la cual será centrada en $\hat{Y}_k$ y cuya longitud dependerá de la nueva varianza.

$$\hat{Y}_k \pm t(v, 0.975). \left\{ 1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right\}^{1/2} s.$$

donde v es el número de grados de libertad basados en s^2 (es igual a n - 2)

Un intervalo de confianza para la media de q nuevas observaciones cerda de $\hat{Y}_k$, es obtenido similarmente como sigue:

Sea $\bar{Y}_q$ la media de las q observaciones futuras de X_k (donde q podría ser igual a uno como en el caso anterior).

Entonces:

$$\bar{Y}_q \sim N(\beta_o + \beta_1 X_k, \sigma^2 q)$$

$$\hat{Y}_k \sim N(\beta_o + \beta_1 X_k, V(\hat{Y}_k))$$

tal que

$$\bar{Y}_q - \bar{Y}_k \sim N(0,\ \sigma_q^2 + V(\hat{Y}_k))$$

y $[(\bar{Y}_q - \hat{Y}_k)/d.t.e(\bar{Y}_q - \hat{Y}_k)]$ es distribuida como una variable t(v), donde v es el número de grados de libertad en el cual está basado s^2, estimador de σ^2.

Ahora

$$\sigma_q^2 = V(\bar{Y}_q) = V\left(\frac{\sum_{i=1}^{q} Y_i}{q}\right) = \frac{1}{q^2} \sum_{i=1}^{q} V(Y_i) = \frac{1}{q^2} \sum_{i=1}^{q} \sigma^2$$

$$= \frac{1}{q^2} \cdot q.\sigma^2 = \frac{\sigma^2}{q} \quad \therefore \quad \sigma_q^2 = V(\bar{Y}_q) = \frac{\sigma^2}{q}$$

La varianza de $\bar{Y}_q - \hat{Y}_k$ es $\sigma_q^2 + V(\hat{Y}_k)$

$$\sigma_q^2 + V(\hat{Y}_k) = \frac{\sigma^2}{q} + \frac{\sigma^2}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\ \sigma^2$$

$$= \sigma^2\left(\frac{1}{q} + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)$$

siendo la desviación típica estimada igual a

$s\left(\frac{1}{q} + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)^{1/2}$, donde s^2 es estimador de σ^2

Entonces

$$\text{prob}\left\{|\bar{Y}_q - \hat{Y}_k| \leq t(v,0.975).\left[s^2\left(\frac{1}{q} + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)\right]^{1/2}\right\} = 0.95$$

de donde podemos obtener los límites de confianza al 95% para $\bar{Y}_q$ alrededor de $\hat{Y}_k$ de

$$\hat{Y}_k \pm t(v,0.975).\left[\frac{1}{q} + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]^{1/2}.s$$

Estos límites nos dan un rango amplio para la media de Y para un X_k dado, dentro de los cuales se espera que el 95% de las -

futuras observaciones de X_k (o futuras medias de q observaciones de X_k como el caso puede ser). son esperadas que esten.

PRUEBA F PARA SIGNIFICACION DE REGRESION.

Los Y_i son variables aleatorias, cualquier función de ellas es también una variable aleatoria, dos funciones particulares son MS_R, media de cuadrados debida a la regresión y s^2, la media - de cuadrados debido a la variación residual, las cuales surgen en el analisis de varianza. Esas funciones entonces tienen su propia distribución, con media, varianza y momentos, donde la media de esas variables vienen dadas por

$$E(MS_R) = \sigma^2 + \beta_1^2 \Sigma (X - \bar{X})^2$$

$$E(s^2) = \sigma^2 \qquad 1.4.11$$

donde, si z es una variable aleatoria, E(z) denota el valor medio o valor esperado. Supongamos que los errores ε_i son variables independientes con distribucion $N(0,\sigma^2)$. Entonces puede - ser demostrado que si $\beta_1 = 0$, la variable MS_R multiplicado por sus grados de libertad (aqui uno) y dividido por σ^2 sigue una - distribución X^2 con el mismo número de grados de libertad (1). En resumen, $(n-2)\ s^2/\sigma^2$ tiene una distribución X^2, con (n-2) - grados de libertad.

La razón

$$F = \frac{MS_R}{s^2} \qquad 1.4.12$$

tiene una F distribución con 1 y n-2 grados de libertad, a condición que $\beta_1 = 0$. Por lo que puede ser utilizado como prueba de $\beta_1 = 0$. comparamos el valor de $F = \frac{MS_R}{s^2}$ con el $100(1-\alpha)\%$ -

del valor tabulado por la distribución F(1,n-2). Para determinar si β_1 puede ser diferente de cero en base a los datos obtenidos.

EJEMPLO

Con los datos del ejemplo continuado.

De la tabla 1.3 podemos ver que $MS_R = 45.59$ y $s^2 = 0.7926$ por lo que $F = \frac{45.59}{0.7926} = 57.52$, donde el porcentaje de puntos al 95% para $F(1,23,0.95) = 4.28$, por lo que el F calculado excede el valor crítico, así $F = 57.52 > 4.28$, por lo que rechazamos la hipótesis H_o: $\beta_1 = 0$, corriendo el riesgo de menos del 5% de estar equivocados.

OBSERVACION:

En el caso de ajustar una línea recta, esta prueba F para regresión es el mismo que la prueba t, para el $\beta_1 = 0$, dado anteriormente. Esto es porque,

$$\frac{MS_R}{s^2} = \frac{b_1 \Sigma(X-\bar{X})(Y-\bar{Y})}{s^2} = \frac{b_1^2 \Sigma(X-\bar{X})}{s^2} \quad \text{por 1.3.3 y 1.4.12}$$

$$= \left[\frac{b_1\{\Sigma(X_i-\bar{X})^2\}^{1/2}}{s}\right]^2 = t^2 \quad \text{por 1.4.3}$$

donde la variable F(1,n-2) es el cuadrado de la variable t, con n-2 grados de libertad, da exacatamente el mismo resultado, cuando se trata de un valor individual. Así el valor de F=57.52 y el valor de t^2 es $(-7.60)^2 = 57.56$, donde pudo haber un error de redondeo, de lo contrario sería igual al valor de F.

PORCENTAJE DE VARIACION EXPLICADA

Para ello definimos R^2 = (sc debido a la regresión)/(total sc, corregido para la media). Entonces R^2 mide la " proporción de variación total acerca de la media $\bar{Y}$ explicada por la regre--- sión ". Implica que la respuesta es cero, cuando todas las variables independientes son cero, Esta es una suposición muy - fuerte la cual es injustificada. En un modelo de línea recta.

$Y = \beta_o + \beta_1 X + \varepsilon$, la omisión de β_o implica que la línea pasa a traves de $(X,Y) = (0,0)$., esto es que su intercepto es cero ó $\beta_o = 0$ cuando $X = 0$. Puede desaparecer también siempre que - sea posible centrar los datos, como se hará en capítulo 5, pero esto es enteramente diferente de colocar $\beta_o = 0$.

Por ejemplo si escribimos

$$Y = \beta_o + \beta_1 X + \varepsilon \qquad \text{y restando } \bar{Y} \text{ tenemos}$$

$$Y - \bar{Y} = \beta_o + \beta_1 X + \varepsilon - \bar{Y}$$

$$Y - \bar{Y} = \beta_o + \beta_1 X + \varepsilon - \bar{Y} + \beta_1 \bar{X} - \beta_1 \bar{X}$$

sumando y restando $\beta_1 \bar{X}$, agrupando términos tenemos

$$Y - \bar{Y} = \beta_o + \beta_1 \bar{X} - \bar{Y} + \beta_1 X - \beta_1 \bar{X} + \varepsilon$$

$$Y - \bar{Y} = (\beta_o + \beta_1 \bar{X} - \bar{Y}) + \beta_1 (X - \bar{X}) + \varepsilon$$

ó $$y = \beta_o' + \beta_1 x + \varepsilon$$

donde $y = Y - \bar{Y}$, $\beta_o' = \beta_o + \beta_1 \bar{X} - \bar{Y}$ y $x = X - \bar{X}$; entonces los mejores estimadores de β_o' y β_1' por mínimos cuadrados estan dados por

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

que es idéntica a 1.2.9, donde

$b_o' = \bar{Y} - b_1\bar{X} = 0$, por $\bar{X} = \bar{Y} = 0$

entonces podemos escribir nuestro modelo

$Y - \bar{Y} = \beta_1(X - \bar{X}) + \varepsilon$, ya que $\beta_o' = 0$.

EJEMPLO (continuado)

de la tabla 1.3

$R^2 = \frac{45.59}{63.82} \times 100 = 71.44\%$

Así, la ecuación obtenida explica el 71.44% de la variación total.

1.5 FUERA DE AJUSTE Y ERROR PURO

Tenemos que recordar que la línea de regresión ajustada está basada en un modelo supuesto, el problema se reduce a revisar si el modelo es ó nó correcto, entonces se puede examinar las consecuencias de un modelo incorrecto, para ello llamaremos $e_i = Y_i - \hat{Y}_i$, que es el residuo de X_i, se ha probado anteriormente que $\Sigma e_i = 0$. los residuos contienen información disponible sobre la forma en las cuales el modelo ajustado falla propiamente en la explicación de la variación de los valores observados de Y.

Sea $n_i = E(Y_i)$, el valor dado por el modelo verdadero, para un $X = X_i$. Entonces podemos escribir.

$$Y_i - \hat{Y}_i = (Y_i - \hat{Y}_i) - E(Y_i - \hat{Y}_i) + E(Y_i - \hat{Y}_i)$$

$$= (Y_i - \hat{Y}_i) - (E(Y_i) - E(\hat{Y}_i)) + (E(Y_i) - E(\hat{Y}_i))$$

$$= (Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i)) + (n_i - E(\hat{Y}_i))$$

$$= \{(Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i))\} + (n_i - E(\hat{Y}_i))$$

$$Y_i - \hat{Y}_i = q_i + B_i$$

donde

$$q_i = (Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i))$$
$$B_i = n_i - E(\hat{Y}_i)$$

La cantidad B_i es el error de sesgo de $X = X_i$, si el módulo es correcto, entonces $E(\hat{Y}_i) = n_i$, por lo que

$B_i = n_i - E(\hat{Y}_i) = n_i - n_i = 0 \quad \therefore \quad B_i = 0.$

Si el modelo es incorrecto, entonces $B_i \neq 0$, ya que $E(\hat{Y}_i) \neq n_i$, pero tiene un valor que depende de X_i.

La cantidad q_i es una variable aleatoria que tiene media cero.

$q_i = (Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i))$, aplicando esperanza

$$E(q_i) = E\left[(Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i))\right]$$
$$= E(Y_i - \hat{Y}_i) - E(n_i - E(\hat{Y}_i))$$
$$= E(Y_i) - E(\hat{Y}_i) - (n_i - E(\hat{Y}_i))$$
$$= n_i - E(\hat{Y}_i) - n_i + E(\hat{Y}_i)$$

$\therefore \quad E(q_i) = 0$

siendo verdadero este valor, sea correcto o nó el modelo, los q_i son corelacionados y la cantidad $q_1^2 + q_2^2 + \ldots + q_n^2$ tiene - valor esperado $(n-2)\,\sigma^2$., donde $V(Y_i) = V(\varepsilon_i) = \sigma^2$, es el error de la varianza.

Pureba

$$\Sigma q_i^2 = \Sigma\left[(Y_i - \hat{Y}_i) - (n_i - E(\hat{Y}_i))\right]^2$$
$$= \Sigma\left[(Y_i - \hat{Y}_i) - E(Y_i - \hat{Y}_i)\right]^2$$
$$\Sigma q_i^2 = \Sigma\{(Y_i - \hat{Y}_i)^2 - 2(Y_i - \hat{Y}_i).E(Y_i - \hat{Y}_i) + \left[E(Y_i - \hat{Y}_i)\right]^2\}$$

Aplicando esperanza y propiedades tenemos

$$E(\Sigma q_i^2) = E(\Sigma\{(Y_i - \hat{Y}_i)^2 - 2(Y_i - \hat{Y}_i).E(Y_i - \hat{Y}_i) + \left[E(Y_i - \hat{Y}_i)\right]^2\}$$
$$= \Sigma\{E(Y_i - \hat{Y}_i)^2 - 2E(Y_i - \hat{Y}_i).E(Y_i - \hat{Y}_i) + \left[E(Y_i - \hat{Y}_i)\right]^2\}$$
$$= \Sigma\{E(Y_i - \hat{Y}_i)^2 - 2\left[E(Y_i - \hat{Y}_i)\right]^2 + \left[E(Y_i - \hat{Y}_i)\right]^2\}$$

$$= \Sigma\{E(Y_i-\hat{Y}_i)^2 - [E(Y_i-\hat{Y}_i)]^2\}$$

$$= \Sigma V(Y_i-\hat{Y}_i) \quad \text{por definición de varianza}$$

$$= \Sigma(V(Y_i) - V(\hat{Y}_i))$$

$$= \Sigma\left[\sigma^2 - \frac{\sigma^2}{n} - \frac{(X_i-\bar{X})^2\sigma^2}{\Sigma(X_i-\bar{X})^2}\right], \quad V(Y_i) = \sigma^2, \quad \text{y por 1.4.9}$$

$$= \Sigma\sigma^2 - \frac{1}{n}\Sigma\sigma^2 - \sigma^2 \cdot \frac{\Sigma(X_i-\bar{X})^2}{\Sigma(X_i-\bar{X})^2}$$

$$= n\sigma^2 - \frac{n\sigma^2}{n} - \sigma^2$$

$$= n\sigma^2 - \sigma^2 - \sigma^2$$

$$= (n - 1 - 1)\,\sigma^2$$

$$= (n - 2)\,\sigma^2$$

de donde $E(\Sigma q_i^2) = (n - 2)\sigma^2$ 1.5.1

De esto puede demostrarse además que la media de los cuadrados de los residuos, es decir el valor

$\frac{1}{n-2}\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\}$ tiene como media σ^2, si el modelo postulado es correcto. Prueba.

Se sabe que $Y_i-\hat{Y}_i = q_i + B_i$, si el modelo es correcto $B_i = 0$, por lo tanto $Y_i-\hat{Y}_i = q_i$, así

$$(Y_i-\hat{Y}_i)^2 = q_i^2$$

$$\Sigma(Y_i-\hat{Y}_i)^2 = \Sigma q_i^2$$

$$\frac{\Sigma(Y_i-\hat{Y}_i)^2}{n-2} = \frac{\Sigma q_i^2}{n-2}$$

Aplicando esperanza tenemos.

$$E\left[\frac{\Sigma(Y_i-\hat{Y}_i|^2}{n-2}\right] = E\left[\frac{\Sigma q_i^2}{n-2}\right] = \frac{1}{n-2}E(\Sigma q_i^2) = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

así se tiene lo afirmado, habiendo aplicado 1.5.1

Ahora si el modelo es incorrecto, entonces su esperanza viene dada por $\sigma^2 + \frac{\Sigma B_i^2}{n-2}$, por lo que $B_i \neq 0$,

PRUEBA

Se tiene que $Y_i - \hat{Y}_i = q_i + B_i$

elevando al cuadrado y aplicando sumatoria tenemos

$$\Sigma(Y_i-\hat{Y}_i)^2 = \Sigma(q_i+B_i)^2 = \Sigma(q_i^2 + 2q_iB_i + B_i^2) = \Sigma q_i^2 + 2\Sigma q_iB_i + \Sigma B_i^2$$

dividiendo por (n-2), así

$$\frac{\Sigma(Y_i-\hat{Y}_i)^2}{n-2} = \frac{\Sigma q_i^2}{n-2} + \frac{2\Sigma q_iB_i}{n-2} + \frac{\Sigma B_i^2}{n-2}$$

Aplicando la esperanza y propiedades tenemos

$$E\left[\frac{\Sigma(Y_i-\hat{Y}_i)^2}{n-2}\right] = \frac{E(\Sigma q_i^2)}{n-2} + 2\frac{\Sigma B_i E(q_i)}{n-2} + \frac{\Sigma B_i^2}{n-2}$$

sabemos que $E(q_i) = E(Y_i-\hat{Y}_i) = 0$. y

por 1.5.1

$$= \frac{(n-2)\sigma^2}{(n-2)} + \frac{\Sigma B_i^2}{n-2}$$

$$= \sigma^2 + \frac{\Sigma B_i^2}{n-2}$$

de donde $E\left[\frac{\Sigma(Y_i-\hat{Y}_i)^2}{n-2}\right] = \sigma^2 + \frac{\Sigma B_i^2}{n-2}$

Si el modelo es correcto, los residuos son (correlacionados) -- desviaciones aleatorias q_i y la media de los cuadrados de los - residuos puede ser usado como un estimador de la varianza. Sin embargo si el modelo no es correcto, esto es $B_i \neq 0$. Entonces - los residuos contienen a q_i y B_i, siendo error de varianza y - sesgo de error para los residuos, debiendo tener cuidado que no sea grande la media de cuadrados de residuos debido al sesgo.

Con el caso simple de ajustar una línea recta, el error de sesgo puede ser detectado en un diagrama de los datos, pero cuando el modelo es más complicado o implica más variables no es posible hacer eso. Si se dispone de un estimador de σ^2 (obtenido de experiencias anteriores) podemos ver si o no la media de cuadrados es significante, más grande que el estimador anterior, si - es significativamente grande, decimos que está fuera de ajuste y deberíamos reconsiderar el modelo, el cual es inadecuado para este caso. Si no hay estimador previo de σ^2 disponible, pero repetidas medidas de Y (*i.e.* dos o más), han sido hechas al mismo - valor de X, podemos usar esas repeticiones para obtener un estimador de σ^2.

Tal estimador decímos representar " el error puro " porque, si el valor de X es idéntico para dos observaciones, solamente la variación aleatoria puede influenciar los resultados y proveer diferencia entre ellos. Tales diferencias usualmente proveerán un estimador de σ^2, el cual es mucho más confiable que podemos obtener de alguna otra fuente.

El error puro del estimador de σ^2, es encontrado como sigue (la misma fórmula es aplicada cuando hay más de una variable inde-- pendiente). Supongamos

$Y_{11}, Y_{21}, \ldots, Y_{1n}$ son n_1 observaciones repetidas de X_1

$Y_{21}, Y_{22}, \ldots, Y_{2n}$ son n_2 observaciones repetidas de X_2

. . .

. . .

. . .

$Y_{k1}, Y_{k2}, \ldots, Y_{kn}$ son n_k observaciones repetidas de X_k.

La contribución al error puro a la suma de cuadrados de X_1 lecturas es.

entonces

$$\sum_{u=1}^{n_1}(Y_{1u}-\bar{Y}_1)^2 = \sum_{u=1}^{n_1}(Y_{1u}^2-2Y_{1u}\bar{Y}_1+(\bar{Y}_1)^2) = \sum_{u=1}^{n_1}Y_{1u}^2-2\bar{Y}_1\sum_{i=1}^{n_1}Y_{1u}+\Sigma(\bar{Y})^2$$

$$= \sum_{u=1}^{n_1}Y_{1u}^2-2\bar{Y}_1\, n_1\bar{Y}_1+n_1(\bar{Y}_1)^2 = \sum_{u=1}^{n_1}Y_{1u}^2 - n_1(\bar{Y}_1)^2$$

de donde

$$\sum_{u=1}^{n_1}(Y_{1u} - \bar{Y}_1)^2 = \sum_{u=1}^{n_1}Y_{1u}^2 - n_1(\bar{Y}_1)^2. \qquad 1.5.2$$

con $\bar{Y}_1 = (Y_{11} + Y_{12} + \ldots + Y_1n_1)/n_1$. Esta suma de cuadrados - tiene (n_1-1) grados de libertad. Similares cantidades pueden - ser encontrados para los otros conjuntos de yes ó el total sc - (error puro) $= \sum_{i=1}^{k}\sum_{u=1}^{n_i}(Y_{iu}-\hat{Y}_i)^2$, con el total de grados de li-- bertad $= \sum_{i=1}^{k}(n_i-1) = \sum_{i=1}^{k}n_i-k = n_e$, decimos entonces que la media de los cuadrados para el error puro es

$$S_e^2 = \frac{\sum_{i=1}^{k}\sum_{u=1}^{n_i}(Y_{iu}-\bar{Y}_i)^2}{\sum_{i=1}^{n}n_i - k} \qquad 1.5.3$$

y es un estimador de σ^2. Esta cantidad es el total de la suma - de cuadrados " dentro de repeticiones " dividido por el total - de grados de libertad.

(Nota: si hay solamente dos observaciones Y_{11}, Y_{12} para el punto X_i, entonces

$\sum_{u=1}^{2}(Y_{iu}-\bar{Y}_i)^2 = \frac{1}{2}(Y_{i1}-Y_{i2})^2$. Esta suma tiene 1 grado de libertad.

PRUEBA

$$\sum_{u=1}^{2}(Y_{iu}-\bar{Y}_i)^2 = (Y_{i1}-\bar{Y}_i)^2 + (Y_{i2}-\bar{Y}_i)^2$$

$$= (Y_{i_1} - \frac{Y_{i_1}+Y_{i_2}}{2})^2 + (Y_{i_2} - \frac{Y_{i_1}+Y_{i_2}}{2})^2$$

$$= (\frac{2Y_{i_1} - Y_{i_1} - Y_{i_2}}{2})^2 + (\frac{2Y_{i_2} - Y_{i_1} - Y_{i_2}}{2})^2$$

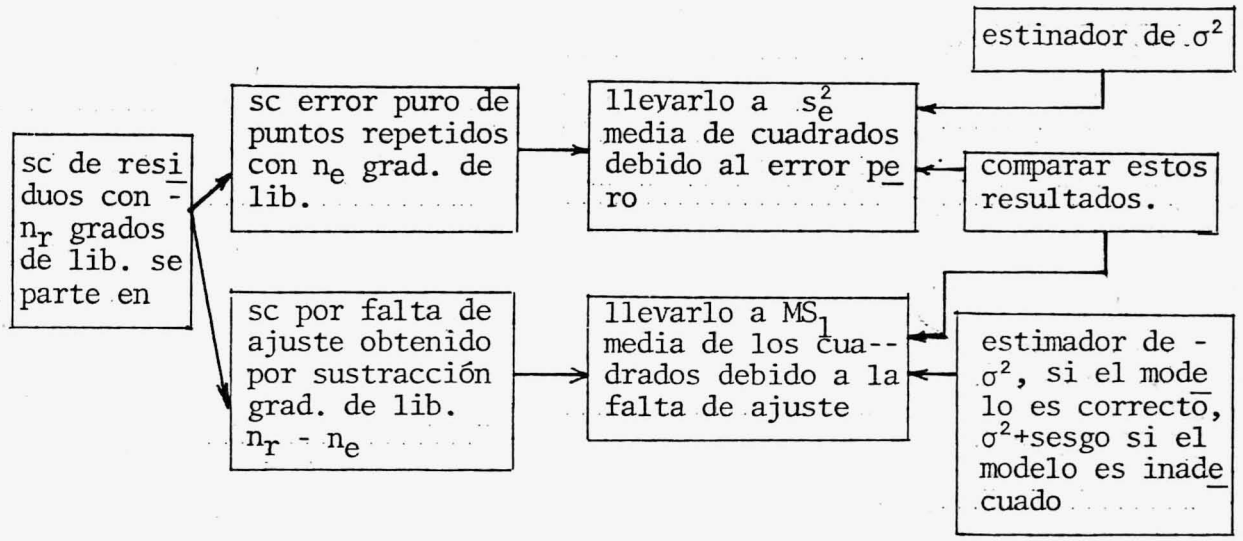
$$= (\frac{Y_{i_1} - Y_{i_2}}{2})^2 + (\frac{Y_{i_2} - Y_{i_1}}{2})^2$$

$$= \frac{(Y_{i_1} - Y_{i_2})^2}{4} + \frac{(Y_{i_1} - Y_{i_2})^2}{4}$$

$$= \frac{1}{2}(Y_{i_1} - Y_{i_2})^2$$

$\therefore \sum_{u=1}^{2} (Y_{iu} - \bar{Y}_i)^2 = \frac{1}{2}(Y_{i_1} - Y_{i_2})^2$, 1.5.4 lo que hace fácil el cálculo, para dos observaciones.

La suma de cuadrados del error puro, puede ser introducido en una tabla de analisis de varianza como se muestra en la siguiente figura



donde el procedimiento usual es comparar la razón $F = \frac{MS_1}{s_e^2}$, con el valor de 100 $(1 - \alpha)$ % de una distribución F con $(n_r - n_e)$ y n_e grados de libertad. Si la razón es

1. Significante. Esto indica que el modelo parece ser inadecuado, la prueba debería ser hecha para descubrir donde y como ocurre lo inadecuado.

2. No significante. Esto indica que allí parece ser no razonable dudar del modelo adecuado y ambos, media de cuadrados de error puro y falta de ajuste pueden ser usados como estimadores de σ^2.

Un nuevo estimador de σ^2 puede ser obtenido combinando el error puro y la suma de cuadrados por falta de ajuste dentro de la suma de residuos y dividiendola por el número de grados de libertad, es decir.

$$s^2 = \frac{\text{sc error puro + sc por falta de ajuste}}{n_r} = \frac{\text{sc residuos}}{n_r}$$

EJEMPLO

Para los siguientes datos se ilustrará la falta de ajuste y el error puro.

OBSERV.	X	Y	OBSERV.	X	Y	OBSERV.	X	Y
1	1.3	2.3	9	3.7	1.7	17	5.3	3.5
2	1.3	1.8	10	4.0	2.8	18	5.3	2.8
3	2.0	2.8	11	4.0	2.8	19	5.3	2.1
4	2.0	1.5	12	4.0	2.2	20	5.7	3.4
5	2.7	2.2	13	4.7	5.4	21	6.0	3.2
6	3.3	3.8	14	4.7	3.2	22	6.0	3.0
7	3.3	1.8	15	4.7	1.9	23	6.3	3.0
8	3.7	3.7	16	5.0	1.8	24	6.7	5.9

cuya línea de regresión se encuentra así

$\Sigma Y = 686$, $\Sigma X = 101$, $\Sigma XY = 307.41$, $\Sigma X^2 = 480.44$, $n = 24$.

$$b_1 = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - (\Sigma X)^2/n} = \frac{307.41 - \frac{(101)(68.6)}{24}}{480.44 - (101)^2/24} = 0.3379 \qquad 0.338$$

cálculo de la recta

$\hat{Y} = \overline{Y} + b_1(X - \overline{X})$, con $\overline{Y} = \frac{68.6}{24} = 2.858$, $\overline{X} = \frac{101}{24} = 4.208$

$\hat{Y} = 2.858 + 0.338(X - 4.208)$, por lo que la recta es

$\hat{Y} = 1.436 + 0.338\,X$.

Ahora se construirá la tabla de analisis de varianza, donde se utilizará $\Sigma Y^2 = 223.6$

TABLA ANALISIS DE VARIANZA

FUENTE	Grad. de Lib.	sc	mc	Razón F
Total (corregido)	23	27.518		
Regresión	1	6.326	6.326	6.569 significante al nivel α=0.05
Residuos	22	21.192	s^2=0.963	

cálculo por filas

a) $\sum_{i=1}^{24} (Y_i - \overline{Y})^2 = \sum_{i=1}^{24} Y_i^2 - \frac{(\sum_{i=1}^{24} Y_i)^2}{24} = 223.6 - \frac{(68.6)^2}{24} = 27.518 = sc$

b) $b_1 (\Sigma X_i Y_i - \frac{\Sigma X \Sigma Y}{n}) = 0.338 \; (307.41 - \frac{(101)(68.6)}{24}) = 6.326$

$mc = \frac{6.326}{1} = 6.326$, $F = \frac{6.326}{0.963} = 6.569$

$F(1,22,0.95) = 4.30$

c) sc residuos = a) - b) = 27.518 - 6.326 = 21.192

$s^2 = mc = \frac{21.192}{22} = 0.963$

Ahora encontraremos el error puro y por lo tanto la falta de - ajuste

1. La sc del erro puro de datos repetidos de X = 1.3 es

$\frac{1}{2}(2.3 - 1.8) = 0.125$, con 1 grado de libertad, por 1.5.4

2. La sc del error puro de datos repetidos de X = 4.7 es

$(5.4)^2 + (3.2)^2 + (1.9)^2 - 3(\frac{5.4 + 3.2 + 1.9}{3}) = 6.26$, por 1.5.3

con 2 grados de libertad.

De manera similar se calculan para los otros datos repetidos y se obtiene la siguiente tabla.

NIVEL DE X	$\sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2$	grados de libertad
1.3	0.125	1
2.0	0.845	1
3.3	2.000	1
3.7	2.000	1
4.0	0.240	2
4.7	6.260	2
5.3	0.980	2
6.0	0.020	1
TOTALES	$\sum_{i=1}^{k} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i) = 12.470$	$n_e = 11$

Vamos ahora a reescribir el cuadro de analisis de varianza para encontrar MS_1 y s_e^2

TABLA ANAVA - (mostrando la falta de ajuste)

FUENTE	Grados de Lib.	sc	mc	razón F
TOTAL	23	27.518		
REGRESION	1	6.326	6.326	6.569 significante $\alpha = 0.05$
RESIDUOS	22	21.192	s^2=0.963	
ERROR PURO	11	12.470	s_e^2=1.134	
FALTA DE AJUSTE	11	8.722	MS_1=0.793	0.699 no sifnificante.

donde los datos agregados de la nueva tabla son:

$$s_e^2 = \frac{\sum_{i=1}^{k} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)}{n_e} = \frac{12.470}{11} = 1.134;$$

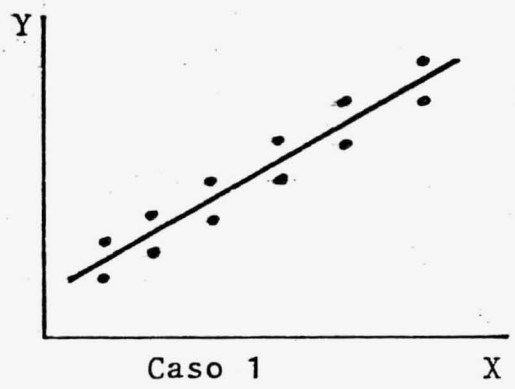
para la falta de ajuste = sc residuos - sc error puro

$$= 21.192 - 12.470 = 8.722$$

asi $MS_1 = \frac{\text{sc falta de ajuste}}{n_r - n_e} = \frac{8.722}{22-11} = 0.7929 \quad 0.793$

la razon $F = \frac{MS_1}{s_e^2} = \frac{0.793}{1.134} = 0.699$, siendo este no significante, además es menor que la unidad. Entonces en base a esta prueba - de mínimo, no tenemos razón de dudar adecuadamente de nuestro - modelo y podemos usar $s^2 = 0.963$ como un estimador de σ^2.

Los siguientes diagramas ilustran algunas situaciones que pue-- den presentarse, cuando una línea recta es encontrada de los datos y la acción consecuente ha sido tomada.



Caso 1

Caso 1

1) Ensayando $Y = \beta_o + \beta_1 X + \varepsilon$
2) No hay falta de ajuste
3) Regresión lineal significante.
4) Use modelo $\hat{Y} = b_o + b_1 X$

donde los datos agregados de la nueva tabla son:

$$s_e^2 = \frac{\sum_{i=1}^{k}\sum_{u=1}^{n_i}(Y_{iu}-\bar{Y}_i)}{n_e} = \frac{12.470}{11} = 1.134;$$

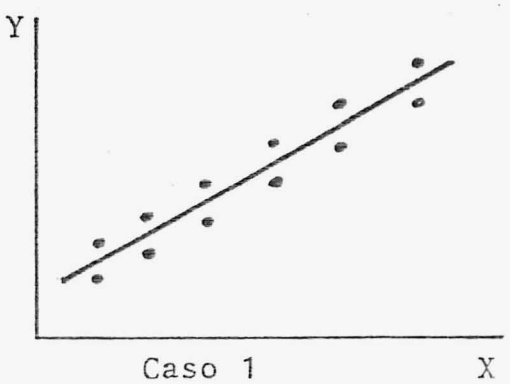
para la falta de ajuste = sc residuos - sc error puro

$$= 21.192 - 12.470 = 8.722$$

asi $MS_1 = \frac{\text{sc falta de ajuste}}{n_r - n_e} = \frac{8.722}{22-11} = 0.7929 \quad 0.793$

la razon $F = \frac{MS_1}{s_e^2} = \frac{0.793}{1.134} = 0.699$, siendo este no significante, además es menor que la unidad. Entonces en base a esta prueba - de mínimo, no tenemos razón de dudar adecuadamente de nuestro - modelo y podemos usar $s^2 = 0.963$ como un estimador de σ^2.

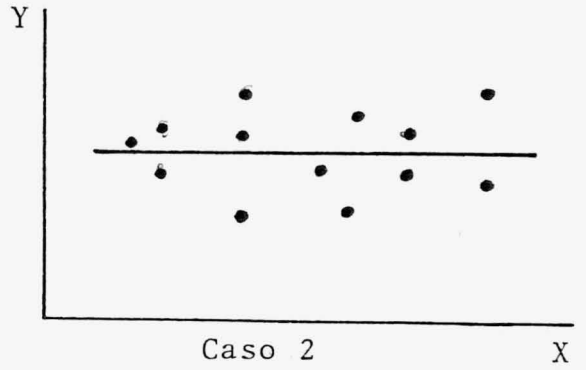
Los siguientes diagramas ilustran algunas situaciones que pue-- den presentarse, cuando una línea recta es encontrada de los da- tos y la acción consecuente ha sido tomada.



Caso 1

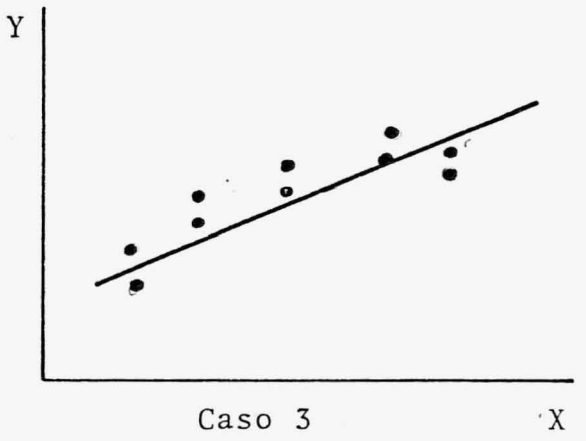
Caso 1

1) Ensayando $Y = \beta_o + \beta_1 X + \varepsilon$
2) No hay falta de ajuste
3) Regresión lineal significan- te.
4) Use modelo $\hat{Y} = b_o + b_1 X$



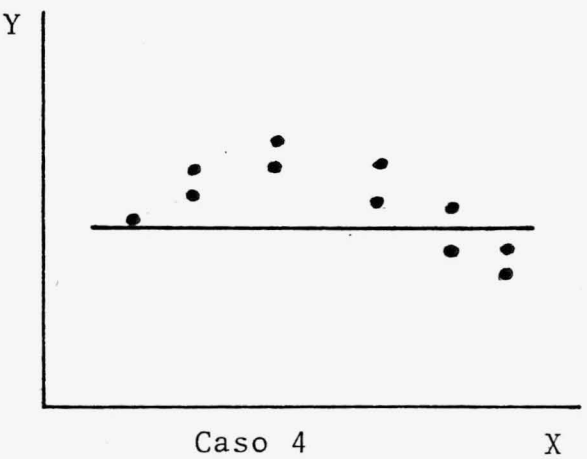
Caso 2

1) Ensayo $Y = \beta_o + \beta_1 X + \varepsilon$
2) No hay falta de ajuste
3) Regresión lineal no significante
4) Use el modelo $\hat{Y} = \overline{Y}$



Caso 3

1) Ensayo $Y = \beta_o + \beta_1 X + \varepsilon$
2) Significante la falta de ajuste
3) Regresión lineal significante.
4) Use el modelo $Y = \beta_o + \beta_1 X + \beta_{11} X^2 + \varepsilon$



Caso 4

1) Ensayo $Y = \beta_o + \beta_1 X + \varepsilon$
2) Significante la falta de ajuste
3) Regresión lineal no significante.
4) Ensaye el modelo
   $Y = \beta_o + \beta_1 X + \beta_{11} X^2 + \varepsilon$

1.6 CORRELACION ENTRE X e Y.

Si X e Y son ambas variables aleatorias siguiendo una distribución (desconocida) bivariante, entonces podemos definir el coeficiente de correlación entre X e Y como sigue:

$$\rho_{XY} = \frac{\text{covarianza } (X,Y)}{\{V(X).V(Y)\}^{1/2}} \qquad 1.6.1$$

En donde, si f(X,Y) es una distribución de probabilidad contínua de X e Y:

$$\text{covar } (X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{Y-E(Y)\}\{X - E(X)\}f(X,Y)\, d\, x\, d_Y$$

$$\text{var } (Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (Y - E(Y))^2 f(X,Y)\, d_Y d_X, \quad \text{donde}$$

$$E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Y\, f(X,Y) d_Y d_X \qquad 1.6.2$$

de manera similar se define V(X), E(X). (Si las distribuciones son discretas " $\int$ " se cambia por " Σ ").

Esta cantidad ρ_{XY} es una medida de asociación de las variables X e Y, siendo su valor $-1 \leq \rho_{XY} \leq 1.$, así si $\rho_{XY} = 1$, X e Y estan correladas perfectamente por una línea recta de pendiente positiva; si $\rho_{XY} = 0$, entonces no hay correlación de línea recta para X e Y; si $\rho_{XY} = -1$, X e Y estan correlacionadas correctamente en una línea recta de pendiente negativa.

En una muestra de tamaño n, (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n). son evaluados en la distribución, la cantidad

$$r_{XY} = \frac{\sum_{i=1}^{n} (X_i-\bar{X})(Y_i-\bar{Y})}{\{\sum_{i=1}^{n} (X_i-\bar{X})^2\}^{1\ 2}\{\sum_{i=1}^{n} (Y_i-\bar{Y})^2\}^{1/2}} \qquad 1.6.3$$

llamando coeficiente de correlación simple entre X e Y y es un estimador de ρ_{XY} y provee una medida empirica de la asociación entre X e Y.

(Si el factor $\frac{1}{n-1}$ es colocado antes de todas las sumas de r_{XY} se convierte en ρ_{XY} con varianza y covarianza, reemplazados por

valores muestrales). Cuando los X_i e Y_i, $i = 1,\ldots, n.$, son todos constantes, más bien que valores de una muestra de alguna distribución, deben ser considerados como datos problacionales entonces, r_{XY} puede ser utilizado como una medida de su asociación, donde el conjunto de valores (X_i, Y_i), $i = 1,2,.. n.$, puede ser considerado como una distribución finita, γ_{XY}, es efectivamente un parámetro poblacional, más que un valor muestral, esto es, $r_{XY} = \rho_{XY}$ en este caso.

Si γ es una variable aleatoria y $X_1, X_2, \ldots, X_n$ representa los valores de una distribución X finita. El coeficiente de correlación ρ_{XY} es definido por 1.6.1 y la expresión 1.6.3 puede ser usada como un estimador de ρ_{XY}, por r_{XY} si disponemos de una muestra de observaciones $Y_1, Y_2, \ldots, Y_n$ y sus correspondientes valores $X_1, Y_2, \ldots, X_n$.

CORRELACION Y REGRESION.

Supongamos que tenemos los datos (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$. Podemos obtener $r_{YX} = r_{XY}$, por aplicación de 1.6.3, si proponemos un modelo $Y = \beta_o + \beta_1 X + \varepsilon$, podemos obtener un coeficiente de regresión estimado de b_1 dado por 1.2.9.

Al comparar las dos expresiones vemos que:

$$b_1 = \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2} = \frac{\{\Sigma(Y_i-\bar{Y})^2\}^{1/2}\ \Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\{\Sigma(Y_i-\bar{Y})^2\}^{1/2}\ \{(X_i-\bar{X})^2\}^{1\ 2}.\{\Sigma(X_i-\bar{X})^2\}^{1/2}}$$

$$b_1 = \frac{\{\Sigma(Y_i-\bar{Y})^2\}^{1/2}}{\{\Sigma(X_i-\bar{X})^2\}^{1/2}} \cdot \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\{\Sigma(X_i-\bar{X})^2\}^{1/2}\{\Sigma(Y_i-\bar{Y})^2\}^{1/2}}$$

$$b_1 = \frac{\{\Sigma(Y_i-\bar{Y})^2\}^{1/2}}{\{\Sigma(X_i-\bar{X})^2\}^{1/2}} \cdot r_{XY}\ , \quad 1.6.4, \quad \text{por} \quad 1.6.3$$

de donde las sumas son de $i = 1$ a n.

En otras palabras b_1 es una versión a escala de r_{XY}, de otra - forma

$$s_Y^2 = \frac{\Sigma(Y_i-\bar{Y})^2}{n-1} \Longrightarrow (n-1)s_Y^2 = \Sigma(Y_i-\bar{Y})^2 \Longrightarrow \sqrt{n-1}\ .\ s_Y = \sqrt{\Sigma(Y_i-\bar{Y})^2}$$

$$s_X^2 = \frac{\Sigma(X_i-\bar{X})^2}{n-1} \Longrightarrow (n-1)s_X^2 = \Sigma(X_i-\bar{X})^2 \Longrightarrow \sqrt{n-1}\ .\ s_X = \sqrt{\Sigma(X_i-\bar{X})^2}$$

Asi en 1.6.4

$$b_1 = \frac{\sqrt{n-1}\ s_Y}{\sqrt{n-1}\ s_X} . r_{XY} = \frac{s_Y}{s_X} r_{XY} \cdot \therefore b_1 = \frac{s_Y}{s_X} r_{XY}$$

Entonces b_1 y r_{XY} estan cercanamente relacionados, pero nos pro- porcionará diferentes interpretaciones. La correlación r_{XY} mide la asociación entre X e Y, mientras que b_1 mide la cantidad del cambio en Y, el cual puede ser predicho cuando una unidad de - cambio es hecha en X.

C A P I T U L O I I

LA MATRIZ DE APROXIMACIÓN DE LA LÍNEA DE REGRESIÓN

Presentaremos ahora el ejemplo utilizado en el capítulo anterior utilizando el algebra de matrices, tenemos ventajas, ya que una vez que el problema es escrito y resuelto en términos matriciales, la solución puede ser aplicada al problema de regresión no importando cuantos términos haya en la ecuación de regresión.

2.1 AJUSTANDO UNA LINEA RECTA EN TERMINOS MATRICIALES: LOS ESTIMADORES DE β_o Y β_1

Nuestro objeto es encontrar β_o y β_1, del modelo $Y = \beta_o + \beta_1 X + \varepsilon_i$ por lo que se define Y como el rector de observaciones Y, X es la matriz de variables independientes, $\boldsymbol{\beta}$ es el vector de los parámetros a ser estimados y ε el vector de errores utilizando la tabla 1.1, con n = 25 observaciones, tenemos que para

$y_i = \beta_o + \beta_i X + \varepsilon_i, \quad i = 1, \ldots, 25$ es:

$$\mathbf{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \vdots \\ 10.36 \\ 11.08 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ 1 & 30.8 \\ & \\ & \\ 1 & 33.4 \\ 1 & 28.6 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \\ \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix} \qquad 2.1.1$$

Se sabe observar que

$\mathbf{Y}$ es un 25 x 1 vector fila; $\mathbf{X}$ es una matriz 25 x 2;

$\boldsymbol{\beta}$ es un vector fila; 2 x 1; $\boldsymbol{\varepsilon}$ es un vector fila 25 x 1

Puediendose realizarse las distintas operaciones entre ellas, - realizando el producto de X, tenemos

$$\mathbf{X\beta} = \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ & \\ 1 & 28.6 \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_o + 35.3\beta_1 \\ \beta_o + 29.7\beta_1 \\ \\ \beta_o + 28.6\beta_1 \end{bmatrix}$$

podemos ahora realizar la suma

$$\mathbf{X\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} \beta_o + 35.3\beta_1 \\ \beta_o + 29.7\beta_1 \\ \\ \beta_o + 28.6\beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_{25} \end{bmatrix} = \begin{bmatrix} \beta_o + 35.3\ \beta_1 + \varepsilon_1 \\ \beta_o + 29.7\ \beta_1 + \varepsilon_2 \\ \\ \beta_o + 28.6\ \beta_1 + \varepsilon_{25} \end{bmatrix}$$

Así para

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, 25 \text{ tenemos}$$

$$10.8 = \beta_0 + 35.3\ \beta_1 + \varepsilon_1$$

$$11.13 = \beta_0 + 29.7\ \beta_1 + \varepsilon_2$$

$$11.08 = \beta_0 + 28.6\ \beta_1 + \varepsilon_n$$

que en forma matricial es:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad 2.1.2$$

Ahora encontraremos b_o, b_1 estimadores de β_o y β_1 respectivamente a partir de las ecuaciones normales

$$\begin{aligned} \Sigma Y &= b_o n + b_1 \Sigma X_1 \\ \Sigma XY &= b_o \Sigma X + b_1 \Sigma X_2 \end{aligned} \qquad 2.1.3$$

que para nuestro ejemplo nos resultó

$$235.60 = 25b_o + 1315b_1$$
$$11821.432 = 1315b_o + 76323.43b_1 \qquad 2.1.4$$

Se necesitará para ello, la traspuesta de una matriz que aplica da a las matrices 2.1.1 tenemos

$$\mathbf{Y}' = (10.98,\ 11.13,\ 12.51,\ \ldots,\ 10.36,\ 11.08)$$

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 35.3 & 29.7 & 30.8 & 33.4 & 28.6 \end{bmatrix} \qquad 2.1.5$$

$$\boldsymbol{\beta}' = (\beta_o,\ \beta_1)$$

$$\boldsymbol{\varepsilon}' = (\varepsilon_1,\ \varepsilon_2,\ \ldots,\ \varepsilon_{25})$$

Para b_o, b_1 estimadores de β_o y β_1, hacemos

$\mathbf{b} = \begin{bmatrix} b_o \\ b_1 \end{bmatrix}$ estimador de $= \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix}$, siendo su traspuesta --

$\mathbf{b}' = (b_o,\ b_1)$

Pueden realizarse operaciones con 2.1.1 y 2.1.5 por ejemplo

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2,\ \ldots,\ \varepsilon_n) \cdot \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon_2^1 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$$

$$\mathbf{Y}'\mathbf{Y} = (Y_1,\ Y_2,\ \ldots,\ Y_n) \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = Y_1^2 + Y_2^2 + \ldots + Y_n^2 \qquad 2.1.6$$

además.

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 \\ 35.3 & 29.7 & 28.6 \end{bmatrix} \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ 1 & 28.6 \end{bmatrix} = \begin{bmatrix} 25 & 1315 \\ 1315 & 76323.42 \end{bmatrix} \quad 2.1.7$$

Que al comparar con 2.1.4, es la matriz de coeficientes del lado derecho, siendo de manera general, al compararla con 2.1.3 -

$$\mathbf{X'X} = \begin{bmatrix} n & \Sigma X \\ \Sigma X & X^2 \end{bmatrix} \quad 2.1.8$$

en suma,

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & 1 \\ 35.3 & 29.7 & 28.6 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ 11.08 \end{bmatrix} = \begin{bmatrix} 235.60 \\ 18821.432 \end{bmatrix}$$

que al compararla con 2.1.4 es la parte izquierda, siendo de manera general

$$\mathbf{X'X} = \begin{bmatrix} \Sigma X \\ \Sigma XY \end{bmatrix} \quad 2.1.9$$

Asi el sistema 2.1.3 puede ser escrito

$$\mathbf{X'Y} = \mathbf{X'X}\beta \quad 2.1.10$$

Para conocer el vector **b**, (estimador de β, que se obtiene por - mínimos cuadrados). Hay necesidad de conocer matrices inversas, que se obtienen de matrices cuadradas y su determinante es diferente de cero, denominándolas singulares.

Deseamos invertir la matriz **X'X**, que es 2 x 2, por lo que primero encontramos su determinante,

$$|\mathbf{X'X}| = \begin{vmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{vmatrix} = (n\Sigma X^2 - (\Sigma X)^2 = n(\Sigma X^2 - \frac{(\Sigma X)^2}{n} = n\Sigma(X-\bar{X})^2 \neq 0$$

luego podemos encontrar su inversa

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} \frac{\Sigma X^2}{n\Sigma(X-\bar{X})^2} & -\frac{\Sigma X}{n\Sigma(X-\bar{X})^2} \\ -\frac{\Sigma X}{n(\Sigma X-\bar{X})^2} & \frac{n}{n\Sigma(X-\bar{X})^2} \end{bmatrix} = \begin{bmatrix} \frac{\Sigma X^2}{n\Sigma(X-\bar{X})^2} & -\frac{\bar{X}}{\Sigma(X-X)^2} \\ \frac{\bar{X}}{\Sigma(X-\bar{X})^2} & \frac{1}{\Sigma(X-\bar{X})^2} \end{bmatrix}$$

2.1.11

o de otra forma

$$(\mathbf{X'X})^{-1} = \frac{1}{n\Sigma(X-\bar{X})^2} \begin{bmatrix} \Sigma X^2 & -\Sigma X \\ -\Sigma X & n \end{bmatrix} \qquad 2.1.12$$

que de aplicarla 2.1.11 a la matriz de 2.1.7, solamente, necesitamos conocer $|\mathbf{X'X}| = n\Sigma(X-\bar{X})^2 = 25\ (7154.42) = 178860.5$ ahora tenemos

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} \frac{76323.42}{178860.5} & -\frac{315}{178860.5} \\ -\frac{1315}{178860.5} & \frac{25}{178860.5} \end{bmatrix} = \begin{bmatrix} 0.4267203 & -0.007352098 \\ -0.007352098 & 0.000137973 \end{bmatrix}$$

Así para aplicarlo al 2.1.10

$$\mathbf{X'Y} = \mathbf{X'X\,b} \longrightarrow (\mathbf{X'X})\,\mathbf{b} = (\mathbf{X'Y})$$

aplicando $(\mathbf{X'X})^{-1}$ tenemos

$$(\mathbf{X'X})^{-1}(\mathbf{X'X})\,\mathbf{b} = (\mathbf{X'X})^{-1}(\mathbf{X'Y}) \quad ; \quad (\mathbf{X'X})^{-1} = (\mathbf{X'X}) = \mathbf{I}$$

$$\mathbf{b} = (\mathbf{X'X})^{-1}(\mathbf{X'Y}) \qquad 2.1.13$$

De aquí, cualquier, problema de regresión puede ser resuelto.

Aplicando 2.1.14, siempre que **X'X** es no singular.

Usándola para nuestro problema tenemos

$$\mathbf{b} = \begin{bmatrix} 0.4267203 & -0.007352098 \\ -0.007352098 & 0.000139773 \end{bmatrix} \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix} = \begin{bmatrix} 13.63414 \\ -0.0806772 \end{bmatrix}$$

Al comparar estos resultados con los obtenidos en 1.2.14 hay - discrepancia que ocurren frecuentemente por los redondeos de - las cifras hechas en los cálculos, pudiendo causar serios errores, dependiendo de los números involucrados, pudiendose tomar como un descuido, o podría ser debido a las calculadoras que -- pueden dar discrepancias al realizar el procedimiento de las diferentes operaciones así por ejemplo si se utiliza la fórmula - 2.1.12, es decir, por último la división al realizar al producto, tenemos

$$(\mathbf{X'X})^{-1} = \frac{1}{178860.5} \begin{bmatrix} 76323.42 & -1315 \\ -315 & 25 \end{bmatrix}$$

y aplicando 2.1.12 tenemos:

$$\mathbf{b} = \frac{1}{178860.5} \begin{bmatrix} 76323.42 & -1315 \\ -1315 & 25 \end{bmatrix} \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix}$$

$$= \frac{1}{178860.5} \begin{bmatrix} 2436614.672 \\ -14278.2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 13.622989 \\ -0.079829 \end{bmatrix}$$

Haciendo las comparaciones de los tres valores obtenidos tene-- mos.

	FORMULAS SECCION 1.2.14	MATRIZ INVERSA	MATRIZ INVERSA (DIVISION DE ULTIMO)
b_o	13.623005	13.62414	13.622989
b_1	-0.079829	-0.0806772	-0.079829

Al observar el cuadro vemos que el método último nos resulta - más preciso. Pero se darán errores cuando diferentes personas - trabajen en el mismo problema con diferentes calculadoras.

RESUMEN

Podemos expresar el modelo de línea recta de los datos de nuestro ejemplo en la forma

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

los estimadores de mínimos cuadrados de (β_o, β_1) esto es, de β son dados por

$$\mathbf{b} = \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Se debe obervar que el valor ajustado $\hat{Y}_i = b_o + b_1 X_i$ puede ser - obtenido por $\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{b}$

2.2 ANALISIS DE VARIANZA EN TERMINOS MATRICIALES

Como debemos recordar de la tabla de analisis de varianza escribimos

$$sc(b_1/b_o) = b_1\left(\Sigma X_i Y_i - \frac{\Sigma X_i \cdot \Sigma Y_i}{n}\right) = b_1(\Sigma X_i Y_i - n\bar{X}\bar{Y})$$

$$sc(b_o) = \text{corrección para la media} = \frac{(\Sigma X_i)^2}{n} = n\bar{Y}^2. \qquad 2.2.1$$

donde ambos tienen un grado de libertad.

Total no corregido = ΣY_i^2 en forma matricial $\mathbf{Y'Y}$

Ahora $sc(b_1/b_o) + sc(b_o) = b_1\Sigma X_i Y_i - b_1 n\bar{X}\bar{Y} + n\bar{Y}^2$

$$= b_1\Sigma X_i Y_i + n\bar{Y}(\bar{Y} - b\bar{X})$$

$$\bar{Y} = b_o + b_1\bar{X} \longrightarrow \bar{Y} - b_1\bar{X} = b_o$$

$$= b_1\Sigma X_i Y_i + n\bar{Y}b_o$$

$$\bar{Y} = \frac{\Sigma Y_i}{n} \longrightarrow \Sigma Y_i = n\bar{Y}$$

$$= b_1\Sigma X_i Y_i + b_o\Sigma Y_i$$

$$= b_o\Sigma Y_i + b_1\Sigma X_i Y_i$$

$$= (b_o, b_1)\begin{bmatrix}\Sigma Y_i \\ \Sigma X_i Y_i\end{bmatrix}$$

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

en términos matriciales, con 2 grados de libertad. El cuadro de analisis de varianza puede ser escrito asi:

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS
$\mathbf{b'}(b_o,b_1)$	$\mathbf{b'X'Y}$	2	
Residuos	$\mathbf{Y'Y - b'X'Y}$	n-2	s^2
TOTAL (No corregido)	$\mathbf{Y'Y}$	n	

De esta manera podemos partir la variación total $\mathbf{Y'Y}$ en dos partes, uno debido a la recta que hemos estimado, nominándolo $\mathbf{b'X'Y}$ y en un residuo, el cual muestra la variación de los puntos alrededor de la recta de regresión, con el fin de encontrar que parte de la variación total puede ser atribuida a la adición del término $\beta_i X_i$ al modelo simple $Y_i = \beta_o + \varepsilon_i$; deberíamos de -

justificar la sustracción del factor de corrección $n\bar{Y}^2$ de la suma de cuadrados de $\mathbf{b'X'Y}$ con el fin de obtener $ss(b_1/b_0)$ como antes. La cantidad $n\bar{Y}^2$ sería $sc(b_0)$ si el modelo $Y_i = \beta_0 + \varepsilon_i$ fue ajustado. El resto de $\mathbf{b'X'Y}$ mide la suma extra de cuadrados removidos por b_1, cuando el modelo $Y_i = \beta_0 + \beta_1 Y_i + \varepsilon_i$ es usado.

EJEMPLO

Para nuestro ejemplo tuvimos

$$\mathbf{b} = \begin{bmatrix} 13.62 \\ -0.0798 \end{bmatrix}; \qquad \mathbf{X'Y} = \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix}$$

$$\text{entonces } sc(b) = \mathbf{b'X'Y} = (13.62 \quad -0.0798) \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix} = 2265.5217$$

$$sc(b_0) = \frac{(\Sigma X_i)^2}{n} = \frac{(235.60)^2}{25} = 2220.2944$$

$$sc(b_1/b_0) = sc(b_1, \text{ después de asignado para } b_0)$$

$$= \mathbf{b'X'Y} - sc(b_0)$$

$$= 2265.5217 - 2220.2944$$

$$sc(b_1/b_0) = 45.2273$$

Habiendo obtenido anteriormente 45.59 para la suma de cuadrados debido a la regresión, que al compararlos se obtiene una discrepancia de 0.36 que es muy grande y que se puede deber al redondeo de números cuando se hacen los cálculos.

2.3 LA VARIANZA Y COVARIANZA DE b_0 y b_1 DE LA MATRIZ CALCULADA

Recordemos que $V(b_1) = \dfrac{\sigma^2}{\Sigma(X_i - \bar{X})^2}$. Además

$$V(b_o) = V(\bar{Y} - b_1\bar{X}) = V(\bar{Y}) + V((-\bar{X})b_1) = V(\bar{Y}) + \bar{X}^2 V(b_1) = \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{(X_i - \bar{X})^2}$$

$$= \sigma^2\left(\frac{1}{n} - \frac{\bar{X}}{\Sigma(X_i - \bar{X})^2}\right) = \sigma^2\left(\frac{\Sigma(X - \bar{X})^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}\right)$$

$$= \sigma^2\left(\frac{\Sigma X_i^2 - n\bar{X}^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}\right) = \sigma^2\frac{\Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2}, \text{ de donde}$$

$$V(b_o) = \frac{\sigma^2\Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2}$$

Anteriormente se demostró que $cov(\bar{Y}, b_1) = 0$, considerando las - equis constantes. En resumen

$$cov(b_o, b_1) = cov(\bar{Y} - b_1\bar{X},\ b_1)$$

$$= cov(\bar{Y},\ b_1) - cov(b_1\bar{X},\ b_1)$$

$$= 0 - \bar{X}\, v(b_1)$$

$$= -\bar{X}.\ \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} = -\frac{\bar{X}\sigma^2}{\Sigma(X_i - \bar{X})^2}$$

entonces podemos escribir la matriz de varianza-covarianza del vector **b** de la forma siguiente

$$V(\mathbf{b}) = V\begin{pmatrix} b_o \\ b_1 \end{pmatrix} = \begin{bmatrix} V(b_o) & cov(b_o, b_1) \\ cov(b_o, b_1) & V(b_1) \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\sigma^2 X^2}{n\Sigma(X_i - \bar{X})^2} & -\dfrac{\bar{X}\sigma^2}{\Sigma(X_i - \bar{X})^2} \\ -\dfrac{\bar{X}\sigma^2}{\Sigma(X_i - \bar{X})^2} & \dfrac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \end{bmatrix} \qquad 2.3.1$$

$$= \begin{bmatrix} \dfrac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2} & -\dfrac{\bar{X}}{\Sigma(X_i-\bar{X})^2} \\ -\dfrac{\bar{X}}{\Sigma(X_i-\bar{X})^2} & \dfrac{1}{\Sigma(X_i-\bar{X})^2} \end{bmatrix} \sigma^2$$

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \qquad 2.3.2$$

Cuando σ^2 es desconocido, usamos s^2 estimador de σ^2, obtenido del cuadro de analisis de varianza, si no hay falta de ajuste, o s_e^2, la media de cuadrados del error puro si hay desajuste.

2.4 VARIANZA DE $\hat{Y}$ USANDO LA MATRIZ DE APROXIMACION

Sea X_k un valor seleccionado de X. El valor medio predictado de Y para este valor de X es

$$\hat{Y}_k = b_o + b_1 X_k$$

Definimos el vector $\mathbf{X}_k$ como $\mathbf{X}_k' = (1, X_k)$

Podemos entonces escribir

$$\hat{Y}_k = (1, X_k) \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \mathbf{X}_k'\mathbf{b} = \mathbf{b}'\mathbf{X}_k$$

donde $\hat{Y}_k$ es una combinación lineal de las variables aleatorias b_o y b_1

$$V(\hat{Y}_k) = V(\mathbf{X}_k'\mathbf{b}) = (\mathbf{X}_k')^2 V(\mathbf{b}) = \mathbf{X}_k'\,\mathbf{X}_k'\,V(\mathbf{b})$$

pero $[V(\mathbf{b})]' = V(\mathbf{b})$

además $[\mathbf{X}_k'.V(\mathbf{b})]' = [V(\mathbf{b})'].\mathbf{X}_k = V(\mathbf{b}).\mathbf{X}_k$

Así, $V(\hat{Y}) = \mathbf{X}_k'.\mathbf{X}_k'.V(\mathbf{b}) = \mathbf{X}_k'.V(\mathbf{b})\,\mathbf{X}_k$

$$= \mathbf{X}_k'(\mathbf{X}'\mathbf{X})^{-1}\sigma^2.\mathbf{X}_k$$

$$= \mathbf{X}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_k.\sigma^2$$

Otra forma alternativa sería

$$V(\hat{Y}_k) = \mathbf{X}'_k . (\mathbf{X}' . \mathbf{X})^{-1} . \sigma^2 . \mathbf{X}_k$$

$$= \mathbf{X}'_k . V(b) . \mathbf{X}_k$$

$$= (1 \; X_k)_{1x2} \begin{bmatrix} V(b_o) & cov(b_o,b_1) \\ cov(b_o,b_1) & V(b_1) \end{bmatrix} \begin{bmatrix} 1 \\ X_k \end{bmatrix}_{2x1}$$

$$= \underset{1x2}{\left[V(b_o) + X_k \, cov(b_o,b_1) \quad cov(b_o,b_1) + X_k V(b_1) \right]} \begin{bmatrix} 1 \\ X_k \end{bmatrix}_{2x1}$$

$$= V(b_o) + X_k^2 \, cov(b_o,b_1) \; X_k \quad cov(b_o,b_1) + X_k^2 \, V(b_1)$$

$$V(\hat{Y}_k) = V(b_o) + 2X_k \, cov(b_o,b_1) + X_k^2 V(b_1)$$

2.5 SUMARIO DE LA MATRIZ DE APROXIMACION PARA AJUSTAR UNA LINEA RECTA.

1. Expresar el modelo en la forma $Y = \mathbf{X} . \beta + \varepsilon$
2. Encontrar $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$ obtener por mínimos cuadrados **b** - estimador de β, obtenido de los datos.
3. Construir $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ la suma de cuadrados debido a coeficientes y además obtener el analisis básico de analisis de varianza como sigue:

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIB.	CUADRADOS MEDIOS
REGRESION	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	2	
RESIDUOS	$\mathbf{Y}'\mathbf{Y}-\mathbf{b}'\mathbf{X}'\mathbf{Y}$	n-2	s^2(estimador de σ^2 si el modelo es correcto)
TOTAL	$\mathbf{Y}'\mathbf{Y}$	n	

Una subdivisión adicional de la suma de cuadrados, es llevada a cabo para encontrar $sc(b_1/b_o)$, la suma extra de cuadrados debido a b_1, e introduciendo el error puro, lo cual se da en la siguiente tabla

FUENTE		SUMA DE CUADRADOS	GRADOS DE LIB.	CUADRADOS MEDIOS
sc(b)	Media (b_o)	$n\bar{Y}^2$	1	
	$sc(b_1/b_o)$	$\mathbf{b'X'Y} - n\bar{Y}^2$	1	
Residuos	Falta de ajuste	$\mathbf{Y'Y} - \mathbf{b'X'Y} - sc(e.p)$	$n-2-n_e$	MS_e
	Error puro	$sc(e.p)$	n_e	s_e
	TOTAL	$\mathbf{Y'Y}$	n	

s^2

Una medida adicional de la regresión es dada por la razón:

$$R^2 = \frac{\mathbf{b'X'Y} - n\bar{Y}^2}{\mathbf{Y'Y} - n\bar{Y}^2}$$

4. Si no hay evidencia de falta de ajuste $(\mathbf{X'X})^{-1}s^2$ proveerá un un estimador de $V(b_o)$, $V(b_1)$ y $cov(b_o,b_1)$ y facilita la prueba de los coeficientes individuales.

5. Las siguientes cantidades pueden ser encontradas.

   El vector de valores fijos: $\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{b}$

   Una predicción de Y de X_k: $\hat{Y}_k = \mathbf{X}_k'\mathbf{b} = \mathbf{b'X}$

   con varianza : $V(\hat{Y}_k) = \mathbf{X}_k'(\mathbf{X'X})^{-1}.\mathbf{X}_k\sigma^2$

2.6 CASO GENERAL DE REGRESION

El uso de matrices para ajustar una línea de regresión, nos permite encontrar algún modelo con parámetros β_o, β_1,..., β_p, por mínimos cuadrados, siendo los cálculos similares como cuando en la línea recta hay que encontrar solamente a β_o y β_1.

Supongamos que tenemos un modelo bajo consideración, el cual - puede ser escrito en la forma:

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

donde $\mathbf{Y}$ es un vector de observaciones n x 1

$\mathbf{X}$ es una matriz de forma conocida n x p

β es un vector de parámetros p x 1

ε es un vector de errores n x 1

y donde $E(\varepsilon) = 0$, $V(\varepsilon) = \mathbf{I}\sigma^2$ tq. los elementos de ε son no correlacionados. Puesto que $E(\varepsilon) = 0$, una alternativa de escribir el modelo es

$$E(\mathbf{Y}) = \mathbf{X}\beta \qquad 2.6.1$$

La suma de cuadrados del error es entonces

$$\begin{aligned}\varepsilon'\varepsilon &= (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{Y}' - (\mathbf{X}\beta)')(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{Y}' - \beta'\mathbf{X}')(\mathbf{Y} - \mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

Pero $\beta'\mathbf{X}'\mathbf{Y}$ es una matriz 1 x 1, lo que la hace un escalar , cuya traspuesta tiene el mismo valor, así

$\beta'\mathbf{X}'\mathbf{Y} = (\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'(\beta'\mathbf{X}') = \mathbf{Y}'\mathbf{X}\beta$, así podemos obtener de 2.6.2

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta$$

El estimador de β es el valor $\mathbf{b}$, el cual cuando es substituido en la ecuación 2.6.2 minimiza $\varepsilon'\varepsilon$. Puede ser determinado diferenciando la ecuación 2.6.2 con respecto a β e igualando la ecuación matricial a cero, y al mismo tiempo reemplazando β por $\mathbf{b}$

$$\varepsilon'\varepsilon = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta$$

derivando tenemos

$\mathbf{0} = 0 - 0 - \mathbf{Y'X} + \boldsymbol{\beta}\mathbf{'X'X}$ de aquí

$\boldsymbol{\beta}\mathbf{'X'X} = \mathbf{Y'X}$ aplicando traspuesta a ambos lados tenemos

$\mathbf{X'}(\boldsymbol{\beta}\mathbf{'X'})' = \mathbf{X'}(\mathbf{Y'})'$

$\mathbf{X'}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X'Y}$ reemplazando $\boldsymbol{\beta}$ por $\mathbf{b}$ tenemos las ecuaciones normales

$$\mathbf{X'Xb} = \mathbf{X'Y} \qquad 2.6.3$$

De aquí se pueden presentar dos casos: la ecuación anterior consiste de p ecuaciones independientes con p parámetros, desconocidos; o estas ecuaciones normales pueden depender unas de otras, por lo que $\mathbf{X'X}$ es singular, es decir $(\mathbf{X'X})^{-1}$ no existe; por lo que el modelo tendrían que hacerse restricciones en los parámetros o ser expresado en términos de pocos parámetros. Pero si las p ecuaciones son independientes, $\mathbf{X'X}$ es no singular, es decir que $(\mathbf{X'X})^{-1}$ existe y la solución para las ecuaciones normales viene dada por

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \qquad 2.6.4$$

La solución anterior cumple las siguientes propiedades:

1. Es un estimador de $\boldsymbol{\beta}$ el cual minimiza la suma de cuadrados de el error $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$., independiente de alguna propiedad de los errores.

2. Los elementos de $\mathbf{b}$ son funciones lineales de las observaciones $Y_1, Y_2, \ldots, Y_n$, y nos proporcionan estimadores insesgado de los elementos de $\boldsymbol{\beta}$ los cuales tienen varianza minima, independiente de las propiedades de la distribución de errores.

3. Si los errores son independientes y $\varepsilon_i \sim N(0,\sigma^2)$, entonces $\mathbf{b}$ es el estimador de máxima verosimilitud de $\boldsymbol{\beta}$ (en términos vectoriales podemos

escribir $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, significando que $\boldsymbol{\varepsilon}$ sigue una distribu--- ción normal n - dimensional multivariante con $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ (donde $\mathbf{0}$, es el vector $\mathbf{0}$, donde sus componentes son todas ceros) y $V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$, esto es, $\boldsymbol{\varepsilon}$ tiene una matriz varianza-covarianza cuyos elementos diagonales son $V(\varepsilon_i)$, $i = 1, 2, \ldots, n$, son todos σ^2 y cuyos elemen tos fuera de la diagonal, covarianza $(\varepsilon_i, \varepsilon_j)$, $i \neq j = 1, \ldots, n$. son todos ceros).

La función de verosimilitud para la muestra $Y_1, Y_2, \ldots, Y_n$ está definido en este caso como el producto.

$$\prod_{i=1}^{n} \frac{1}{\sigma(2\Pi)^{1/2}} \mathbf{e}^{-\varepsilon_i^2/2\sigma^2} = \frac{1}{\sigma^n (2\Pi)^{n/2}} \mathbf{e}^{-\Sigma\varepsilon_i^2/2\sigma^2} = \frac{1}{\sigma^n (2\ )^{n/2}} \mathbf{e}^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2\sigma^2} \qquad 2.6.5$$

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \Sigma\varepsilon_i^2$$

así ya que para un valor fijo de σ, maximizando la función de ve- rosimilitud es equivalente a minimizar la cantidad $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, lo que - nos provee una justificación para el procedimiento de minimos -- cuadrados, ya que en muchas situaciones físicas la suposición de que los errores son distribuidos normalmente es bastante razona- ble.

Supongamos que hemos usado el método de mínimos cuadrados para - estimar $\boldsymbol{\beta}$ por $\mathbf{b}$. Podemos proceder con los siguientes pasos ya - sea que los errores sean o no distribuidos normalmente.

1. Los valores ajustados son obtenidos de $\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{b}$
2. El vector de residuos es dado por $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. (El cual examina remos en el siguiente capítulo). Es verdad que $\sum_{i=1}^{n} \mathbf{e}_i \hat{\mathbf{Y}}_i = 0$ - cualquiera sea el modelo. Esto puede ser visto multiplicando

la j-esima ecuación normal por el j-esima β y sumando los resultados si hay un término β_o en el modelo, es también cierto que $\sum_{i=1}^{n} \mathbf{e}_i = 0$. (El $\mathbf{e}_i$ y $\hat{Y}_i$, $i = 1,2,\ldots, n$ son los i-esimos elementos de los vectores e y $\hat{Y}$, respectivamente)

3. $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ proporciona las varianzas (en la diagonal) y covarianzas (fuera de la diagonal) del estimador. (Un estimador de σ^2 es obtenido como se describe abajo).

4. Supongamos que $\mathbf{X}_o'$ es un vector 1 x p es especificado cuyos elementos son de la misma forma como una fila de $\mathbf{X}$ tq. $\hat{\mathbf{Y}} = \mathbf{X}_o'\mathbf{b}' = \mathbf{b}'\mathbf{X}_o$ es el valor ajustado de un punto específico $\mathbf{X}_o$ por ejemplo, si el modelo fue $Y = \beta_o + \beta_1 X + \beta_{11} X^2 + \varepsilon$ entonces $\mathbf{X}_o' = (1, X_o, X_o^2)$ para un valor dado X_o.

   Entonces $\hat{Y}_o$ es el valor predicho de $\mathbf{X}_o$ por la ecuación de regresión y tiene varianza.

$$V(\hat{Y}_o) = \mathbf{X}_o'.V(\mathbf{b}).\mathbf{X}_o = \mathbf{X}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_o\sigma^2 \qquad 2.6.8$$

   Si usamos s_v^2 como un estimador de σ^2, limites de confianza de $1 - \alpha$ para el valor medio de $\mathbf{Y}$ en $\mathbf{X}_o$ es obtenido de la forma siguiente:

$$\hat{Y}_o \pm t(v, 1 - \tfrac{\alpha}{2})\, s\sqrt{\mathbf{X}_o'(\mathbf{X}_o'\mathbf{X})^{-1}\mathbf{X}_o}$$

5. Un análisis básico de varianza puede ser construido como sigue:

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS
REGRESION	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	p	MS_R
RESIDUOS	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	n - p	MS_E
TOTAL	$\mathbf{Y}'\mathbf{Y}$	n	

Una subdivisión de las partes de esta tabla puede ser llevada a cabo como sigue

5a) Si un término β_o está en el modelo podemos subdividir la suma de regresión en

$$sc(b_o) = \frac{(\Sigma Y_i)^2}{n} = n\bar{Y}^2$$

$$sc(\text{Regresión } / b_o) = sc(R/b_o) = \mathbf{b'X'Y} - \frac{(\Sigma Y_i)^2}{n} \qquad 2.6.10$$

Esas sumas de cuadrados tienen 1 y p-1 grados de liberttad respectivamente (una subdivisión más extensiva de la suma de cuadrados será discutida más adelante.

5b) Si se tienen observaciones repetidas, entonces la suma de cuadrados de residuos se puede dividir en sc(error puro) -- con n_e grados de libertad el cual estima σ^2 y sc(falta de ajuste) con $n-p-n_e$ grados de libertad, cuando todas las coordenadas $X_1, X_2, \ldots, X_k$ tengan repeticiones, nos proporciona el cuadro de analisis de varianza siguiente:

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	CUADRADOS MEDIOS	
b_o	$sc(b_o)$	1		MS
REGRESION	$sc(R/b_o)$	p-1	$MS(R/b_o)$	MR
FALTA DE AJUSTE	sc(f.d.a)	$n-p-n_e$	MS(f.d.a)	MS
ERROR PURO	sc(e.p)	n_e	MS(e.p)	ME
TOTAL	$\mathbf{Y'Y}$	n		

La razón

$$R^2 = \frac{sc(R/b_o)}{\mathbf{Y'Y} - ss(b_o)} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

es una extensión de la cantidad definida para la línea recta de

regresión y es el cuadrado del coeficiente de correlación múltiple o llamado también coeficiente de determinación múltiple.

Si $\hat{Y}_i = Y_i$ es decir si la predicciones son perfectas, entonces $R^2 = 1$.

Si $\hat{Y}_i = \bar{Y}$, esto es $b_1 = b_2 = \ldots = b_{p-1} = 0$. (Es decir, que el modelo $Y = \beta_o + \varepsilon$ ha sido ajustado), entonces $R^2 = 0$. Entonces R^2 es una medida de la utilidad de los términos, más que β_o en el modelo. Es importante observar que R^2 puede ser hecho uno, simplemente seleccionando n coeficientes propiamente en el modelo, de manera que ajuste a todos los datos exactamente. (Por ejemplo, si tenemos una observación de Y de cuatro valores diferentes de X un polinomio cúbico $Y = \beta_o + \beta_1 X + \beta^2 X^2 + \beta_3 X^3$ pasa exactamente a través de los cuatro puntos).

R^2 es usado como una medida conveniente, de buen éxito explicando la variación de los datos en la ecuación de regresión, y el valor de R^2 puede ser mejorado cuando no lo sea, adicionando un término extra al modelo para dar un mejor ajuste de los datos y no es debido al número de observaciones.

La partición del cuadro de análisis de varianza es una igualdad algebraica solamente y no depende de las propiedades de la distribución de los errores.

Sin embargo si se asume que $\varepsilon_i \sim N(0,\sigma^2)$ y que los son independientes, esto es que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$ podemos hacer lo siguiente

1. La prueba de falta de ajuste para tratar la razón

$$\left[\frac{sc(fda)/(n - p - n_e)}{sc(error\ puro)/n_e}\right] \qquad 2.6.12$$

Como una variable $F[(n-p-n_e), n_e]$, y comparando su valor con $F[(n-p-n_e), n_e, 1 - \alpha]$ Si no hay falta de ajuste, $ss(Residuos)/(n-p) = MS_E$ usualmente llamando s^2 es un estimador ensayado de σ^2, si la falta de ajuste no puede ser probada, usar s^2 en el modelo, como estimador de σ^2 implica que este es correcto.

2. Prueba de la ecuación de regresión con todos los parámetros (más específicamente H_o: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$, contra H_1: no todos $\beta_i = 0$) tratando la razón de los cuadrados medios.

$$\frac{[sc(R/b_o)/(p - 1)]}{s^2} \qquad 2.6.13$$

Como una variable $F(p-1,\nu)$, donde $\nu = n - p$

Para un nivel α. Se observa que si la razón de cuadrados medios excede a $F(p-1,\nu,1-\alpha)$, significa que la regresión obtenida es - " estadísticamente significante "; en otras palabras la proporción de variación observada en los datos, los cuales han sido - computarizados por la ecuación, es tan grande que sería esperada una probabilidad de $100(1-\alpha)\%$ en conjuntos similares de da-- tos con los mismos valores de n y $\mathbf{X}$.

3. Establecido que $\mathbf{b} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$

4. Obtener una región de confianza $100(1-\alpha)\%$ para todos los pa rámetros de la ecuación.

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}\ (\boldsymbol{\beta} - \mathbf{b}) \leq ps^2\ F(p,\nu,\ 1 - \alpha)$$

donde $F(p,\nu,1-\alpha)$ es el $1 - \alpha$ puntos (α puntos altos) de la dis-- tribución $F(p,\nu)$, y donde s^2 tiene el mismo significado, que en

(1) anterior y si el modelo asumido es el correcto. En general esto será útil solamente cuando p es pequeño, digamos 2,3 ó 4, a menos que la información sea presentada cuidadosamente que pueda ser entendida fácilmente. La última desigualdad nos da una región cuyo entorno da una "figura elíptica"., de muchas dimensiones, p, como parámetros β haya.

Podemos obtener intervalos de confianza para los parámetros separadamente por

$$b_i \pm t(v, 1-\tfrac{1}{2}\alpha) \text{ (estimación d.t.}(b_i))$$

donde el "estimados d.t.(b_i)"es la raiz cuadrada del término iésimo de la diagonal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}s^2$.

Por ejemplo cuando hay dos parámetros β_o y β_1 se utiliza la matriz (2.3.1) así para $i = 0$, tenemos

$$b_o \pm t(v, 1 - \tfrac{1}{2}\alpha).\ s.\sqrt{\frac{\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2}}$$

habiendo sustituido s^2 por σ^2 ya que $V(b_o)$ es

$$\frac{\sigma^2\Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2}$$

Cuando dos parámetros son considerados la fig. 2.1, ilustra la situación si la región es considerada al 95% de confianza, los parámetros verdaderos β_1 y β_2, como puntos, estarán dentro de la elipse, en lo cual los datos estiman convenientemente a los parámetros.

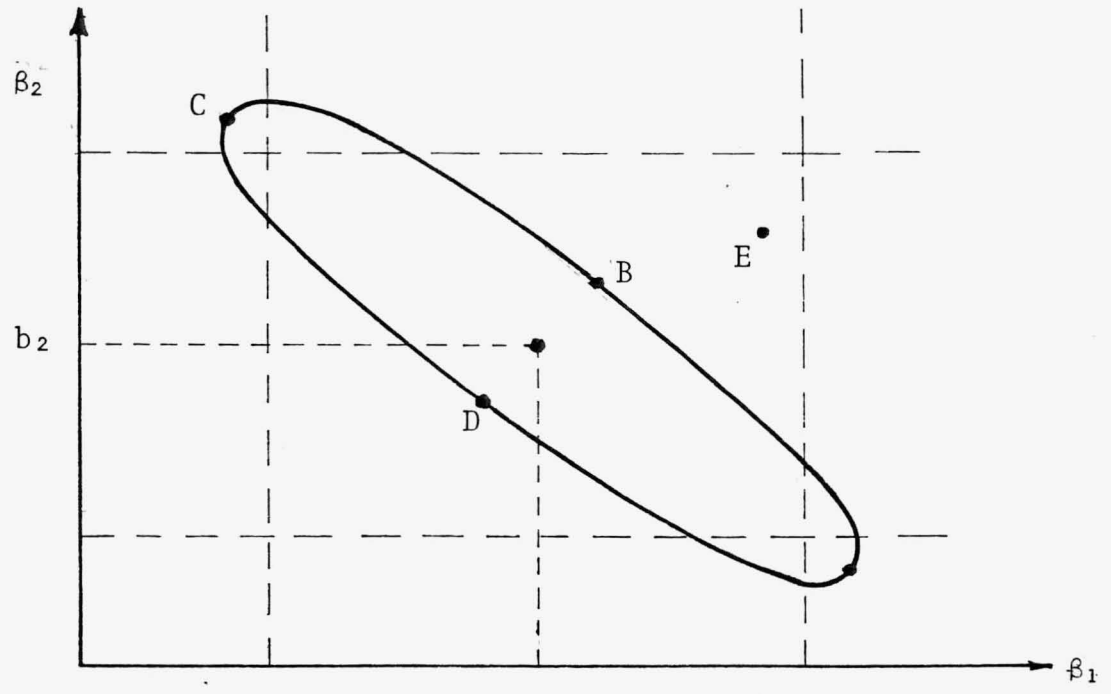


Figura 2.1

Si tomamos en cuenta la correlación entre los estimadores b_1 y b_2. El 95% de los intervalos de confianza individuales para β_1 y β_2 separadamente, son apropiados para especificar rangos, para los parámetros individuales independiente del valor de los otros parámetros.

Los intervalos definen un rectángulo que se ve en la figura, 2.1

Se podría considerar las coordenadas del punto E como razonables para (β_1, β_2), pero no cae en la región de confianza, por lo que no es razonable tomarlo. Cuando se tienen dos parámetros, no es difícil dibujar la elipse de confianza, una forma sería encontrar las coordenadas de los puntos al final al eje mayor de la región., en la figura 2.1 serian A,B,C y D. Pero para cada intervalo, si se debe tener la medida de $V(b_i)$ y la medida de la covarianza de b_i y b_j. Cuando b_i y b_j tienen diferentes medidas en sus varianzas, la correlación entre b_i y b_j viene dada por

$$\rho_{ij} = \frac{\text{cov}(b_i - b_j)}{\left[V(b_i)V(b_j)\right]^{1/2}}$$

Si este valor no es pequeño, entonces ocurre la situación ilustrada en la figura 2.1

Si ρ_{ij} es cercano a cero, entonces la región rectangular definida por los intervalos de confianza individuales, aproximará a la correcta región de confianza.

La elongación de la región depende de los valores de $V(b_i)$ y $V(b_j)$ como se muestra en la figura 2.2

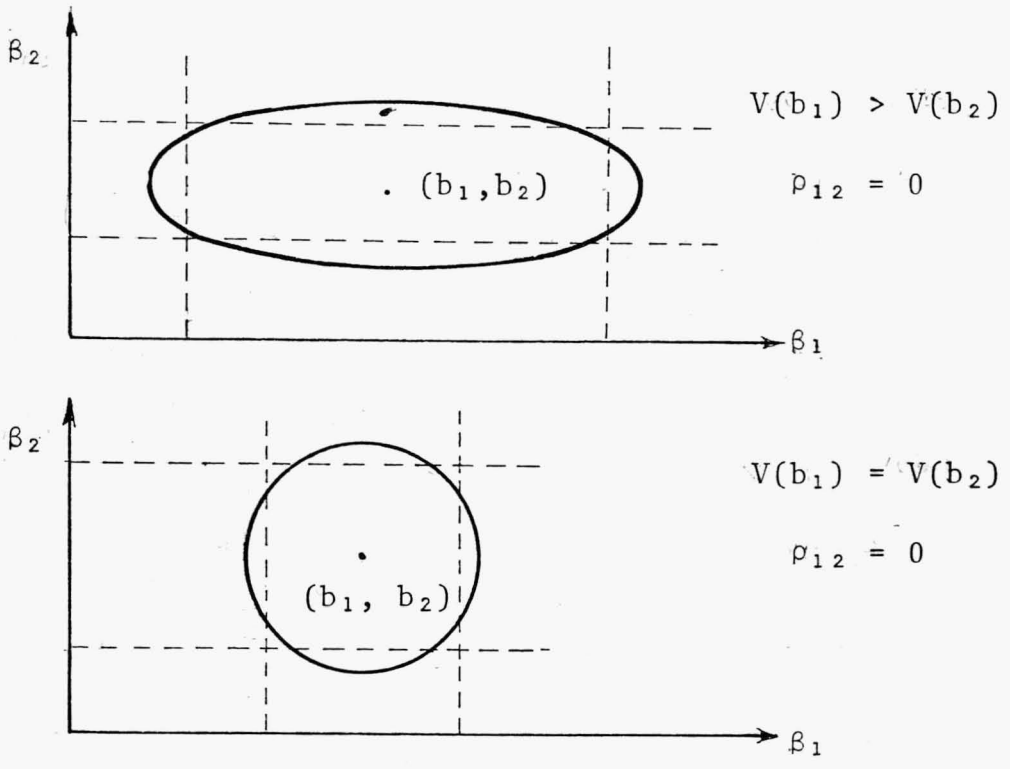


Figura 2.2

Nota: Una forma equivalente de escribir el modelo original es de la forma

$$E(Y - \bar{Y}) = \beta_1(X_1 - \bar{X}_1) + \beta_2(X_2 - \bar{X}_2) + \cdots + \beta_k(X_k - \bar{X}_k)^2$$

donde $\bar{Y}$, $\bar{X}_1$, $\bar{X}_2, \ldots, \bar{X}_k$ son las medias observadas de los datos actuales, entonces los intervalos de confianza no involucran a

β_o, el cual algunas veces es de poco interés.

2.7 EL PRINCIPIO DE LA " SUMA EXTRA DE CUADRADOS "

Ahora surge la siguiente pregunta, cuán válido sería ser contestada, examinando la correspondiente suma de cuadrados de la porción correspondiente al término agregado, entonces la media de cuadrados derivada puede ser comparada con s^2, estimador de σ^2, si es pequeña, el término será removido; si no, queda en el modelo. Por ejemplo cuando se ajusta, la línea recta, al agrupar el término $\beta_1 X$, la suma de cuadrados de ese término se representó por $sc(b_1/b_o)$

Exponemos ahora el procedimiento más general.

Supongamos las funciones z_1, z_2, ..., z_p son funciones conocidas de las variables básicas X_1, X_2, ..., y supongamos que se dispone de los valores de las equis y la correspondiente respuesta Y.

Consideraremos los modelos siguientes

1. $Y = \beta_o + \beta_1 z + \beta_2 z_2 + \ldots\ldots + \beta_p z_p + \varepsilon$

Supongamos que obtenemos los siguientes estimadores por mínimos cuadrados: $b_o(1)$, $b_1(1)$, ..., $b_p(1)$ y supongamos que $sc(b_o(1), b_1(1), \ldots, b_p(1)) = s_1$ y no hay fuera de ajuste, tomemos como estimador de σ^2 a s^2, obtenido de los residuos del modelo (1).

2. $Y = \beta_o + \beta_1 z_1 + \beta_2 z_2 + \ldots\ldots\ldots + \beta_q z_q + \varepsilon \qquad (q < p)$

Las zetas en el modelo (2) son las mismas que en el modelo (1). En este segundo modelo hay pocos términos y supongamos que ob-

tenemos los estimadores por mínimos cuadrados

$b_o(2), b_2(2), \ldots, b_q(2)$

Nota. Estos estimadores pueden ser o no los mismos que $b_o(1), b_1(1), \ldots, b_p(1)$ anteriores. Si ellos son idénticos entonces $b_i(1)$ y $b_j(1)$ son funciones lineales ortoonales para $1 \leq i \leq q$ y $q + 1 \leq j \leq p$. Esto sucede en el modelo 1 cuando las q columnas de la X matriz son todas ortogonales a las últimas p - q columnas.

Supongamos que $sc(b_o(2), b_1(2), \ldots, b_q(2)) = s_2$.; hagamos $s_1 - s_2$ es la suma extra de cuadrados debido a la inclusión de $\beta_q + 1\ z_q + 1 + \ldots, + \beta_p\ z_p$. Puesto que s_1 tiene p + 1 grados de libertad y s_2 tiene q + 1, $s_1 - s_2$ tiene p - q grados de libertad.

Puede ser demostrado que si $\beta_q + 1 = \beta_q + 2 = \ldots = \beta_p = 0$ entonces $E\{(s_1 - s_2)/(p - q)\} = \sigma^2$. En resumen si los errores son normalmente distribuidos, $s_1 - s_2$ serán distribuidos como $\sigma^2 x_{p-q}$ independientemente de s^2. Esto significa que si comparamos $s_1 - s_2/p - q$ con s^2 por una prueba $F(p - q, v)$, donde v es el número de grados de libertad en el cual se basa s^2, para probar la hipótesis:

$H_o: \beta_q + 1 = \beta_q + 2 = \ldots = \beta_p = 0$

Podemos escribir $s_1 - s_2$ convenientemente como $sc(b_q + 1, \ldots, b_p/b_o, b, \ldots, b_q)$ que se debe leer como la suma de cuadrados $b_q + 1, \ldots, b_q$ dados $b_o, b_1, \ldots, b_q$; debiendo siempre tener en cuenta que son dos modelos involucrados. así para un modelo de regresión, podemos obtener en forma conti

nuada $sc(b_o)$, $sc(b_1/b_o)$, $sc(b_2,/b_o, b_1),\ldots,$ $sc(b_p/b_o,\ldots,b_p-1)$ todas esas sumas de cuadrados se distribuyen independientemente de s^2, teniendo cada una un grado de libertad, pudiendose comparar sus cuadrados medios, con s^2 por una prueba F, para cada caso.

Cuando los términos ocurren en grupos naturales, como en los polinomios, por ejemplo a) β_o b) término de primer órden c) término de segundo órden, se puede construir la suma de cuadrados así, sc(b_o), sc(b de primer orden/b_o), sc(b segundo orden/ b_o, b de primer oden) y comparar estos con s^2.

El número de grados de libertad, es el número de parámetros en la suma de cuadrados antes del signo de división, excepto cuando los estimadores son linealmente dependientes, tomandose entonces el número de ecuaciones lineales independientes, obtenidos sus cuadrados medios así (suma de cuadrados)/(grados de libertad), pudiendose obtener una prueba-F dividiendola entre s^2.

Este principio es un caso especial de la prueba de hipotesis general lineal. En el tratamiento más general, la suma extra de cuadrados es calculada de la suma de cuadrados residuales y no de la suma de cuadrados de la regresión, puesto que la suma **Y'Y** es el mismo para ambos cálculos de regresión.

2.8 COLUMNAS ORTOGONALES EN LA MATRIZ **X**

Supongamos que tenemos un problema de regresión, involucrado parámetros β_o, β_1 y β_2. Usando el principio de suma extra de cuadrados podemos calcular el número de cantidades tales como :

$sc(b_2)$ de el modelo $Y = \beta_2 X_2 + \varepsilon$

$sc(b_2/b_o)$ de el modelo $Y = \beta_o + \beta_2 X_2 + \varepsilon$

$sc(b_2/b_o, b_1)$ de el modelo $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Estos usualmente tendrán, completamente valores numéricos diferentes, excepto cuando la columna " β_2 " de la $\mathbf{X}$ matriz es - ortogonal a las columnas " β_o " y " β_1 " cuando esto sucede se puede hablar de $sc(b_2)$.

Ahora examinaremos la situacion con más detalle.

Supongamos que en el modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, dividimos la matriz $\mathbf{X}$ en un conjunto de t columnas en forma de matriz:

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t)$$

Una corespondiente división para $\boldsymbol{\beta}$ puede ser hecha

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_t \end{bmatrix}$$

donde el número de columnas de $\mathbf{X}_i$ es igual al número de filas en $\boldsymbol{\beta}_i$, $i = 1, \ldots, Y$. Entonces el modelo puede ser escrito.

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t) \begin{bmatrix} \beta_1 \\ \beta_2 \\ \\ \beta_t \end{bmatrix} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \ldots + \mathbf{X}_t\boldsymbol{\beta}_t$$

Supongamos que $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \\ b_t \end{bmatrix}$

es el vector estimador de $\boldsymbol{\beta}$ para este modelo, obtenido de las ecuaciones normales.

$$\mathbf{X'Xb} = \mathbf{X'Y}$$

RESULTADO. Si las columnas de $\mathbf{X}_i$ son ortogonales a las columnas $\mathbf{X}_j$ para todo i, $j = 1,2, \ldots, t (i \neq j)$, esto es, $\mathbf{X}_i'\mathbf{X}_j = 0$, se sigue entonces que

$$sc(\mathbf{b}) = sc(b_1) + sc(b_2) + \ldots\ldots + sc(b_t)$$

$$= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} + \mathbf{b}_2\mathbf{X}_2\mathbf{Y} + \ldots\ldots\ldots + \mathbf{b}_t\mathbf{X}_t\mathbf{Y}$$

y $\mathbf{b}_i$ es el estimador de mínimos cuadrados de $\boldsymbol{\beta}_i$, y $sc(\mathbf{b}_i) = \mathbf{b}_i'\mathbf{X}_i'\mathbf{Y}$ ya sea que algunos de los otros términos esten en el modelo o no, Entonces

$$sc(\mathbf{b}_j / \text{ algún conjunto } b_j\text{'s}, j \neq i)$$

(Note que no es necesario para las columnas de $\mathbf{X}_i$ sean ortogonales a todas las otras de $\mathbf{X}$.)

Considerando el caso $t = 2$.

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

donde $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2'\mathbf{X}_1 = 0$ (esto significa que todas las columnas de $\mathbf{X}_1$ son ortogonales en todas las columnas de $\mathbf{X}_2$). El modelo se escribe así.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\mathbf{X}_1, \mathbf{X}_2) \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Las ecuaciones normales con: $\mathbf{X'Xb} = \mathbf{X'Y}$

$$\mathbf{X'Xb} = \mathbf{X'Y} \quad , \quad \mathbf{X}' = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{bmatrix}$$

$$\begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \quad (X_1, X_2) \quad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} Y$$

$$\begin{bmatrix} X_1' X_1 & X_1' X_2 \\ X_2' X_1 & X_2' X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1' Y \\ X_2' Y \end{bmatrix}$$

Puesto que los términos fuera de la diagonal son ceros.

$(X_1' X_2 = 0 = X_2' X_1)$, tenemos

$$\begin{bmatrix} X_1' X_1 & 0 \\ 0 & X_2' X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1' Y \\ X_2' Y \end{bmatrix}$$

$$\begin{bmatrix} X_1' X_1 b_1 \\ X_2' X_2 b \end{bmatrix} = \begin{bmatrix} X_1' Y \\ X_2' Y \end{bmatrix}$$

Tenemos las ecuaciones normales

$$X_1' X_1 b_1 = X_1' Y$$

$$X_2' X_2 b_2 = X_2' Y$$

con soluciones

$$b_1 = (X_1' X_1)^{-1} X_1 Y$$

$$b_2 = (X_2' X_2)^{-1} X_2 Y$$

Asumiendo que las matrices invertidas son no singulares.

Entonces b_1 es el estimador de β_1, ya sea que β_2 este en el modelo o no. Ahora

$$sc(b_1) = b_1' X_1' Y \qquad sc(b_2) = b_2' X_2' Y$$

Entonces

$$sc(\mathbf{b}) = sc(b_1, b_2) = \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

$$= (\mathbf{b}_1', \mathbf{b}_2')(\mathbf{X}_1, \mathbf{X}_1)'\mathbf{Y}$$

$$= (\mathbf{b}_1', \mathbf{b}_2')\begin{pmatrix}\mathbf{X}_1' \\ \mathbf{X}_2'\end{pmatrix}\mathbf{Y}$$

$$= (\mathbf{b}_1', \mathbf{b}_2')\begin{pmatrix}\mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y}\end{pmatrix}$$

$$= \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} + \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y}$$

$$sc(\mathbf{b}) = sc(\mathbf{b}_1) + sc(\mathbf{b}_2)$$

se sigue que

$$sc(\mathbf{b}_1/\mathbf{b}_2) = sc(\mathbf{b}_1, \mathbf{b}_2) - sc(\mathbf{b}_2) = sc(\mathbf{b}_1)$$

así

$$sc(\mathbf{b}_2/\mathbf{b}_1) = sc(\mathbf{b}_1, \mathbf{b}_2) - sc(\mathbf{b}_1) = sc(\mathbf{b}_2)$$

Y esto solamente depende de la ortogonalidad de $\mathbf{X}_1$ y $\mathbf{X}_2$, para casos mayores que t = 2 es similar.

2.9 PRUEBA-F PARCIAL Y PRUEBA-F SECUENCIAS

Anteriormente hemos obtenido la suma extra de cuadrados para uno o más coeficientes de un modelo dados otros coeficiente de otro modelo. Si tenemos varios términos en un modelo, de regresión, podemos pensar si cada uno de ellos ingresa a alguna ecuación deseada, entonces

$ss(b_i/b_o, b_1, \ldots, b_i\text{-}1, b_i\text{+}1, \ldots, b_k)$, $i = 1, 2, \ldots, k$

tendremos un grado de libertad, la contribución de cada b_i, a la suma de cuadrados, cuando ingresa al nuevo modelo que no

contiene β_i, es decir la medida del valor de adicionar un término β_i, al modelo que no lo tiene, su cuadrado medio es igual a la suma de cuadrados con un grado de libertad, puede ser comparada con s^2 por una prueba-F. Este tipo de prueba es llamado prueba-F parcial para β_i.

Así por ejemplo, si el término a considerar es X_t, podemos hablar de una prueba-F parcial en la variable X_t, aunque sepamos que la prueba-F actua sobre el coeficiente β_t. Cuando un modelo conveniente está siendo " construido " la prueba-F, es un -- criterio útil para adicionar o remover términos en el modelo. El efecto de una sola variable (digamos X_q), en una determinada respuesta puede ser muy grande si solo incluye a X_q en la ecuación de regresión, pero puede afectar muy poco a la respuesta, si se agrega luego que se hayan agregado otras variables anteriormente y se debe a que X_q está altamente correlacionada con-ellas en la ecuación de regresión.

La prueba-F parcial puede ser hecha para todos los coeficientes de regresión, como si cada variable correspondiente fué la última en entrar a la ecuación, para ver los efectos relativos últimos al entrar en la ecuación. Esta información puede ser - combinada por otra información, si una selección de variables - se necesita, supongamos que X_1, o X_2 (una u otra) será usada - para obtener una respuesta Y de la ecuación de regresión, y supongamos que el uso de X_1 proporciona un pequeño error de predicción que el uso de X_2. Usaríamos por lo tanto X_1 como valor predictivo.

Sin embargo, si X_2 fuera una variable a través de la cual el nivel de respuesta es controlado, mientras que X_1 no. Y si el control fuera importante, sería mejor utilizar X_2, más bién que X_1, como una variable independiente en trabajos futuros.

Cuando unas variables son adicionadas, una por una en etapas a una ecuación de regresión, podemos hablar de la prueba-F secuencial.

2.10 ENSAYANDO UNA HIPOTESIS LINEAL GENERAL EN REGRESION

Cuando estamos haciendo un trabajo de regresión es a menudo necesario una prueba estadística de una hipótesis H_o, el cual implica funciones lineales de los verdaderos coeficientes de regresión $\beta_0, \beta_1, \ldots, \beta_k$

EJEMPLO 1

Modelo 1: $E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2$

H_o: $\beta_1 = 0, \beta_2 = 0$ (dos funciones lineales, independientes)

Ejemplo 2:

Modelos 2: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

H_o: $\beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$ (k funciones lineales independientes)

Ejemplo 3:

Modelo: $E(Y) = \beta_o + \beta_1 X_1 + \ldots + \beta_k X_k$

H_o: $\beta_1 - \beta_2 = 0, \beta_2 - \beta_3 = 0, \quad , \beta_k^{-1} \beta k = 0$

(k-1 funciones lineales independientes)

Note que esto expresa la hipótesis

H_o: $\beta_1 = \beta_2 = \ldots = \beta_k = \beta$

Ejemplo 4 (Caso general)

Modelo: $E(Y) = \beta_o + \beta_1 X_1 + \quad + \beta_k X_k$

$$H_o : \begin{aligned} c_{11}\beta_1 + \ldots + c_{1k}\beta_k &= 0 \\ c_{21}\beta_1 + \ldots + c_{2k}\beta_k &= 0 \\ &\vdots \\ cm_1\beta_1 + \ldots + cm_k\beta_k &= 0 \end{aligned}$$

En esta hipótesis hay m ecuaciones lineales de $\beta_1, \beta_2, \ldots, \beta_k$, todos los cuales no pueden ser independientes. H_o puede ser expresado en forma matricial así:

$$H_o : \mathbf{C\beta}$$

Donde

$$\mathbf{\varepsilon} \quad \begin{bmatrix} c_{11} \ldots c_{1k} \\ c_{21} \ldots c_{2k} \\ \cdots\cdots\cdots \\ c_{m1} \ldots c_{mk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Supondremos en lo que sigue que las m ecuaciones son dependientes, y que si tuviese ecuaciones independientes, las otras m-q se tomarían como combinaciones lineales de las otras q.

Explicación del caso general.

Ensayando una hipotesis general lineal, $\mathbf{C\beta} = \mathbf{0}$

Supongamos que el modelo bajo consideración, es correcto y es

$$E(\mathbf{Y}) = \mathbf{X\beta}$$

donde $\mathbf{Y}$ es n x 1, $\mathbf{X}$ es n x p, y $\boldsymbol{\beta}$ es p x 1. Si $\mathbf{X'X}$ es inversible podemos estimar $\boldsymbol{\beta}$ así.

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

La suma de cuadrados de residuos para este analisis es

$$scr = \mathbf{Y'Y} - \mathbf{b'X'Y}$$

esta suma de cuadrados tiene n-p grados de libertad.

La hipótesis lineal al ser ensayada es H_o: $\mathbf{C\beta} = Q$. que proporciona q ecuaciones independientes en los parámetros $\beta_o, \beta_1, \ldots, \beta_k$ y m-q ecuaciones dependientes, (combinaciones lineales de las q primeras) utilizando las q ecuaciones independientes para encontrar las q betas, en términos de las otras p-q de ellas, sustituyendo esas ecuaciones en el modelo original obtenemos el modelo reducido

$$E(\mathbf{Y}) = \mathbf{Z}\,\alpha$$

donde α es un vector de parámetros a ser estimados.

Hay p-q parámetros. El lado derecho de la igualdad **Z** es nx(**p-q**) y α (p-q)x1, representa el resultado de sustitución **X** para las betas dependientes.

Podemos ahora estimar el vector parámetro α en el nuevo modelo por

$$\mathbf{a} = (\mathbf{Z'Z})^{-1}\mathbf{Z'Y}\ ,\ \text{si } \mathbf{Z'Z} \text{ es no singular.}$$

La suma de cuadrados de residuos se obtiene así

$$scw = \mathbf{Y'Y} - \mathbf{a'Z'Y}$$

esta suma tiene n-p+q grados de libertad.

Puesto que menos parámetros son involucrados, en este segundo analisis, la **scw** será mayor a **scr**. La diferencia scw - scr es llamada la suma de cuadrados de residuos debido a la hipótesis $\mathbf{C\beta} = \mathbf{0}$ y tiene q grados de libertad [n-p+q - (n-p) = q]

Una prueba de la hipótesis H_o: $\mathbf{C}\beta = 0$, es obtenido considerando la razón

$$\frac{scw - scr}{q} \quad / \quad \frac{scr}{n-p} \qquad 2.10.1$$

refiriendose a la distribución F(q,m-p) de la manera usual. Si los errores son distribuidos normalmente y son independientes, la prueba exacta.

La prueba apropiada para los ejemplos 1 y 2 está dada por

$[sc(R/b_o) /p-1] / s^2$ dado por 2.6.13 con k = p-1 con F(p-1,v),

donde q = p-1 y v = n-p. El modelo reducido en ambos casos - es, (puesto que $\beta_i = 0$, $i = 1, \ldots, n$)

$$E(\mathbf{Y}) = \mathbf{J}'\beta_o$$

donde $\mathbf{J}' = (1,1, \ldots, 1)$. Otra manera de escribir este modelo es

$$E(Y_i) = \beta_{oo}, \ldots, \quad i = 1, \ldots, n$$

puesto que $b_o = \bar{Y}$, $scw = \mathbf{Y}'\mathbf{Y} - n\bar{Y}_2$, con n-1 grados de libertad; mientras que $scr = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$ con (n-k-1) grados de libertad tal que la razón para la prueba $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (para - ejemplo 2, k = 2) es simplemente

$$\frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{k} \quad / \quad \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}}{n - k - 1} \qquad 2.10.2$$

(Para el numerador por aplicación de 2.10.1, tenemos

$$\frac{(\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2) - (\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y})}{(n - 1) - (n - k - 1)} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}_2}{k}$$

esto referido a la distribución F(k,n-k-1). Este es exactamente el procedimiento de la ecuación 2.6.13 con, k = p-1, v = n-k-1 y $s^2 = MS_E = scr./v$

Procedimiento en un simple, pero no típico caso.

Ejemplo

Dado el modelo $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, probar la hipótesis

H_o: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ donde

$\mathbf{Y}' = (1,4,8,9,3,8,9)$

$\boldsymbol{\beta}' = (\beta_o, \beta_1, \beta_2, \beta_{11})$

$$\mathbf{X} = \begin{array}{c} \begin{matrix} 1 & X_1 & X_2 & X_1^2 \end{matrix} \\ \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \end{bmatrix} \end{array} \quad \text{y} \quad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 2 & -2 & 93 \end{bmatrix}$$

Solución

Encontraremos scr. para el modelo original, cuando es ajustado el modelo

$$E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 7 & 0 & 3 & 4 \\ 0 & 4 & 0 & 0 \\ 3 & 0 & 9 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & -1/6 & -1/2 \\ 0 & \frac{1}{4} & 0 & 0 \\ -\frac{1}{6} & 0 & \frac{1}{6} & 1/6 \\ -\frac{1}{2} & 0 & \frac{1}{6} & 3/4 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 42 \\ 4 \\ 38 \\ 22 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X-Y} = \begin{bmatrix} 11/3 \\ 1 \\ 3 \\ 11/6 \end{bmatrix}$$

$\mathbf{b'X'Y} = 312.33$, $\mathbf{Y'Y} = 316.$

scr = $\mathbf{Y'Y} - \mathbf{b'X'Y} = 316 - 312.33 = 3.67$

Ecuaciones para la hipótesis nula H_o: $\mathbf{C}\,\boldsymbol{\beta} = 0$

$$\begin{aligned} \beta_{11} &= 0 \\ \beta_1 - \beta_2 &= 0 \\ \beta_1 - \beta_2 + \beta_{11} &= 0 \\ 2\beta_1 - 2\beta_2 + 93\beta_{11} &= 0 \end{aligned}$$

Estas hipótesis pueden ser expresadas simplemente como:

H: $\beta_{11} = 0$, $\beta_1 = \beta_2 = \beta.$, ya que la tercera y cuarta ecuaciones son combinaciones de las dos primeras.

Sustituyendo estas ecuaciones en el modelo, da el modelo reduci do.

$$E(\mathbf{Y}) = \beta_o + \beta(\mathbf{X}_1 + \mathbf{X}_2) = \alpha_o + \alpha\,\mathbf{Z}$$

donde $\alpha_o = \boldsymbol{\beta}_o$, $\alpha = \beta$, $\mathbf{Z} = \mathbf{X}_1 + \mathbf{X}_2$

entonces.

$$Z = \begin{bmatrix} 1 & (-1-1) \\ 1 & (1-1) \\ 1 & (-1+1) \\ 1 & (1+1) \\ 1 & (0+0) \\ 1 & (0+1) \\ 1 & (0+2) \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$$

Ahora encontremos

$$(Z'Z)^{-1} = \begin{bmatrix} 7 & 3 \\ 3 & 13 \end{bmatrix}^{-1} = \frac{1}{82} \begin{bmatrix} 13 & -3 \\ -3 & 7 \end{bmatrix}$$

$$(Z'Y) = \begin{bmatrix} 42 \\ 42 \end{bmatrix} \quad ; \quad a = (Z'Z)^{-1} Z'Y = \frac{21}{41} \begin{bmatrix} 10 \\ 4 \end{bmatrix}$$

$a'Z'Y = 301.17$

$scw = Y'Y - a'z'Y = 316 - 301.17 = 14.83$

Ahora $p = 4$, $n = 7$, $q = 2$, $n-p = 3$ y

$scw - scr = 14.83-3.67 = 11.16 = sc$ debido a la hipotesis

El estadístico apropiado para H_o, es entonces

$11.16/2 \div 3.67/3 = 4.56$. Puesto que $F(2,3,0.95) = 9.55$ no rechazamos H_o; puesto que el modelo original fue $E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X^2$ y la hipótesis no rechazada implica que $\beta_{11} = 0$, $\beta_1 = \beta_2 = \beta$, un modelo más apropiado sería $E(Y) = \beta_o + \beta(X_1 + X_2)$.

2.11 ASIGNANDO PESOS EN MINIMOS CUADRADOS

Algunas veces sucede que algunas de las observaciones usadas en

análisis de regresión son " menos confiables " que otras, lo - que usualmente significa es que la varianza de las observaciones no son iguales, en otras palabras la matriz $\mathbf{V}(\varepsilon)$ no es igual a $\mathbf{I}\sigma^2$, teniendo elementos diferentes en su diagonal. Puede suceder, en algunos problemas, que los elementos fuera de la diagonal de $V(\varepsilon)$ no ser cero, esto es, que las observaciones son correlacionadas.

Cuando uno u otro o ambas de esos eventos ocurre, la estimación ordinaria por mínimos cuadrados, es decir, para $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ no es aplicable y es necesario corregir el procedimiento, de la -- siguiente manera, transformando las observaciones $\mathbf{Y}$ a otras variables $\mathbf{Z}$, la cual debe satisfacer las suposiciones de la regresión [es decir $\mathbf{Z} = \mathbf{Q}\beta + \mathbf{f}$, con $E(\mathbf{f}) = 0$ $V(\mathbf{f}) = \mathbf{I}\sigma^2$, para la prueba-F, intervalos de confianza, debe ser válido que $\mathbf{f} \sim N(0,\mathbf{I}\sigma^2)$ y aplicar el análisis usado anteriormente para encontrar el estimador, pudiendose posteriormente expresarlo en términos de $\mathbf{Y}$ se descubrirá como transformar el procedimiento usual.

Supongamos que el modelo en consideración es:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \qquad 2.11.1$$

donde

$$E(\varepsilon) = 0, \quad V(\varepsilon) = \mathbf{V}\sigma^2 \quad y \quad \varepsilon \sim N(0,\mathbf{V}\sigma^2) \qquad 2.11.2$$

Es posible encontrar una matriz simétrica no singular $\mathbf{p}$, tq.

$$\mathbf{p}'\mathbf{p} = \mathbf{p}.\mathbf{p} = \mathbf{p}^2 = \mathbf{V} \qquad 2.11.3$$

Escribamos

$$\mathbf{f} = \mathbf{p}^{-1}.\varepsilon \quad , \quad tq \quad E(\mathbf{f}) = 0 \qquad 2.11.4$$

$[E(\mathbf{f}) = E(\mathbf{p}^{-1}\boldsymbol{\varepsilon}) = \mathbf{p}^{-1}\, E(\boldsymbol{\varepsilon}) = 0, \quad \text{por } 2.11.2]$

Ahora si se tiene que, si $\mathbf{f}$ es un vector de variable aleatoria tq $E(\mathbf{f}) = \mathbf{0}$, entonces $E(\mathbf{f}\,\mathbf{f}') = V(\mathbf{f})$, donde la esperanza es tomada separadamente para cada término en la matriz cuadrada n x n. Entonces.

$$V(\mathbf{f}) = E(\mathbf{f}.\mathbf{f}') = E(\mathbf{p}^{-1}\boldsymbol{\varepsilon}.\boldsymbol{\varepsilon}'\mathbf{p}^{-1}), \quad \text{ya que } (\mathbf{p}^{-1}\boldsymbol{\varepsilon})' = \boldsymbol{\varepsilon}'\mathbf{p}^{-1}$$

puesto que $(\mathbf{p}^{-1})' = \mathbf{p}^{-1}$

$$= \mathbf{p}^{-1}E(\boldsymbol{\varepsilon}.\boldsymbol{\varepsilon}')\mathbf{p}^{-1} \quad - [\text{por } E(\boldsymbol{\varepsilon}) + 0 \Rightarrow E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = V(\boldsymbol{\varepsilon})]$$

$$= \mathbf{p}^{-1}V(\boldsymbol{\varepsilon})\mathbf{p}^{-1}$$

$$= \mathbf{p}^{-1}.\ \mathbf{V}\sigma^2\mathbf{p}^{-1} \qquad \text{por } 2.11.2$$

$$= \mathbf{p}^{-1}\mathbf{V}\mathbf{p}^{-1}\sigma^2$$

$$= \mathbf{p}^{-1}\mathbf{p}\,\mathbf{p}\mathbf{p}^{-1}\sigma^2 \qquad \text{por } 2.11.3$$

$$V(\mathbf{f}) = \mathbf{I}\sigma^2 \qquad 2.11.5$$

$\mathbf{f}$ es distribuida normalmente, es decir, $\mathbf{f} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, puesto que $\mathbf{f}$ consiste de combinaciones lineales de los elementos $\boldsymbol{\varepsilon}$, los cuales si son distribuidos normalmente.

Si premultiplicamos la ecuación 2.11.1 por $\mathbf{p}^{-1}$ obtenemos un nuevo modelo.

$$\mathbf{p}^{-1}\mathbf{Y} = \mathbf{p}^{-1}\mathbf{X}\,\boldsymbol{\beta} + \mathbf{p}^{-1}\boldsymbol{\varepsilon} \qquad 2.11.6$$

ó

$$\mathbf{z} = \mathbf{Q}\,\boldsymbol{\beta} + \mathbf{f} \qquad 2.11.7$$

Al comparar las dos últimas ecuaciones se tiene que

$$\mathbf{z} = \mathbf{p}^{-1}\mathbf{Y} \quad y \quad \mathbf{Q} = \mathbf{p}^{-1}\mathbf{X}.$$

Ahora es posible aplicar la teoría básica de mínimos cuadrados

a 2.11.7. La suma de cuadrados de residuos es

$$f'f = \varepsilon' p^{-1} . p^{-1} . \varepsilon = \varepsilon' (p.p)^{-1} \varepsilon = \varepsilon' . V^{-1} \varepsilon$$

$$f'f = \varepsilon' V^{-1} \varepsilon = (Y - X\beta)' V^{-1} (Y - X\beta) \qquad 2.11.8$$

A partir de la última igualdad se obtienen las ecuaciones normales, diferenciando con respecto a β, e igualando a cero, y posteriormente sustituyendo β por b.

$$ff' = (Y - X\beta)' V^{-1} (Y - X\beta)$$

$$= (Y' - \beta' X') V^{-1} (Y - X\beta)$$

$$= Y' V^{-1} Y - Y' V^{-1} X\beta - \beta' X' V^{-1} Y + \beta' X' V^{-1} X\beta$$

Derivando con respecto a β

$$0 = -Y' V^{-1} X + \beta' X' V^{-1} X$$

$$\beta' X' V^{-1} X = Y' V^{-1} X \quad \text{aplicando traspuesta.}$$

$$X' (\beta' X' V^{-1})' = X' (Y' V^{-1}) \quad , \quad \text{puesto que } (V^{-1})' = V^{-1}$$

tenemos

$$X' V^{-1} (\beta' X')' = X' V^{-1} Y$$

$$X' V^{-1} X \beta = X' V^{-1} Y \quad \text{de aquí} \quad X' V^{-1} Xb = X' V^{-1} Y \qquad 2.11.9$$

o sea propiamente $Q'Qb = Q'.z$

Puesto que $Q = p^{-1} X$ y $z = p^{-1} Y$

Tenemos $Q'Qb = Q'z$

$$X' p^{-1} p^{-1} X . b = X p^{-1} p^{-1} Y$$

$$X' (p.p)^{-1} Xb = X (p.p)^{-1} Y$$

$$X' V^{-1} Xb = X V^{-1} Y \qquad \text{por 2.11.3}$$

donde 2.11.9 tiene como solución $b = (X' V^{-1} X)^{-1} X' V^{-1} Y$ 2.11.10

siempre que la matriz invertida sea no singular.

La suma de cuadrados de la regresión es

$$\mathbf{b'Q'z} = \mathbf{Y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}Y} \qquad 2.11.11$$

Prueba por sustitución de $\mathbf{b'}$, $\mathbf{Q'}$ y $\mathbf{z}$, ya que

$\mathbf{b} = (\mathbf{X'V^{-1}b})^{-1}.\mathbf{X'V^{-1}Y}$ y su traspuesta es

$\mathbf{b'} = [\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}Y}]' = (\mathbf{X'V^{-1}Y})'\,[(\mathbf{X'V^{-1}X})^{-1}]' = \mathbf{Y'V^{-1}X(X'V^{-1}X}$

y $\mathbf{Q} = \mathbf{p^{-1}X}$ su traspuesta es $\mathbf{Q'} = \mathbf{X'p^{-1}}$, y $\mathbf{z} = \mathbf{p^{-1}Y}$

Ahora

$$\mathbf{b'Q'z} = \mathbf{Y'V^{-1}X(X'V^{-1}X)^{-1}X'p^{-1}p^{-1}Y}$$

$$\mathbf{b'Q'z} = \mathbf{Y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}Y}$$

La suma de cuadrados es

$$\mathbf{z'z} = \mathbf{Y'p^{-1}p^{-1}Y} = \mathbf{Y'V^{-1}Y} \qquad 2.11.12$$

La diferencia entre las ecuaciones 2.11.11 y 2.11.12 nos da la - suma de cuadrados de residuos. La suma de cuadrados debido a la media es $(\sum z_i)^2/n$, donde a $_i$ son los n elementos del vector $\mathbf{z}$

La matriz de varianza-covarianza de $\mathbf{b}$ es.

$$\mathbf{V(b)} = (\mathbf{Q'Q})^{-1}\sigma^2 = (\mathbf{X'V^{-1}X})^{-1}\sigma^2 \qquad 2.11.13$$

Una región de confianza para todos los parámetros puede ser obtenido de

$$(\mathbf{b}-\boldsymbol{\beta})\ \mathbf{Q'Q}(\mathbf{b}-\boldsymbol{\beta}) = \left[\frac{p}{n-p}\right]\ (\mathbf{z'z}-\mathbf{b'Q'z})\ F(p,n-p,1-\alpha) \qquad 2.11.14$$

tras de sustiuir de ecuaciones 2.11.11 y 2.11.12 y estableciendo que $\mathbf{Q} = \mathbf{p^{-1}X}$ si se desea.

Una aplicación simple de " asignar pesos en mínimos cuadrados " ocurre cuando las observaciones son independientes, pero tienen

diferentes varianzas tq.

$$\mathbf{V}\sigma^2 = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sigma_n^2 \end{bmatrix}$$

donde algunas de las σ_i^2 pueden ser iguales.

En problemas prácticos es a menudo dificultoso obtener información específica de la forma de $\mathbf{V}$ primero.

Por esta razón muchas veces se toma $\mathbf{V} = \mathbf{I}$ (sabiendo ser erroneo) y entonces se ensaya descubrir algo de $\mathbf{V}$ examinando los residuos del analisis de regresión.

Si una asignación de pesos de mínimos cuadrados debería realizarce, pero se hace un analisis ordinario de mínimos cuadrados - el estimador obtenido podría ser insesgado pero no tendría varianza mínima, ya que estimadores de mínima varianza se obtienen al asignar correctamente los pesos, en mínimos cuadrados.

Si el modelo de mínimos cuadrados es usado, entonces los estimadores son obtenidos de

$$\mathbf{b}_o = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad y$$

$$E(\mathbf{b}_o) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta \text{ pero,}$$

$$V(\mathbf{b}_o) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; V(\mathbf{Y})\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; \mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Recordamos de 2.11.13 que si el analisis correcto es llevado a cabo

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{V}\,\mathbf{X})^{-1}\sigma^2$$

cuya matriz nos proporciona las más pequeñas varianzas, ambas - para coeficientes individuales y para funciones lineales de los

coeficientes.

Un ejemplo de asignación de pesos en mínimos cuadrados.

Es para el modelo $E(y) = \beta x$ el cual deseamos ajustar

Por lo que supongamos que

$$V\sigma^2 = V(y) = \begin{bmatrix} 1/w_1 & & & 0 \\ & 1/w_2 & & \\ & & \ddots & \\ 0 & & & 1/w_n \end{bmatrix} \sigma^2 .$$

donde w's son los pesos a ser encontrados. Esto significa que

$$\mathbf{V}^{-1} = \begin{bmatrix} w_1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix}$$

Vamos ahora a encontrar **b** de

$$\mathbf{X'V^{-1}X\,b = X'V^{-1}Y} \qquad 2.11.15$$

$$(x_1, x_2 \ \ldots, x_n) \begin{bmatrix} w & & & \\ & \ddots & 0 & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mathbf{b} = (x_1, x_2, \ldots, x_n) \begin{bmatrix} w_1 & & & \\ & \ddots & 0 & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$(x_1 w_1, x_2 w_2, \ \ldots, \ x_n w_n \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \mathbf{b} = (x_1 w_1, \ \ldots, \ x_n w_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$(x_1^2 w_1 + {}^2_2 w_2 + \ldots + x_n^2 w_n)\,\mathbf{b} = (x_1 w_1 y_1 + x_2 w_2 y_2 + \quad + x_n w_n y_n)$$

$$\left(\sum_{i=1}^{n} w_i x_i^2\right) \mathbf{b} = \sum_{i=1}^{n} (w_i x_i y_i) \qquad 2.11.16$$

y se llega a

$$\mathbf{b} = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i x_i^2}$$

Caso 1

Supongamos que $\sigma_i^2 = V(y_i) = kx_i$, esto es, la varianza de los y_i es proporcional a la media de los correspondientes x_i. Entonces

$w_i = \frac{\sigma^2}{kx_i}$ que resulta de

$$\mathbf{V}\sigma^2 = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \sigma_2^2 & \\ \mathbf{0} & & \sigma^2 \end{bmatrix} = V(y) = \begin{bmatrix} \frac{1}{w_1} & & \mathbf{0} \\ & \frac{1}{w_2} & \\ \mathbf{0} & & \frac{1}{w_n} \end{bmatrix} \cdot \sigma^2 = \begin{bmatrix} \frac{\sigma^2}{w_1} & & \mathbf{0} \\ & \frac{\sigma^2}{w_2} & \\ \mathbf{0} & & \frac{\sigma^2}{w_n} \end{bmatrix}$$

Al comparar la primera matriz en el término a_{11} con el de la última matriz tenemos

$$\sigma_1^2 = \frac{\sigma^2}{w_1}, \quad \text{pero} \quad \sigma_1^2 = kx_1 \Rightarrow kx_1 = \frac{\sigma^2}{w^1} \Rightarrow w_1 = \frac{\sigma^2}{kx_1}$$

y asi para todos los términos de la diagonal. Entonces

$$b = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2} = \frac{\sum \frac{\sigma^2}{kx_i} x_i y_i}{\sum \frac{\sigma^2}{kx_i} x_i^2} = \frac{\frac{\sigma^2}{k} \cdot \sum y_i}{\frac{\sigma^2}{k} \sum x_i} = \frac{\sum y_i}{\sum x_i} = \frac{\frac{\sum y_i}{n}}{\frac{\sum x_i}{n}} = \frac{\bar{Y}}{\bar{X}}$$

Entonces, si la varianza de y_i es proporcional a x_i, el mejor estimador de los coeficientes de regresión es la media de los y_i dividido por la medida de los x_i. En resumen

$$V(b) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2 = (\sum w_i x_i^2)^{-1}\sigma^2 = \frac{\sigma^2}{\sum w_i x_i^2} \therefore V(b) = \frac{\sigma^2}{\sum w_i x_i^2}$$

además

$$V(b) = \frac{\sigma^2}{\sum w_i X_i^2} = \frac{\sigma^2}{\sum \frac{\sigma^2}{kx_i} X_i} = \frac{\sigma^2}{\frac{\sigma^2}{x} \sum} = \frac{x}{\sum x_i}$$

Caso 2

Supongamos $\sigma_i^2 = V(y_i) = kx_i^2$, esto es, la varianza de y_i es proporcional al cuadrado del correspondiente x_i, entonces $w_i = \sigma^2/kx_i^2$ por lo que :

$$\mathbf{b} = \frac{\Sigma w_i x_i y_i}{\Sigma w_i x_i} = \frac{\sum \frac{\sigma^2}{kx_i^2} \cdot x_i y_i}{\frac{\sigma^2}{kx_i} \cdot x_i} = \frac{\frac{\sigma^2}{k} \frac{y_i}{x_i}}{\frac{\sigma^2}{k} \sum 1} = \frac{y_i/x_i}{n}$$

Entonces si la varianza de los y_i es proporcional a x_i^2, el mayor estimador de los coeficientes de regresión es el promedio de las n pendientes obtenidas de cada par de observaciones - y_i/x_i, también

$$V(b) = \frac{\sigma^2}{\Sigma w_i x_i^2} = \frac{\sigma^2}{\sum \frac{\sigma^2}{kx_i^2} x_i^2} = \frac{\sigma^2}{\frac{\sigma^2}{k} \Sigma 1} = \frac{k}{n}$$

2.12 SESGO EN LOS ESTIMADORES DE REGRESION

Anteriormente afirmabamos que el estimador de mínimos cuadrados $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ de $\boldsymbol{\beta}$ en el modelo $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ es un estimador insesgado esto significa que

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

Esto es si se considera la distribución de **b** (tomando diferentes muestra de **Y**, con los **X** fijos y estimando $\boldsymbol{\beta}$ por cada muestra) entonces la media de esa distribución es $\boldsymbol{\beta}$.

Ahora si el modelo postulado no es correcto; esto es $E(\mathbf{b}) \neq \boldsymbol{\beta}$.

es decir el estimador es sesgado, donde el sesgo no solamente depende del modelo planteado y el modelo verdadero, sino tam-- bién en los valores de las variables **X** que entran en los cálcu los. Cuando un experimento proyectado es empleado, el sesgo de pende del proyecto experimental, como del modelo.

El modelo tratado será no singular, para trabajarlo en forma - matricial.

Supongamos el modelo planteado.

$$E(Y) = X_1\beta_1 \qquad 2.12.1$$

esto nos lleva a los estimadores de mínimos cuadrados

$$b_1 = (X_1'X)^{-1}X\,Y \qquad 2.12.2$$

donde X_1 es una matriz de constantes, podemos aplicar la esperanza a b_1 y a **Y** que son variables aleatorias.

$$E(b_1) = (X_1'X_1)^{-1}X_1'E(Y) = (X_1'\,X_1)^{-1}\,X_1'\,X_1.\beta_1 = \beta_1 \qquad 2.12.3$$

por lo que b_1 es un estimador insesgado de β_1

Ahora vamos a portular de nuevo el modelo 2.12.1
tq **b** como definido en 2.12.2, es el vector de estimadores de los coeficientes de regresión.

Supongamos ahora sin embargo, que la verdadera respuesta no es la ecuación 2.12.1 sino que

$$E(Y) = X_1\beta_1 + X_2\beta_2 \qquad 2.12.2$$

Esto es, existe un término $X_2\beta_2$, el cual no tomamos en cuenta en nuestro procedimiento de estimación.

Ahora sigue que

$$E(\mathbf{b}_1) = (\mathbf{X}_1'\mathbf{X}_1)\mathbf{X}_1' \mathbf{E}(\mathbf{Y}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2)$$

$$= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_{1\ 1} + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$$

$$= \beta_1 + \mathbf{A}\beta_2$$

donde $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$

Se debe ovservar que el sesgo no depende solamente del modelo - planteado y del modelo verdadero, si no también de las matrices

$\mathbf{X}_1$ y $\mathbf{X}_2$

Ejemplo 1 Supongamos postulamos el modelo

$E(Y) = \beta_o + \beta_1 X$, pero el modelo $E(Y) = \beta_o + \beta_1 X + \beta_{11} X^2$ es actualmente la verdadera función respuesta, desconoscamosla. Usemos las observaciones de Y de X = -1, 0 y 1 a estimar β_o y β_1 - en el modelo postulado, cuál será el sesgo introducido? esto es, será el estimador b_o y b_1 actualmente estimado?. El modelo verdadero, en términos de las observaciones es.

$$E(\mathbf{Y}) = E\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{matrix} \underline{1} \\ \\ \\ \end{matrix}\begin{matrix} \underline{X} \\ \\ \\ \end{matrix}\begin{matrix} \underline{X^2} \\ \\ \\ \end{matrix}$$

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

$$= \begin{matrix} \underline{1} & \underline{X} \\ \\ \\ \end{matrix}\begin{matrix} \\ \\ \\ \end{matrix}\begin{matrix} X^2 \\ \\ \\ \end{matrix}$$

$$= \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\beta_{11}$$

$$= \mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2$$

de esa forma se tiene de la ecuación 2.12.4 con ecuación 2.12.1 el modelo postulado, se sigue que

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \left(\begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

$$\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Entonces $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 1/3 & 0 \\ 0 & 1/2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 2/3 \\ 0 \end{bmatrix}$

Aplicando 2.12.5 obtenemos

$$E(\mathbf{b}_1) = E \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \boldsymbol{\beta}_1 + \mathbf{A}\beta_2 = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 2/3 \\ 0 \end{bmatrix} \beta_{11} = \begin{bmatrix} \beta_o + \frac{2}{3}\beta_{11} \\ \beta_1 \end{bmatrix}$$

De donde

$E(b_o) = \beta_o + \frac{2}{3}\beta_{11}$

$E(b_1) = \beta_1$ entonces b_o es sesgado por $\frac{2}{3}\beta_{11}$, y β_1 es insesgado.

Ejemplo 2

Supongamos que el modelo postulado es $E(Y) = \beta_o + \beta_1 X$, pero el modelo real es

$$E(Y) = \beta_o + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3$$

Que sesgos son introducidos tomanda las observaciones de

$X = -3, -2, -1, 0, 1, 2, 3.$

Encontramos

$$E(Y) = E\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{array}{c} \begin{matrix} \underline{1} & \underline{X} & \underline{X^2} & X^3 \end{matrix} \\ \begin{bmatrix} 1 & -3 & 9 & -27 \\ 1 & -2 & 4 & -6 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & -8 \\ 1 & 3 & 9 & -27 \end{bmatrix} \end{array} \cdot \begin{bmatrix} \beta \\ \beta_1 \\ \beta_{11} \\ \beta_{111} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 27 \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{111} \end{bmatrix} = X_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$$

Por lo que

$$\mathbf{X}_1 = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} ; \mathbf{X}_2 = \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 9 \end{bmatrix}$$

Encontremos

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \left\{ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right\}^{-1} = \begin{bmatrix} 7 & 0 \\ 0 & 28 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{7} & 0 \\ 0 & \frac{1}{28} \end{bmatrix}$$

$$\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & -1 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 27 \end{bmatrix} = \begin{bmatrix} 28 & 0 \\ 0 & 196 \end{bmatrix}$$

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} \frac{1}{7} & 0 \\ 0 & 1/28 \end{bmatrix} \begin{bmatrix} 28 & 0 \\ 0 & 196 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 7 \end{bmatrix}$$

Entonces

$$E(\mathbf{b}_1) = E \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \boldsymbol{\beta}_1 + A\boldsymbol{\beta}_2 = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 4 & 0 \\ 0 & 7 \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{11} \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 4\beta_{11} \\ 7\beta_{111} \end{bmatrix} = \begin{bmatrix} \beta_0 + 4\beta_1 \\ \beta_1 + 7\beta_{111} \end{bmatrix}$$

ó $E(b_o) = \beta_o + 4\beta_1$, $E(b\ ) = \beta_1 + 7\beta_{111}$

Usando la fórmula general 2.12.5 podemos encontrar el sesgo en algún estimador una vez postulado el modelo.

EL EFECTO DE SESGO EN EL ANALISIS DE MINIMOS CUADRADOS

Nota. Para esta parte llamaremos $\mathbf{X}$ a la matriz $\mathbf{X}_1$, y $\boldsymbol{\beta}$ al vector llamado $\boldsymbol{\beta}_1$, a $\mathbf{X}_2\boldsymbol{\beta}_2$ será el término extra del modelo verdadero. Ahora resumiremos el efecto del sesgo en el analisis -- usual de mínimos cuadrados.

Supongamos que

1. El modelo postulado $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ conteniendo p parámetros; $V(\mathbf{Y}) = \mathbf{I}\sigma^2$.
2. El modelo verdadero es $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2$, donde $\boldsymbol{\beta}_2$ puede ser $\mathbf{0}$, si el modelo es correcto.
3. El número total de observaciones es n y hay f grados de libertad disponible por falta de ajuste y **e** grados de libertad para error puro, tq $n = p+f+e$. Esto significa que hay - $p + f$ puntos distintos en el diseño.
4. El estimador $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ y el valor ajustado $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$ son - obtenidos de manera usual.
5. $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2$

Entonces el número de resultados son verdaderos como

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	VALOR ESPERADO DE CUADRADOS MEDIOS
Estimadores	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	p	$\sigma^2 + (\beta + \mathbf{A}\beta_2)'\mathbf{X}\mathbf{X}(\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\beta}_2)/p$
Falta de ajuste por diferencia		f	$\sigma^2 + \boldsymbol{\beta}_2'(\mathbf{X}_2 - \mathbf{X}\mathbf{A})'(\mathbf{X}_2 - \mathbf{X}\mathbf{A})\boldsymbol{\beta}_2/f$
Error puto	es_e^2	e	σ^2
TOTAL	$\mathbf{Y}'\mathbf{Y}$	n	

5. Cuando $\beta_2 = 0$, esto es, cuando el modelo postulado es correcto, los anteriores se reducen a lo siguiente:

$$E(\mathbf{b}) = \boldsymbol{\beta} \quad ; \quad \mathbf{E}(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}$$

$$E(\text{cuadrados medios debido a estimadores}) = \sigma^2 + \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{p}$$

(Esto es por que la media de cuadrados debido al estimador es comparado con un estimador de σ^2 a la prueba H_o: $\beta_o = 0$, tiene esperanza σ^2, si $\boldsymbol{\beta} = 0$)

$$E(\text{cuadrados medios por falta de ajuste}) = \sigma^2$$

(Esto es porque el modelo es correcto, lo que nos proporciona un estimador valido de σ^2. Si la falta de ajuste existe, el cuadrado medio de falta de ajuste tiene una esperanza más grande que σ^2)

C A P Í T U L O I I I

RECIDUOS

3.0 INTRODUCCION

Nota. El trabajo de este capítulo es útil y no válido solamente para modelos de regresión lineal, sino también para modelos no lineales y analisis de varianza. Se trabajará siempre que un modelo sea ajustado y presente variación inexplicada (en forma de residuos).

Los residuos son definidos por las n diferencias $e_i = Y_i - \hat{Y}_i$, $i = 1,2,\ldots, n.$; donde Y_i es una observación y $\hat{Y}_i$ es el valor encontrado por la ecuación ajustada. Es decir, lo que la ecuación de regresión no ha sido capaz de explicar. Así podemos decir, que los e_i son errores, para el modelo planteado; anteriormente a los e_i se les ha hecho ciertas suposiciones generales, que son independientes, tienen media cero, una varianza constate σ^2, una distribución normal y una para hacer una prueba F. Así, si el modelo es correcto los residuos mostrarían tendencias a confirmar esas supociciones planteadas, o al mejor no exhibir una negativa a ellas. Podríamos preguntarnos ¿pueden los residuos hacer aparecer nuestras suposiciones como equivocadas?.

Después de examinar los residuos seremos capaces de responder si:

1. Las suposiciones parecen ser violadas
2. Las suposiciones parecen no ser violadas

Es de observar 2, no significa que estemos aceptando que las suposiciones son correctas, sino que en base a los datos, no se tiene razón para decir que son incorrectas.

Para examinar los residuos, se hará por medio de gráficos, fáciles de hacer, pero son reveladores cuando las suposiciones - son violadas. Los principales métodos para plotear los residuos e_i son:

1. Plotearlos todos
2. Plotear una sucesión de tiempo, si el órden es conocido.
3. Plotear contra los valores ajustado $\hat{Y}_i$
4. Plotear contra las varialbes X_{ji}; para $j = 1,2,\ldots, k$.

En adición a ese ploteo básico, los residuos deberían también ploteados.

5. En alguna forma razonalbe, para problemas particulares bajo consideración.

Explicación de los ploteos mediante el siguiente ejemplo

Ejemplo. Un análisis de regresión proporciona once residuos, - $e_1, e_2, \ldots, e_{11}$ con valores 5, -2, -4, 4, 0, -6, 9, -2,-5,3,-2. respectivamente.

3.1 PLOTEARLOS TODOS.

Cuando los residuos del ejemplo son ploteados todos obtenemos el diagrama en la figura 3.1

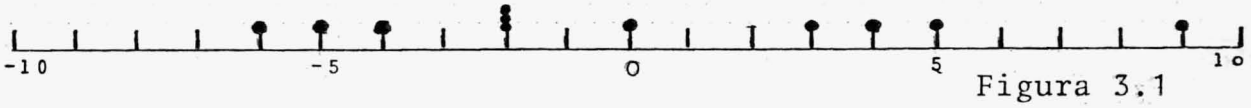


Figura 3.1

Si nuestro modelo es correcto, el diagrama de esos once resi-- duos deberían asemejarse a una distribución normal con media cero ¿contradice nuestra idea plotear todos los residuos?.

Primero, notamos que la media de los residuos es cero, que se obtiene de la primera ecuación normal, 1.2.8, derivándola con -

respecto a β_o, cuando el modelo es ajustado, así si

$$E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

la ecuación normal, puede ser escrita

$$-2\Sigma(Y_i - b_o - b_1 X_{1i} - \ldots - b_k X_{ki}) = 0$$

donde la suma es tomada de $i = 1$ a n, la cual se reduce a

$$\Sigma(Y_i - \hat{Y}_i) = 0; \text{ así } \Sigma e_i = 0$$

Así. $\frac{\Sigma e_i}{n} = \bar{e} = 0$

Mientras que el ploteo exhibe pequeñas irregularidades, no parece anormal para una muestra de once observaciones de una distribución normal ¿como podría justificarse?.

Comparandola con diagramas obtenidos de varias muestra de tamaño once, de una tabla de desviación normal aleatoria.

Así una pequeña práctica de estas ordenaciones, nos proporciona una excelente " forma " de como un gráfico anormal podría verse antes y poder asegurar que contradice la suposición de normalidad.

Las " desviaciones normales unitarias " forma de los residuos.

Usualmente asumimos que $\varepsilon_i \sim N(0,\sigma^2)$ tq $\frac{\varepsilon_i}{\sigma} \sim N(0,1)$

Ahora si el modelo es correcto los cuadrados medios de los residuos

$$s^2 = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

es un estimador de σ^2. (Nota: si ignoramos el error de redondeo $\bar{e} = 0$). La cantidad e_i/s es llamada a menudo forma de las " desviaciones normales unitarias " de los residuos e_i.

Los e_i/s, $i = 1,2,\ldots, n$; puede ser examinados en un ploteo de todos, para ver si ellos muestran que la suposición $\varepsilon_i/\alpha \sim N(0,1)$ es equivocada. Puesto que un 95% de la distribución N(0,1) da como límites (-1.96, 1.96), podríamos esperar aproximadamente el 95% de los e_i/s esten entre los límites (-2,2). Si n-p es pequeño, los límites al 95% pueden ser usados en una distribución t(n-p)

3.2 PLOTEO DE UNA SUCESION DE TIEMPO

Asumimos que los residuos en el ejemplo anterior ocurieron en el órden en un tiempo dado. El ploteo sería como se muestra en la figura 3.2.

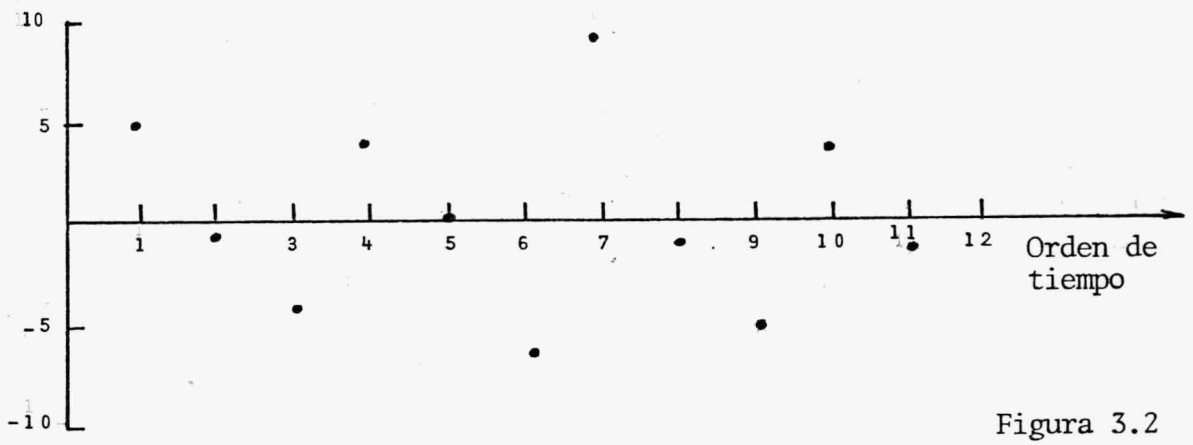


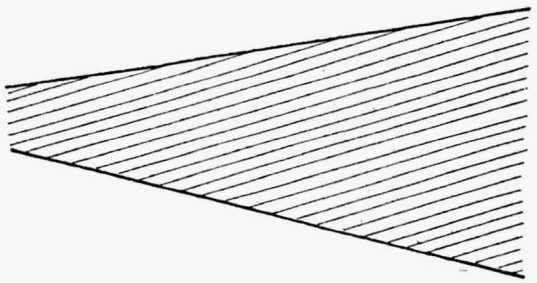
Figura 3.2

Si luego de ploteado " damos pasos atrás ", de este diagrama obtenemos la impresión de una " banda " de residuos tal como se muestra en la figura 3.3.

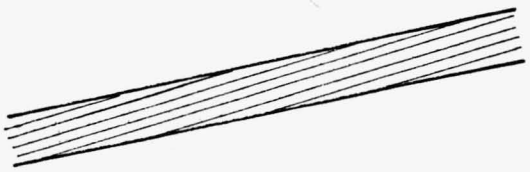
Figura 3.3

Esto indica que en un término largo de tiempo no hay efectos sobre los datos. Sin embargo al " dar pasos atrás " se pueden asemejar a las siguientes gráficas de la figura 3.4

(1)



(2)



(3)

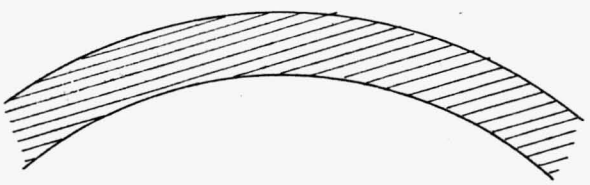


Figura 3.4

donde se hace constar el efecto del tiempo, como sigue

1) La varianza no es constante, pero se incrementa con el tiempo, indicado con esto que una asignación de pesos por mínimos cuadrados podía haberse realizado.

2) Un término lineal en el tiempo tendría que ser incluido en el modelo

3) Términos lineales y cuadráticos en el tiempo tendrían que ser incluidos en el modelo.

Se pueden dar combinaciones o variaciones (como en el caso (2).

donde la inclinación podría ser opuesta)

Mencionamos anteriormente que los residuos revelan tendencias no aparentes, en períodos largos de tiempo.

Una observación más cercana del ploteo en el tiempo demuestra que la observación no puede ser extendida a períodos cortos de dirección de tiempo.

Si tomamos los residuos en grupos de tres, es decir,

(1,2,3), (4,5,6), (7,8,9), (10,11).; podemos observar una línea de tendencia hacia abajo en cada grupo, revelando una clase de ordenamiento por zonas , las cuales pueden ser introducidas en el modelo revisado. Si tentativamente asumimos una inclinación general, para la tendencia lineal pudierámos, por ejemplo, adicionar un término de la forma $\delta\{(t-1)\text{modulo } 3\}$ al modelo, donde δ es un coeficiente de regresión a ser estímado y (t-1) módulo 3 es la variable obtenida por el residuo que resulta despues de dividir el valor de t-1 por 3.

Los valores de las nuevas variables son como siguen

Variable vieja t	1	2	3	4	5	6	7	8	9	10	11
(t-1) Modulo 3	0	1	2	0	1	2	0	1	2	0	1

(Otra alternativa es usar {(t-1) modulo 3-1} como una variable esta nos da los valores -1,0 y 1 en lugar de 0,1 y 2.).

3.3. PLOTEO CONTRA $\hat{Y}_i$

Asumimos que los $\hat{Y}_i$, los cuales corresponden a los e_i dados anteriormente fueron 44, 8, 10, 62, 22, 48, 56, 30, 24, 16, 34. Entonces al plotearlos obtenemos la siguiente figura. 3.5

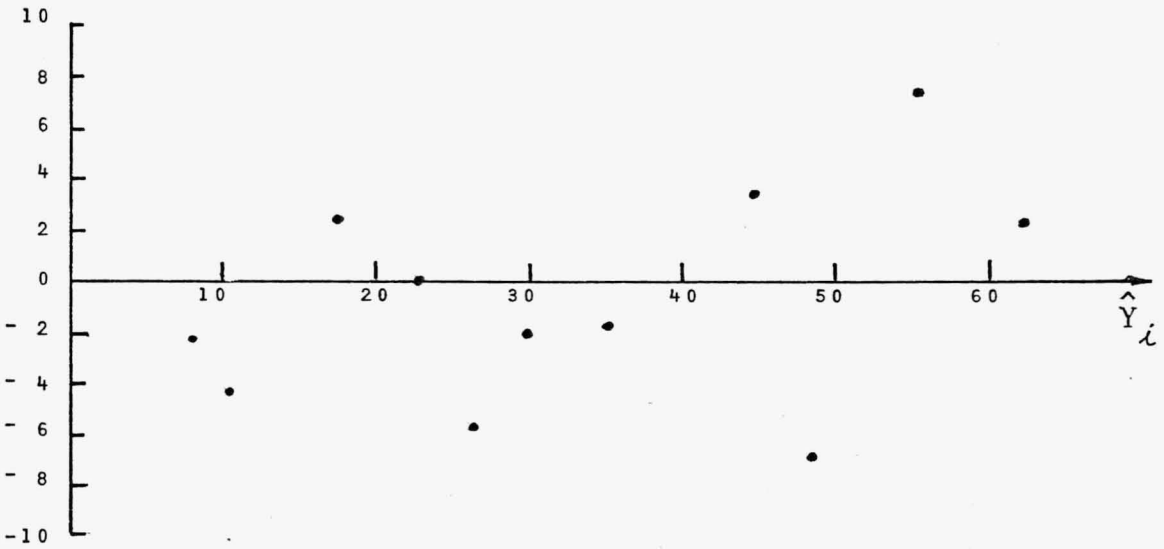


Figura 3.5

La banda horizontal indica la no anormalidad y nuestro analisis de mínimos cuadrados no parecería ser invalidado.

La anormalidad sería indicada por el ploteo de las formas mostradas en la Figura 3.4 para (1), (2) y (3); aquí el ploteo indicaría:

1) Varianza no constante, como se asumió: necesitamos una asignación de pesos por mínimos cuadrados o una transformación en las observaciones Y_i, antes de hacer un analisis de regresión.

2) Error en el analisis; la desviación de la ecuación ajustada en residuos negativos, corresponde a los $\hat{Y}$ bajos, residuos - positivos para valores altos de $\hat{Y}$. El efecto puede ser causado por la omisión equivocada del término β_o en el modelo.

3) Modelo inadecuado: necesita términos extra en el modelo -

(es decir, un cuadrado o producto mezclados) o la necesidad de una transformación en las observaciones Y_i, antes del analisis.

3.4 PLOTEO CONTRA LAS VARIABLES INDEPENDIENTES X_{ji}; $i = 1,2,...n$.

La forma de ese ploteo es el mismo que contra los $\hat{Y}_i$, excepto que usamos (en lugar de los correspondientes valores de $\hat{Y}_i$) -- los valores de los correspondientes X_{ji}; llamandolas X_{j1}, X_{j2} ..., X_{jn}. El gráfico correspondiente de una banda horizontal - de residuos, es considerado como satisfactoria. Las anomalías - en la figura 3.4 indican aquí.

1) Varianza no constante, necesidad de una asignación de pesos por mínimos cuadrados o una fransformación prelimirar en - las yes.
2) Error en cálculos, efecto lineal de X_j no ha sido removido
3) Necesidad de términos extras, por ejemplo, un término cua-- drático X_j en el modelo o una transformación en los yes.

En problemas pequeños de regresión, es decir 2 ó 3 variables - es posible hacer un diagrama del modelo ya sea en 2 ó 3 dimen- ciones colocando cerca de los puntos, los residuos correspon-- dientes, siempre que sea posible.

3.5 PLOTEO DE OTROS RESIDUOS

Con un buen conocimiento sobre el problema, pueden plantearse otros típos de diagramas, como por ejemplo, sugongamos que las once observaciones las cuales nos llevarón a los once residuos anteriores, provienen de tres máquinas A, B, C tq. los residuos agrupados por máquina fueron:

A: -1, -4, -6; B: -2, -5, -2; C: 5, 4, 0, 9, 3

que se muestran en el diagrama contra máquinas

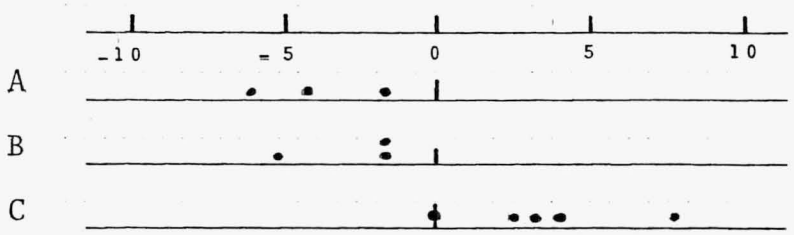


Figura 3.6

que sugiere que hay una diferencia básica en el nivel de res-- puesta Y de la máquina C con las máquinas A y B.

Tal diferencia podría introducirse dentro del modelo con una - variable falsa (se verá posteriormente).

3.6 ESTADISTICOS PARA EXAMINAR RESIDUOS

En las secciones anteriores, mediante diagramas se han utiliza do técnicas visuales para poder determinar las violaciones a - las suposiciones hechas a la regresión, pero no se tenía una - medida número rica del defecto y lo cual es necesario, ya que se tuvieron tres tipos de discrepancias, pudiéndose medir esos de fectos por un estadistico apropiado, por lo que se define

$$T_{pq} = \sum_{i=1}^{n} e_i^p \hat{Y}_i^q$$

entonces

1) $T_{21} = \sum_{i=1}^{n} e_i^2 \hat{Y}_i$ provee una medida para el tipo de defecto de la figura 3.4 (1)

2) $T_{11} = \sum_{i=1}^{n} e_i \hat{Y}_i$ esto siempre sería cero. Esto provee una medi da para el defecto de la figura 3.4 (2)

3) $T_{12} = \sum_{i=1}^{n} e_i \hat{Y}_i^2$ provee una medida para el defecto de la figura

3.4 (3)

3.7 CORRELACION ENTRE RESIDUOS

En una situación general de regresión, cuando p parámetros son estimados de n observaciones, los n residuos son asociados con n-p grados de libertad. Entonces los residuos pueden no ser independientes y existir correlación entre ellos. Si el modelo - postulado es $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, y si $\mathbf{X'X}$ es no singular, entonces los residuos pueden ser escritos en forma matricial así.

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb} = \mathbf{Y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y}$$

$$= [\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}]\,\mathbf{Y}$$

$$= [\mathbf{I} - \mathbf{R}]\,\mathbf{Y}$$

donde $\mathbf{R} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, la cual es importante en regresión.

Puesto que $E(\mathbf{Y}) = \mathbf{X}\beta$, tenemos que

$$\mathbf{e} - \mathbf{E}(\mathbf{e}) = [\mathbf{I} - \mathbf{R}]\,\mathbf{Y} - \mathbf{E}[(\mathbf{I} - \mathbf{R})]\,\mathbf{Y}$$

$$= [\mathbf{I} - \mathbf{R}]\,\mathbf{Y} - (\mathbf{I} - \mathbf{R})\,\mathbf{E}(\mathbf{Y})$$

$$= (\mathbf{I} - \mathbf{R})\mathbf{Y} - (\mathbf{I} - \mathbf{R})\mathbf{X}\beta$$

$$= (\mathbf{I} - \mathbf{R})(\mathbf{Y} - \mathbf{X}\beta)$$

$$= (\mathbf{I} - \mathbf{R})\,\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \text{ del modelo postulado.}$$

y la matriz de varianza-covarianza de $\mathbf{e}$ es definida por

$$V(\mathbf{e}) = E\{[\mathbf{e} - E(\mathbf{e})]\,[\mathbf{e} - E(\mathbf{e})]'\}$$

$$= E\{[(\mathbf{I} - \mathbf{R})\,\boldsymbol{\varepsilon}]\,[(\mathbf{I} - \mathbf{R})\boldsymbol{\varepsilon}]'\}$$

$$= E\{(\mathbf{I} - \mathbf{R})\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'[(\mathbf{I} - \mathbf{R})]'\}$$

Pero $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$, si $E(\boldsymbol{\varepsilon}) = 0$, como se mencionó anterior

ormente. Además

$$(\mathbf{I} - R)' = I' - R' = I - \mathbf{R}' = I - \mathbf{R}$$

puesto que

$$\mathbb{R}' = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\right]' = (\mathbf{X}')' \left[\mathbf{X}.(X'X)^{-1}\right]'$$

$$= \mathbf{X}.((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \mathbf{X}'$$

$$= \mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= \mathbf{R}$$

es decir es simétrica.

Asi

$$V(\mathbf{e}) = (\mathbf{I} - \mathbf{R})\, E(\varepsilon\varepsilon')(\mathbf{I} - \mathbf{R})'$$

$$= (\mathbf{I} - \mathbf{R})\, \mathbf{I}\sigma^2\, (\mathbf{I} - \mathbf{R})$$

$$= (\mathbf{I} - \mathbf{R})\, (\mathbf{I} - \mathbf{R})\sigma^2$$

$$= (\mathbf{I}.\mathbf{I} - \mathbf{I}\mathbf{R} - \mathbf{R}\mathbf{I} + \mathbf{R}\mathbf{R})\, \sigma^2$$

de aquí

$$\mathbf{I}.\mathbf{I} = \mathbf{I}^2 = \mathbf{I} \quad ; \quad \mathbf{I}.\mathbf{R} = \mathbf{R} \quad ; \quad \mathbf{R}.\mathbf{I} = \mathbf{R} \quad y$$

$$\mathbf{R}.\mathbf{R} = \mathbf{R}^2 = \left[\mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}.\mathbf{X}'\right] \left[\mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$$

$$= \mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}.(X'X)^{-1}\mathbf{X}'$$

$$= \mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}).(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= \mathbf{X}.\mathbf{I}.(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= \mathbf{X}.(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$= \mathbf{R} \quad , \text{ por definición de } \mathbf{R}$$

aplicando estas igualdades tenemos

$V(\mathbf{e}) = (\mathbf{I} - \mathbf{R} + \mathbf{R} + \mathbf{R})\ \sigma^2$

$V(\mathbf{e}) = (\mathbf{I} - \mathbf{R})\ \sigma^2$

Entonces la $V(e_i)$ es dado por el iésimo elemento de la diagonal, y $cov(e_i, e_j)$ es dado por el (i,j) -ésimo elemento de la matriz $(\mathbf{I} - \mathbf{R})\ \sigma^2$. la correlación de e_i y e_j es dada por

$$\rho_{ij} = \frac{covar(e_i,\ e_j)}{\{V(e_i).V(e_i)\}^{1/2}}$$

El valor de esta correlación depende solamente de los elemen-- tos de la matriz **X**.

En el analisis de regresión, el efecto de correlación entre residuos no necesita ser considerado, cuando el diagrama esta hecho, excepto cuando la razón (n-p)/n es bastante pequeña. (donde (n-p) es número de grados de libertad de los residuos y n es el número de grados de libertad)

3.8 PUNTOS RAROS

Un punto raro entre residuos, es aquel cuyo valor absoluto es mayor que el resto de los residuos, es muy peculiar, e indica un punto que no es típico del resto de los datos, asi un punto raro debe ser sometido particualrmente a un examen cuidadoso, para ver si la razón de su peculiaridad puede ser determinada.

Se han propuesto reglas para rechazar puntos raros (es decir, - para remover las correspondientes observaciones de los datos, - despues de lo cual los datos son reanalizados sin esas observaciones), pero rechazarlos no siempre es un proceso correcto, - puesto que algunas veces esta proporcionando información que no pueden dar otros puntos, debido al hecho que pueden originarse

de una combinación de circunstancias inusuales, que pueden ser de vital interés y requieren una investigación especial, más - que un rechazo.

Como regla general, un punto raro de que son errores en las observaciones o manejo en una máquina.

3.9 EXAMINADO CORRIDAS EN UN DIAGRMA DE UNA SUCESION DE TIEMPO DE RESIDUOS.

Cuando la sucesión de tiempo de un conjunto de residuos es conocida, pudiendo ser un patrón inusual que se presente en grupos de residuos positivos y negativos, por ejemplo, tomando un caso extremo, si se tienen 30 residuos en sucesión de tiempo, que consiste de 10 negativos, seguidos de veinte residuos positivos, podemos sospechar que una variable no considerada ha -- cambiado los niveles entre el décimo y undécimo dato. Buscaríamos una causa razonable para este comportamiento. Cuando hay - una sucesión de tiempo de tales datos, es útil tener un método, para hacer posible una discusión de la anormalidad del patrón.

Supongamos que tenemos una sucesión de signos como sigue

+ + - + - - - - + + - + + +

donde esos signos representan a los residuos en sucesión de - tiempo, pero " el más o menos " podrían denotar "macho y hembra" " mejor y peor ", "tratamiento A y tratamiento B " o dos niveles de una clasificación dicotómica.

Supongamos que hay n signos por todos, de los cuales n_1 son signos más y n_2 son signos menos y hay u corridas. En el ejemplo anterior $n_1 = 8$, $n_1 = 8$ y hay 7 corridas, indicadas en los pa--

entesis de abajo

(+ +) (-)(+) (- - - -)(+ +)(-)(+ + + +).

Será este arreglo único?

Para esto plantiemos el siguiente ejemplo, si hay 6 signos dos de los cuales son positivos, cuantos arreglos son posibles.

ARREGLOS	NUMEROS DE CORRIDAS
+ + - - - -	2
+ - + - - -	4
+ - - + - -	4
+ - - - + -	4
+ - - - - +	3
- + + - - -	3
- + - + - -	5
- + - - + -	5
- + - - - +	4
- - + + - -	3
- - + - + -	5
- - + - - +	4
- - - + + -	3
- - - + - +	4
- - - - + +	2

La distribución de corridas es como sigue

u	2	3	4	5	
Frecuencia	2	4	6	3	total 15
P(u)	2/15	4/15	6/15	3/15	
Prob. Acumulada	0.133	0.400	0.800	1.000	

Así cuando u = 5 ocurren 3/15 ó 20% de los casos poribles, o con una probabilidad de 0.2. Si obervamos u = 2 en un conjunto de seis residuos, de los cuales dos fueron positivos, entonces al ser observado el evento ocurre con una probabilidad de 0.133.

Para alguna sucesión dada de signos, podemos encontrar la probabilidad que el valor observado u (o un valor más pequeño) ocurrirá ejemplo , cuando $n_1 = 2$, $n_2 = 4$ y $u = 3$); prob ($u \leq 3$) = prob($u = 2$) + prob($u = 3$) = $\frac{4}{15} = 0.4$, y un evento no raro, ocurrirá en un 40% de las veces. En base a tal nivel de probabilidad, podemos decir si el arreglo ha ocurrido o no ha ocurrido.

Podríamos por ejemplo, comparar la probabilidad con un valor preasignado, es decir $\alpha = 0.05$ y rechazar la idea de un arreglo aleatorio si prob($u \leq$ u observado) ≤ 0.5). Cuando $n_1 > 10$ y $n_2 > 10$ valores exactos no se necesitan, puesto que una distribución normal (aproximada) la cual proporciona una precisión satisfactoria.

haciendo

$$u = \frac{2n_1n_2}{n_1+n_2} + 1, \qquad \sigma^2 = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)} \qquad 3.9.1$$

Entonces se define aproximadamente

$$z = \frac{(u-\mu+1/2)}{\sigma} \qquad 3.9.2$$

Es una desviación normal unitaria, donde 1/2 es la corrección para la continuidad de la variable discreta.

Nota. Cuando $n_1 \leq 10$ y $n_1 \leq n_2 \leq 10$ es posible obtener su distribución de la misma forma del último ejemplo.

Ejemplo. Examinado un conjunto de 27 residuos, 15 de los cuales

tienen un signo y 12 signo opuesto, arreglados en sucesión de - tiempo, revela u = 7.

¿Cuántos arreglos parecen ser aleatorios?

$n_1 = 15, \quad n_2 = 12, \quad u = 7$

$$u = \frac{(2)(S)(2)}{15 + 12} + 1 = \frac{43}{3}$$

$$\sigma^2 = \frac{2(15)(12)(2x15x12-15-12)}{(12+15)^2(15+12-1)} = \frac{740}{117}$$

obtenidos al aplicar 3.9.1

el valor de z a partir de 3.9.2 es

$$z = \frac{(7 - \frac{43}{3} + 1/2)}{(740/116)^{1/2}} = -2.713$$

La probalidad de obtener una desviación normal unitaria de valor -2.713 ó más pequeña, es 0.0033 (ó 0.33%) de modo que un número bajo de corridas inusuales parecen haber ocurrido.

Rechazamos la idea que el arreglo de signos es aleatoria. El modelo sería sospechoso y se tendría que examinar el patrón de residuos para ver la causa.

Nota. La prueba anterior es aplicable cuando las ocurrencias -- producen el patrón de corridas son independientes. Para el cálculo de la probabilidad es por la tabla normal.

C A P Í T U L O I V

DOS VARIABLES INDEPENDIENTES

4.0 INTRODUCCION

Anteriormente se ha tratado el modelo de regresión lineal de - primer órden en una variable X

$$Y = \beta_o + \beta_1 X + \varepsilon$$

y demostramos como el análisis de línea recta puede ser expresado en forma matricial. Pudiéndose extender a más variables - por medio de la matriz de aproximación, así, podemos aplicarlo al modelo de primer órden:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Para lo cual seguiremos con el ejemplo del capítulo 1, aplicando una nueva variable X_6, al modelo que tenía la variable X_8; por lo que el modelo será escrito

$$Y = \beta_o X_o + \beta_8 X_8 + \beta_6 X_6 + \varepsilon \qquad 4.0.1$$

donde

Y = respuesta ó números de libras de vapor por mes

X_o= variable falsa, cuyo valor es siempre 1

X_8= promedio de temperatura atmosférica por mes (°F)

X_6= número de días operados en el mes.

Las siguientes matrices pueden ser construidas así

$$\mathbf{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ 8.40 \\ \vdots \\ 10.36 \\ 11.08 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ 1 & 58.8 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 33.4 & 20 \\ 1 & 28.6 & 22 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_8 \\ \beta_6 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}$$

donde

$\mathbf{Y}$ es un vector 25 x 1

$\mathbf{X}$ es una matriz 25 x 3

$\boldsymbol{\beta}$ es un vector 3 x 1

$\boldsymbol{\varepsilon}$ es un vector 25 x 1

Usando resultados anteriores, por mínimos cuadrados obtenemos a $\mathbf{b}$ estimador de $\boldsymbol{\beta}$ dado por

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad , \quad \text{con } \mathbf{X}'\mathbf{X} \text{ es matriz no singular}$$

entonces

$$\mathbf{b} = \begin{bmatrix} b_o \\ b_8 \\ b_6 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & 1 & 1 \\ 35.3 & 29.7 & 30.8 & 28.6 \\ 20 & 20 & 23 & 22 \end{bmatrix} \begin{bmatrix} 1 & 35.3 & 25 \\ 1 & 29.7 & 20 \\ \vdots & & \\ 1 & 28.6 & 22 \end{bmatrix} \right\}^{-1} .$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 35.3 & 29.7 & 28.6 \\ 20 & 20 & 20 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ \vdots \\ 11.08 \end{bmatrix}$$

Multiplicando matricialmente

$$\begin{bmatrix} b_o \\ b_8 \\ b_6 \end{bmatrix} = \begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 26353.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 35.3 & 29.7 & 28.6 \\ 20 & 20 & 22 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ \vdots \\ 11.08 \end{bmatrix}$$

$$\begin{bmatrix} b_o \\ b_8 \\ b_6 \end{bmatrix} = \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ 0.011242 & 0.146207 \times 10\text{-}3 & 0.175467 \times 10\text{-}3 \\ -0.106098 & 0.175467 \times 10\text{-}3 & 0.478599 \times 10\text{-}2 \end{bmatrix} \begin{bmatrix} 235.60 \\ 11821.432 \\ 4831.86 \end{bmatrix} \quad 4.0.2$$

$$\begin{bmatrix} b_o \\ b_8 \\ b_6 \end{bmatrix} = \begin{bmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{bmatrix}$$

entonces la ecuación ajustada por mínimos cuadrados

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6$$

Recordemos la forma algebraica de las ecuaciones normales para dos variables independientes:

$$b_o n + b_1 \Sigma X_{1i} + b_2 \Sigma X_{2i} = \Sigma Y_i$$

$$b_o \Sigma X_{1i} + b_1 \Sigma X_{1i}^2 + b_2 \Sigma X_{1i} X_{2i} = \Sigma X_{1i} Y_i$$

$$b_o \Sigma X_{2i} + b_1 \Sigma X_{2i} X_{1i} + b_2 \Sigma X_{2i}^2 = \Sigma X_{2i} Y_i$$

Para la ecuación ajustada $\hat{Y} = b_o + b_1 X_{1i} + b_2 X_{2i}$.

La ecuación anterior ajustada, actualmente es posible obtener - la misma ecuación a traves de una serie simple de lineas rectas, lo cual es lo que se hará.

4.1 REGRESION MULTIPLE CON DOS VARIABLES INDEPENDIENTES COMO UNA SUCESION DE LINEAS RECTAS DE REGRESION.

Pasos para poderlo realizar

1. Plotear Y (cantidad de vapor) contra X_8 (promedio de temperatura atmosférica)

   Que se obtiene como resultado: si la temperatura sube, entonces la necesidad de vapor disminuye.

2. Regresión de Y contra X_8. Está linea fué ajustada en el cap. 1, (1.2.14) nos dió como ecuación

$$\hat{Y} = 13.6215 - 0.0798\ X_8$$

   Esta ecuación no es buena predictora de Y. Adicionando una nueva variable, X_6 (# de días operados) la ecuación de predicción podría mejorar significativamente la predicción, así se podría relacionar, el número de días operados con la variación no explicada de los datos, después de que los efectos de la temperatura atmosférica hayan sido removidos.

3. Regresión de X_6 contra X_8; calculando los residuos $X_{6i} - \hat{X}_{6i}$, $i = 1,2, \ldots, n$; el ploteo de X_6 contra X_8 está en figura 4.1

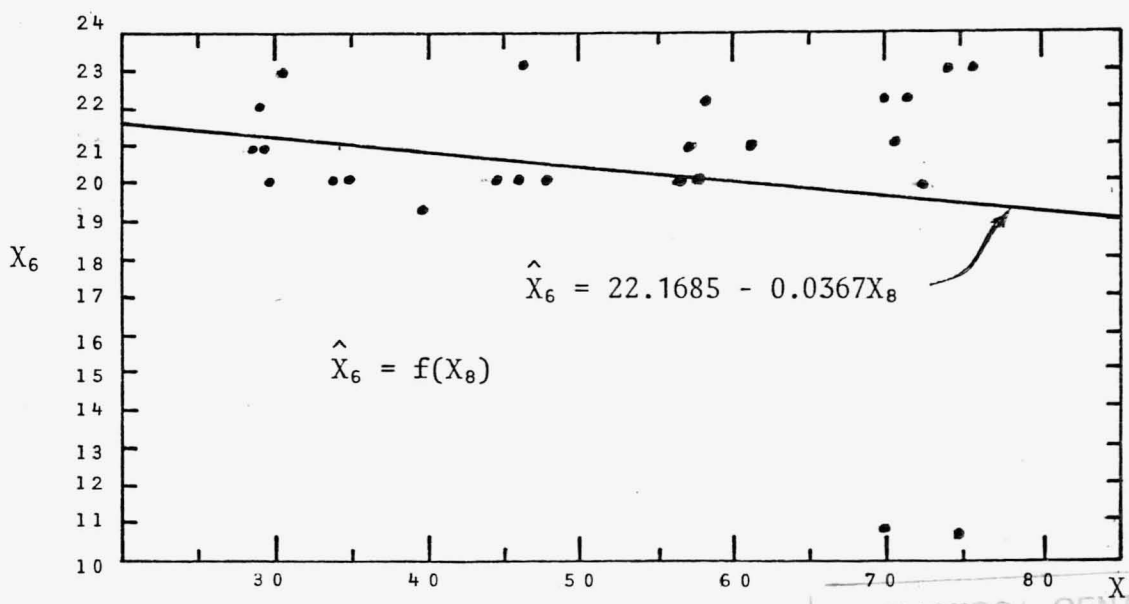


Figura 4.1

Calcularemos b por el método matricial que viene dado por

$$\begin{bmatrix} b_o \\ b_8 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & 1 \\ 35.3 & 29.7 & 28.6 \end{bmatrix} \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 28.6 \end{bmatrix} \right\}^{-1} \begin{bmatrix} 1 & 1 \\ 35.3 & 28.6 \end{bmatrix} \begin{bmatrix} 20 \\ 20 \\ \cdot \\ \cdot \\ \cdot \\ 22 \end{bmatrix}$$

$$\begin{bmatrix} b_o \\ b_8 \end{bmatrix} = \begin{bmatrix} 22.1685 \\ -0.0367 \end{bmatrix}$$

y la ecuación ajustada es $\hat{X}_6 = 22.1685 - 0.0367\ X_8$ y los resi-- duos se muestran en la siguiente tabla.

T A B L A 4 . 1

OBSERVACIONES

i	X_{6i}	$\hat{X}_{6i}$	$X_{6i} - \hat{X}_{6i}$
1	20	20.87	-0.87
2	20	21.08	-1.08
3	23	21.04	1.96
4	20	20.01	-0.01
5	21	19.92	1.08
6	22	19.55	2.45
7	11	19.44	-8.44
8	23	19.36	3.64
9	21	19.58	1.42
10	20	20.06	-0.06
11	20	20.47	-0.47
12	21	21.11	-0.11
13	21	21.14	-0.14
14	19	20.73	-1.73
15	23	20.45	2.55
16	20	20.39	-0.39

Pasa.

17	22	19.99	2.01
18	22	19.60	2.40
19	11	19.60	-8.60
20	23	19.44	3.56
21	20	19.53	0.47
22	21	20.04	0.96
23	20	20.53	-0.53
24	20	20.94	-0.94
25	22	21.12	0.88

Notamos que hay dos residuos, cuyos valores absolutos son considerablemente mayores que el resto de los residuos. Esto resulta, cuando el número de días operados es pequeño (11), considerándolos como puntos raros y podría pensarse en no utilizar esos datos en el análisis. Pero si se desea ajustar una ecuación con una predicción satisfactoria, deberá tomarse en cuentra todos los datos, para desarrollar una ecuación que haga uso de la información que ellas contienen. Si esos meses particulares no fueren tomados en cuenta, el efecto que se produciría en su respuesta sería pequeño y no es porque la variable no afecte la respuesta, sino en un descuido al creer que la variación observada en la variable no podría ejercer un efecto apreciable en la respuesta.

4. Hacer la regresión $Y - \hat{Y}$ contra $X_6 - \hat{X}_6$, para ajustar el modelo.

$$Y_i - \hat{Y}_i = \beta(X_{6i} - X_{6i}) + \varepsilon_i$$

Se debe observar que β_o no es utilizado, ya que se trata específicamente de residuos, cuyas sumas son ceros. Es decir al estimar β, $b_o = 0.$, solamente se estimará $b = b_1$

usando las tablas 1.2 y 4.1, se obtiene:

$$b = \frac{\Sigma(Y_i - \hat{Y}_i)(X_{6i} - \hat{X}_{6i})}{\Sigma(X_{6i} - \hat{X}_{6i})^2} = \frac{42.0821}{208.8523} = 0.2015$$

entonces la ecuación ajustada es

$$\widehat{Y - \hat{Y}} = 0.2015(X_6 - \hat{X}_6)$$

cuyo gráfico se muestra en la figura 4.2

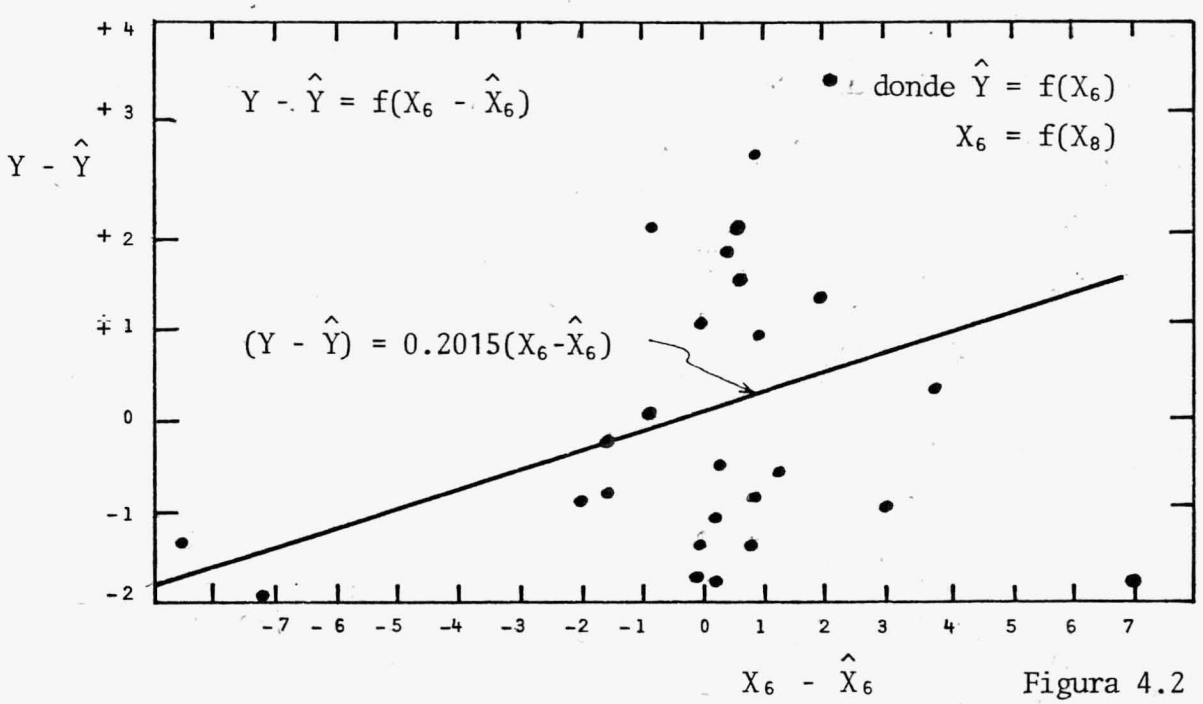


Figura 4.2

En la última ecuación obtenida podemos sustituir $\hat{Y}$ y $\hat{X}_6$ como -- funciones de X_8 asi

$$[Y - (13.6215 - 0.0798\ X_8)] = 0.2015\ [X_6 - (22.1685 - 0.0367\ X_8)]$$

ó

$$\hat{Y} = 9.1545 - 0.0724\ X_8 + 0.2015\ X_6$$

nuestro resultado anterior fué

$$\hat{Y} = 9.1266 - 0.0724\ X_8 + 0.2029\ X_6$$

Al comparar las ecuaciones anteriores, prácticamente son idénti-

cas, las discrepancias observadas son debido a errores de redon deo.

4.2 EXAMINANDO LA ECUACION DE REGRESION

Que tan útil es la ecuación, $Y = f(X_8, X_6)$?

Los residuos de la ecuación ajustada $\hat{Y}$ en función de X_8 y X_6 y los puntos observados estan dados en la siguiente tabla.

T A B L A 4 . 3

OBSERVACION Nº	X_8	X_6	Y	$\hat{Y}$	RESIDUOS
1	35.3	20	10.98	10.63	0.35
2	29.7	20	11.13	11.03	0.10
3	30.8	23	12.51	11.56	0.95
4	58.8	20	8.40	8.93	-0.53
5	61.4	21	9.27	8.94	0.33
6	71.3	22	8.73	8.43	0.30
7	74.4	11	6.36	5.97	0.39
8	76.7	23	8.50	8.24	0.26
9	70.7	21	7.82	8.27	-0.45
10	57.5	20	9.14	9.02	0.12
11	46.4	20	8.24	9.82	-1.58
12	28.9	21	12.19	11.29	0.90
13	28.1	21	11.88	11.35	0.53
14	39.1	19	9.57	10.15	-0.58
15	46.8	23	10.94	10.40	0.54
16	48.5	20	9.58	9.67	-0.09
17	59.3	22	10.09	9.30	0.79
18	70.0	22	8.11	8.52	-0.41
19	70.0	11	6.83	6.29	0.54
20	74.5	23	8.88	8.40	0.48
21	72.1	20	7.68	7.96	-0.28
22	58.1	21	8.47	9.18	-0.71
23	44.6	20	8.86	9.96	-1.10

pasa

24	33.4	20	10.36	16.77	-0.41
25	28.6	22	11.08	11.52	-0.44
	1315	506	235.60		$\Sigma(Y_i - \hat{Y}) = 0$
			$\bar{Y} = 9.424$		$\Sigma(Y_i - \hat{Y})^2 = 9.6432$

Un gráfico de los valores de Y y el valor ajustado $\hat{Y}$ es mostrado en la figura 4.3. El gráfico indica que el modelo ajustado - es un buen predictor del vapor usado mensualmente, sin embargo la adición de X_6 al modelo ha sido útil?

El análisis de varianza de la regresión es como sigue:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Total (no co-- rregido)	25	2284.1102		
Media (b_o)	1	2220.2944		
Total (corre- gido)	24	63.8158		
Regresión (b_o)	2	54.1871	27.0936	61.8999
Residuos	22	9.6287	0.4377	

En base de un nivel de confianza α de 0.05, la ecuación de mínimos cuadrados es

$$\hat{Y} = 9.1266 - 0.0724\ X_8 + 0.2029\ X_6$$

es un buen predictor, el valor calculado F = 61.8999 es más -- grande que el de tablas F(2,22,0.95) = 3.44.

QUE HA PASADO POR LA ADICION DE UNA SEGUNDA VARIABLE INDEPEN--- DIENTE (NOMBRANDOLA X_6)?

Hay varios criterios útiles para responder esta pregunta de los cuales veremos los más importantes.

1. EL CUADRADO DEL COEFICIENTE DE CORRELACION MULTIPLE, R^2.

El cuadrado del coeficiente de correlación múltiple R^2 es definido como sigue

$$R^2 = \frac{\text{suma de cuadrados debido a la regresión } /b_o}{\text{total (corregido) suma de cuadrados.}}$$

Es expresado a veces en porcentaje, 100 R^2, mientras más grande es el valor, mejor es la ecuación ajustada, es decir el modelo explica mejor la variación de los datos. Vamos a comparar el valor de R^2 en cada paso del problema de regresión.

PASO 1. $Y = f(X_8)$

ECUACION DE REGRESION	100 R^2
$\hat{Y} = 13.6215 - 0.0798\ X_8$	$100\ R^2 = 100 \times \frac{45.59}{63.82} = 71.44\%$

PASO 2. $\hat{Y} = f(X_8, X_6)$

ECUACION DE REGRESION	100 R^2
$\hat{Y} = 9.1266 - 0.0724\ X_8 + 0.2029\ X_6$;	$100\ R^2 = 100 \times \frac{54.1871}{63.8158} = 84.91\%$

Entonces hay un incremento sustancial en R^2. Sin embargo este estadístico debe ser usado con precaución, puesto que uno siempre puede tomar $R^2 = 1$. Tal como se mencionó anteriormente cuando la regresión es perfecta .

En resumen, dado que el número de observaciones es mucho más grande que el número de variables X potenciales bajo consideración , la adición de una nueva variable incrementaría el valor de R^2, pero no incrementaría necesariamente, la precisión estimada de la respuesta.

Esto es porque la reducción en la suma de cuadrados de residuos puede ser más bajo que la original, ya que un grado de libertad es removido de los grados de libertad de los residuos.

Veamos el siguiente cuadro.

R^2	VARIABLES	SUMA CUADRADOS	GRADOS LIB.	C.M
39.78	X_2	1922459	15	128163.92
42.23	X_2, X_3	1844400	14	131742.85

Vemos que aunque una variable extra ha sido incluida en el modelo de regresión, la media de cuadrados residuales la ha incrementado puesto que la variable extra produce una reducción en la suma de cuadrados 1922459 - 1844400 = 78059 < 12816392, por la pérdida de un grado de libertad. El valor de R_2 ha incrementado un poco, sin embargo, los cuadrados medios han aumentado, esos resultados los encuentra en el apéndice.

2. EL ERROR ESTANDAR DEL ESTIMADOR S.

Los cuadrados medios de residuos s^2, es un estimador de σ^2_{YX}, la varianza acerca de la regresión. Adicionando una variable antes y despues al modelo, podemos comprobar

El análisis s = $\sqrt{\text{cuadrados medios de residuos}}$

Este estadístico indica que mientras más pequeño sea s, es mejor como estimador. Puesto que s puede ser cero, incluyendo bastantes parámetros en el modelo (así como R^2 puede ser la unidad), este criterio puede ser también usado teniendo mucho ciudado, puesto que hay pocas repeticiones y muchos grados de libertad para el error, las reducciones de s son convenientes.

En nuestro ejemplo

PASO 1 $s = \sqrt{0.7926} = 0.89$

PASO 2 $s = \sqrt{0.4377} = 0.66$

es decir, la adición de X , disminuye el valor de s y se ha mejorado la precisión de estimación.

3. EL ERROR ESTANDAR DEL ESTIMADOR s, COMO UN PORCENTAJE DE LA RESPUESTA MEDIA

Otra manera de ver el decrecimiento de s, es considerarlo en relación a la respuesta. En nuestro ejemplo

PASO 1. s como un porcentaje de la media $\bar{Y}$ es

$$0.89/9.424 = 9.44\ \%$$

PASO 2. s como un porcentaje de la media $\bar{Y}$ es

$$0.66/9.424 = 7.00\ \%$$

Entonces la adición de X_6 ha reducido el error estandar del estimador cerca del 7%, del centro de la respuesta media, para decidir si el nivel de precisión es satisfactorio es necesario tener un conocimiento previo y además tacto personal.

EL CRITERIO DE LA PRUEBA F-SECUENCIAS (DEMOSTRANDO LA CONTRIBUCION ADICIONAL DE X_6 DADO QUE X_8 ESTA YA EN LA ECUACION

Este método de evaluar el valor de X_6 como una variable adicionada a $Y = f(X_8)$ consiste en dividir la suma de cuadrados debido a la regresión en dos partes como sigue:

CUADRO DE ANALISIS DE VARIANZA

FUENTE DE VARIACION	GRADOS LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Total (no corregido)	25	2284.1102		
Media (b_o)	1	2220.2944		
Total (corregido)	24	63.8158		
Regresión	2	54.1871	27.0936	61.8999
Debida b_8/b_o	1	45.5924	45.5924	104.1636
Debida $b_6/b_o,b_8$	1	8.5947	8.5947	19.6361
Residuos	22	9.6287	$0.4377=s^2$	

Puesto que 19.6361 excede a $F(1,22,0.95) = 4.30$ la adición de X_6 ha sido significativa. Esta prueba-F es llamada prueba-F secuencial.

EL CRITERIO DE LA PRUEBA-F PARCIAL

Otra manera de evaluar el valor de X_6 , es considerar el órden de las dos variables en el procedimiento de mínimos cuadrados - por ejemplo, podemos preguntarnos

1. Si hubimos introducido X_6 primero en la ecuación, que contribución habría hecho?
2. Dado que X_6 fué introducida primero, cuál es la contribución de X_8 luego de ser adicionada a la ecuación de regresión?

Esas preguntas son contestada mediante, cálculos hechos en la - siguiente tabla.

CUADRADO DE ANALISIS DE VARIANZA

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	F
Total (no corregido)	25	2284.1102		
Media (b_o)	1	2220.2944		
Total (corregido	24	63.8158		
Regresión/b_o	2	54.1871	27.0936	61.8999
Debida b_6/b_o	1	18.3424	18.3424	41.9063
Debida b_8/b_o, b_6	1	35.8447	35.8447	81.8933
Residuos	22	9.6287	0.4377	

Note que la contribución del X_6 anterior es más importante, que su contribución después de que X_8 ha sido introducida. Esto es observado mediante el valor de F para X_8 así

PASO 1 104.1636

PASO 2 81.8933

Sin embargo X_8, es todavía la variable más importante en ambos casos, puesto que su contribución en la reducción de la suma de cuadrados de residuos es el más significativo, sin tomar en cuenta el orden de introducción de las variables.

ERROR ESTANDAR DE b_i

Usando los resultados dados en la sección 2.6, la matriz de varianza-covarianza de b es $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

Entonces la varianza de $b_i = V(b_i) = C_{ii}\sigma^2$, donde C_{ii} es el -

elemento de la diagonal, en $(\mathbf{X}'\mathbf{X})^{-1}$ correspondiente a la i-ésima variable.

La covarianza de b_i, $b_j = \text{cov}(b_i, b_j) = C_{ij}\,\sigma^2$, donde C_{ij} es el elemento fuera de la diagonal correspondiente a la i-ésima fila con la j-ésima columna, puesto que $(\mathbf{X}'\mathbf{X})^{-1}$ es simétrica;

$C_{ij} = C_{ji}$.

La desviación estandar de b es $\sigma\sqrt{C_{ji}}$, así por ejemplo para las variables X_o, X_8, X_6, el error estandar estimado de b_8 es obtenido así:

$$\text{Valor estimado de la var } b_8 = s^2\mathbf{C}_{88}$$
$$= (0.4377)(0.146207 \times 10^{-3})$$
$$= 0.639948 \times 10^{-4}$$

(donde C_{88} se obtiene de la primera matriz del lado derecho, correspondiendo a la posición (2,2), por ser X_8, la primera variable en la ecuación, matriz de la ecuación 4.0.2)

Entonces la desviación típica estimada de b_8 es

$$\sqrt{\text{var } b_8} = \sqrt{0.639948 \times 10^{-4}} = 0.008$$

LIMITES DE CONFIANZA PARA EL VALOR MEDIO VERDADERO DE **Y**, DADO UN CONJUNTO ESPECIFICO DE EQUIS.

El valor predictado $\hat{Y} = b_o + b_1X_1 + \ldots + b_pX_p$ es un estimador de $E(Y) = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p$.

La varianza de $\hat{Y}$, $V(b_o + b_1X_1 + \ldots + b_pX_p =$

$V(b_o) + X_1^2V(b_1) + \ldots + X_p^2V(b_p) + 2X_1\,\text{covar}(b_o,b_1) + \ldots +$

$2X_{p-1}X_p\,\text{cov}(b_{p-1}, b_p)$.

Esta expresión puede ser escrita convenientemente como sigue

en notación matricial, donde $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$

$$V(\hat{Y}) = \sigma^2 \mathbf{C}\mathbf{X}_o'\mathbf{C}\mathbf{X}_o$$

$$= \sigma^2 \begin{bmatrix} 1 & X_1 & \dots & X_p \end{bmatrix} \begin{bmatrix} C_{oo} & C_{o1} & \dots & C_{op} \\ C_{1o} & C_{11} & \dots & C_{1p} \\ \vdots & \vdots & \ddots & \\ C_{po} & C_{p1} & \dots & C_{pp} \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{bmatrix}$$

entonces, los límites de confianza de $1 - \alpha$ sobre el valor medio verdadero de Y en $\mathbf{X}_o$ son dados por

$$\hat{Y} \pm t\{(n-p-1),\ 1 - \tfrac{1}{2}\alpha\}\ .s.\sqrt{\mathbf{X}'\mathbf{C}\mathbf{X}_o}$$

ejemplo, la varianza de $\hat{Y}$ para el punto $\mathbf{X}$ en el espacio $(X_8 = 32,\ X_6 = 22)$ es obtenido, como sigue:

$$\text{var}\ (\hat{Y}) = s^2(\mathbf{X}_o'\mathbf{C}\mathbf{X}_o)$$

$$= (0.4377)(1,32,22) \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ -0.011242 & 0.146207\text{x}10^{-3} & -0.175467\text{x}10^{-3} \\ -0.106098 & 0.175467\text{x}10^{-3} & 0.478999\text{x}10^{-2} \end{bmatrix} \begin{bmatrix} 1 \\ 32 \\ 22 \end{bmatrix}$$

$$= 0.4377 \times 0.104140 = 0.0045582$$

Los límites de confianza al 95% del valor medio verdadero de Y de $X_8 = 22$ y $X_6 = 22$ son dados por

$$Y \pm t\ (22{,}0975).\ s.\sqrt{\mathbf{X}_o'\mathbf{C}\mathbf{X}_o}$$

$$11.2736 \pm (2.074)\quad \sqrt{0.045582}$$

$$= 11.2736 \pm 0.4418$$

$$\Rightarrow [10.8318,\ 11.7154]$$

Entonces todos los intervalos al 95% de confianza construidos

para el valor medio de Y para $X_8 = 32$, $X_6 = 22$, el 95% de esos intervalos contendrán el valor medio verdadero de Y.

De otra manera, podemos decir que hay una probabilidad de 0.95 que el valor medio verdadero de Y de $X_8 = 32$ y $X_6 = 22$ está entre 10.8318 y 11.7154.

LIMITES DE CONFIANZA PARA LA MEDIA DE g OBSERVACIONES DADAS A - UN CONJUNTO ESPECIFICO DE EQUIS.

Los límites son calculados de

$$\hat{Y} \pm t\,(V,\ 1 - \tfrac{1}{2}\alpha).\ s.\sqrt{1/g + X_o' C X_o}$$

donde, g = # observaciones dadas.

Por ejemplo, los límites de confianza al 95%, para una observación individual para el punto ($X_8 = 32$, $X_6 = 22$) son

$$\hat{Y} \pm t\,(22, 0.975).\ s.\ \sqrt{1 + X_o' C X_o}$$

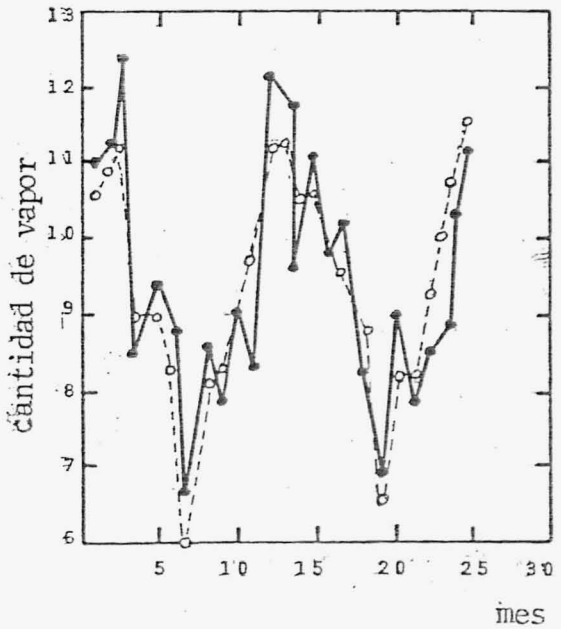
$$11.2736 \pm (2.074)(0.661589)\ \sqrt{1 + 0.10413981}$$

$$11.2736 \pm (2.074)(0.661589)(1.05078)$$

$$11.2736 \pm 1.4418$$

$$\Rightarrow \quad [9.8318,\ 12.7154]$$

Figura 4.3



CAPÍTULO V

MODELOS MÁS COMPLICADOS

5.0

Hasta ahora se ha tratado, regresión lineal, con una variable independiente, la cual fué representada posteriormente en forma matricial, pudiéndose trabajar en esta forma cuando se tienen más variables independientes, teniendo en cuenta que los parámetros a ser estimados eran lineales encontrandoles intervalos de confianza, lo mismo para los Y estimados.

Luego se trabajó algebraicamente, con criterios para examinar una ecuación con dos variables independientes. En este capítulo se verán modelos más complicados, donde algunas de ellas, implican transformaciones en las variables y el uso de variables falsas.

Podemos escribir el tipo más general de modelo lineal de las variables X_1, X_2, ..., X_k en la forma

$$Y = \beta_o Z_o + \beta_1 Z_1 + \quad + \beta_p Z_p + \varepsilon \qquad 5.0.1$$

$Z_o = 1$ es una variable falsa, sin embargo algunas veces es conveniente matemáticamente tener Z_o en el modelo, por ejemplo.

$$(Z_{1i}, Z_{2i}, \ldots, Z_{pi}) \quad , \quad i = 1, 2, \ldots, n.$$

Son n subindices de las variables Z_j, $j = 1, \ldots, p$. correspondientes a observaciones Y_i , $i = 1, \ldots, n$; entonces cuando $j \neq 0$ y $Z_{oi} = 1$

$$\sum_{i=1}^{n} Z_{ji} = \sum_{i=1}^{n} Z_{ji} Z_{oi}$$

y entonces puede ser representado por la expresión general del producto cruz

$$\sum_{i=1}^{n} Z_{ji} Z_{\ell i}$$

Si fueran escritas las ecuaciones normales. Note que $\sum_{i=1}^{n} Z_{oi}^2 = n$.

cada Z_j, $j = 1, \ldots, p$ es una función de $X_1, X_2, \ldots, X_k$, es decir ,

$$Z_j = Z_j(X_1, X_2, \ldots, X_k)$$

y puede tomar cualquier forma, en algún ejemplo Z; puede ser -- función de una variable.

Haciendo reareglos en la ecuación 5.0.1 pueden ser analizados - por los métodos dados en las secciones 2.6. a 2.12

5.1 MODELOS POLINOMIALES DE VARIOS ORDENES EN X_j.

MODELOS DE PRIMER ORDEN.

1. Si $p = 1$, y $Z_1 = X$ en ecuación 5.0.1 obtenemos el modelo simple de primer órden en una variable independiente.

$$Y = \beta_o + \beta_1 X + \varepsilon \qquad 5.1.1$$

2. Si $p = k$, y $Z_j = X_j$ obtenemos un modelo con k varialbes independientes.

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \qquad 5.1.2$$

MODELOS DE SEGUNDO ORDEN

1. Si $p = 2$, $Z_1 = X_1, Z_2 = X^2$ y $\beta_1 = \beta_{11}$, obtenemos el modelo de segundo órden con una variable independiente

$$Y = \beta_o + \beta_1 X + \beta_{11} X^2 + \varepsilon \qquad 5.1.3$$

2. Si $p = 5$, $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = X_1^2$, $Z_4 = X_2^2$, $Z_5 = X_1X_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, $\beta_5 = \beta_{12}$, obtenemos un modelo de segundo órden en dos variables.

$$Y = \beta_o + \beta_1X_1 + \beta_2X_2 + \beta_{11}X_1^2 + \beta_{22}X_2^2 + \beta_{12}X_1X_2 + \varepsilon \qquad 5.1.4$$

Modelos de segundo órden son usados en estudios de superficies de respuesta donde se desea aproximar, las características de algunas superficies de respuesta desconocidas. Puede en algunos casos omitirse algunos términos del modelo, cuando se tiene un conocimiento de ciertos tipos de superficie.

MODELOS DE TERCER ORDEN

1. Si $p = 3$, $Z_1 = X$, $Z_2 = X^2$, $Z_3 = X^3$, $\beta_2 = \beta_{11}$, $\beta_3 = \beta_{111}$ obtenemos el modelo de tercer órden con una variable independiente

$$Y = \beta_o + \beta_1X + \beta_{11}X^2 + \beta_{111}X^3 + \varepsilon \qquad 5.1.5$$

2. Su $p = 9$, el modelo de tercer órden se puede representar con dos variables de la siguiente manera.

$$Y = \beta_o + \beta_1X_1 + \beta_2X_2 + \beta_{11}X_1^2 + \beta_{22}X_2^2 + \beta_{12}X_1X_2 + \beta_{111}X_1^3 + \beta_{222}X_2^3 + \beta_{122}X_1X_2^2 + \beta_{112}X_1^2X_2 + \varepsilon \qquad 5.1.6$$

La forma que se obtienen las betas es de acuerdo al subíndice de las variables así β_{122} es obtenido de $X_1X_2^2 = X_1X_2X_2$.

Un modelo general para k-factores $X_1, X_2, \ldots, X_k$ puede ser obtenido de manera similar. Estos modelos pueden ser usados en superficies de respuesta, aunque menos frecuentemente que los modelos de segundo órden.

Si se quiere utilizar más variables puede hacerce de manera similar a los modelos mostrados.

5.2 MODELOS QUE ENVUELVEN TRANSFORMACIONES DIFERENTES A POTENCIAS DE ENTEROS.

Trataremos aquí modelos no vistos en la sección anterior, pero siempre útiles en regresión.

MODELOS OBTENIDOS POR LA TRANSFORMACION DE X_j

a) TRANSFORMACION RECIPROCA. Si en la ecuación 5.0.1 tomamos $p = 2$, $Z_1 = \frac{1}{X_1}$, $Z_2 = \frac{1}{X_2}$, se obtiene el modelo

$$Y = \beta_o + \beta_1 \frac{1}{X_1} + \beta_2 \frac{1}{X_2} + \varepsilon \qquad 5.2.1$$

b) TRANSFORMACION LOGARITMICA. Usando $p = 2$, $Z_1 = \ell nX$, $Z_2 = \ell nX_2$ de 5.0.1 obtenemos

$$Y = \beta_o + \beta_1 \ell nX_1 + \beta_2 \ell nX_2 + C \qquad 5.2.2$$

c) TRANSFORMACION RAIZ CUADRADA. Usando $p = 2$, $Z_1 = X_1^{1/2}$, $Z_2 = X_2^{1/2}$ de 5.0.1 obtenemos

$$Y = \beta_o + \beta_1 X_1^{1/2} + \beta_2 X_2^{1/2} + \varepsilon \qquad 5.2.3$$

Claramente se pueden hacer muchas transformaciones, teniéndose de esta manera muchos modelos postulados, los cuales contienen nuchos o pocos términos. La selección de alguna transformación puede ser hecha a partir de un conocimiento previo de las variables en el modelo, para poderlas llevar en un momento determinado a un modelo simple, teniéndose después que llevarlas a las variables originales, haciendo la transformación correspondiente.

MODELOS NO LINEALES QUE SON LLEVADOS A UNA FORMA LINEAL

Podemos dividir los modelos no lineales (es decir no lineales - en los parámetros) en dos tipos esencialmente lineales y esen-- cialmente no lineales.

Un modelo esencialmente lineal, puede ser expresado, por conveniencia al transformar las variables, en el modelo lineal nor-- mal de Ec. 5.0.1

Si un modelo no lineal no puede ser expresado en esta forma, entonces es esencialmente no lineal.

Nosotros veremos, modelos esencialmente lineales y así poderlos expresar en forma matricial, donde habrá que utilizar transformaciones sobre las variables dependientes e independientes.

Ejemplos

MODELO MULTIPLICATIVO

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \varepsilon \qquad 5.2.4$$

donde α, β, γ, δ, son parámetros desconocidos, ε es el error - multiplicativo aleatorio, aplicando logaritmos naturales se convierte 5.2.4 en forma lineal.

$\ell n\ Y = \ell n(\alpha X_1^{\beta}\ X_2^{\gamma}\ X_3^{\delta}\ \varepsilon)$, por propiedades de logaritmo se llega - a:

$$\ell n\ Y = \ell n\alpha + \beta \ell n X_1 + \gamma \ell n X_2 + \delta \ell n X_3 + \ell n\varepsilon \qquad 5.2.5$$

siendo éste de la forma 5.0.1 y puede ser resuelto utilizando - los procedimientos usuales de regresión. Puede observar que para obtener pruebas de significación válidas e intervalos de con

fianza para los estimadores, es necesario que $\ell n\ \varepsilon \sim N(o, \mathbf{I}\sigma^2)$, lo cual se verificará al examinar los residuos de la ecuación - ajustada.

MODELO EXPONENCIAL

$$Y = e^{\beta_o + \beta_1 X_1 + \beta_2 X_2} \cdot \varepsilon \qquad 5.2.6$$

tomando logaritmo natural a ambos lados se llega a

$$\ell nY = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ell n\, \varepsilon \qquad 5.2.7$$

MODELO RECIPROCO

$$Y = \frac{1}{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon} \qquad 5.2.8$$

tomando recíproco a ambos lados tenemos

$$\frac{1}{Y} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad 5.2.9$$

Para este caso la respuesta vendrá dada por el recíproco de la variable dependiente.

OTRA FORMA DEL MODELO EXPONENCIAL

$$Y = \frac{1}{1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}} \qquad 5.2.10$$

transformada a la forma lineal tenemos

$$\frac{1}{Y} = 1 + e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

$$\frac{1}{Y} - 1 = e^{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$ aplicando logaritmos se llega a

$$\ell n\left(\frac{1}{Y} - 1\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad 5.2.11$$

Para poder aplicar mínimos cuadrados en los modelos anteriores se tiene la necesidad de transformarlos a la Ec. 5.0.1, para - que los coeficientes así estimados " sean estimadores de míni-- mos cuadrados " en la forma normal y encontrar su valor equivalente en la ecuación original.

Ejemplos de esencialmente no lineales son:

a) $Y = \beta_o + \beta_1 e^{-\beta_2 X} + \varepsilon$

b) $Y = \beta_o + \beta_1 X + \beta_2 \beta_3^X + \varepsilon$

para a) tenemos

$$\ell nY = \ell n(\beta_o + \beta_1 e^{-\beta_2 X} + \varepsilon).$$

no pudiéndose transformar, es decir, no es posible ponerlo en - la forma de la Ec. 5.0.1

5.3 EL USO DE VARIABLES FALSAS EN REGRESION MULTIPLE.

CONCEPTO GENERAL DE UNA VARIABLE FALSA

Las variables consideradas en regresión pueden tomar valores - dentro de rangos contínuos. Ocasionalmente es posible introdu-- cir un factor (término) el cual puede tener dos ó más variables. En tal caso se puede introducir un factor el cual tenga dos ó más niveles por ejemplo los datos pueden provenir de tres máquinas ó dos fábricas ó seis operadores, no pudiéndose establecer una escala contínua para la variable " máquina " " operador " " fábrica " . Se asignan niveles a esas variables debido al hecho de que las máquinas ó fábricas u operadores puedan tener efectos deterministicos en la respuesta, los niveles usualmen-- te (pero no siempre) no tienen relación a algún nivel físico --

que podría existir entre ellas. Un ejemplo de variable falsa - se encuentra en la asignación de la variable X_o del término β_o en el modelo de regresión cuyo valor es la unidad. El uso de - una variable falsa es pura conveniencia, así supongamos que deseamos introducir dentro del modelo la idea de que hay dos tí-- pos de máquinas (A y B), que producen diferentes niveles de respuesta.

Una manera de hacer ésto es adicionar una variable falsa Z y un coeficiente de regresión α, tal que aparece en el modelo un término αZ. Teniéndose que estimar α al mismo tiempo que los betas. Los valores de Z pueden ser:

Z = 0 si la observación es de la máquina A

Z = 1 si la observación es de la máquina B.

Algunos valores distintos de Z podrían ser convenientes, aunque lo anterior es usualmente mejor. Otro ejemplo de asignación para este caso. Supongamos que de un total de n observaciones n_1 viene de la máquina A y $n_2 = n - n_1$ viene de la máquina B. Si - se selecciona los niveles,

$$Z = \frac{- n_2}{[n_1 n_2 (n_1 + n_2)]^{1/2}} \quad \text{para máquina A}$$

$$Z = \frac{n_1}{[n_1 n_2 (n_1 + n_2)]^{1/2}} \quad \text{para máquina B.}$$

Se encontrará que la columna correspondiente de la matriz **X** es ortogonal a la " columna β_o " y tiene suma de cuadrados unidad,

NOTA

Si fué deseado tomar en cuenta tres máquinas distintas dos va-- riables falsas serían requeridas. Entonces tendríamos

$(Z_1, Z_2) = (1, 0)$ para máquina A

$= (0, 1)$ para máquina B

$= (0, 0)$ para máquina C

y al modelo se le incluirían los términos $\alpha_1 Z_1 + \alpha_2 Z_2$ con coe ficientes α_1 y α_2 a ser estimados. Si deseamos, columnas las - cuales sean ortogonales a la " columna β_o " y que tienen suma de cuadrados unidad.

$$(Z_1, Z_2) = \left\{ \frac{-n_3}{[n_1 n_2 (n_1 + n_3)]^{1/2}}, 0 \right\} \text{ para máquina A}$$

$$= \left\{ 0, \frac{-n_3}{[n_2 n_3 (n_2 + n_3)]^{1/2}} \right\} \text{ para máquina B}$$

$$= \left\{ \frac{n_1}{[n_1 n_3 (n_1 + n_3)]^{1/2}}, \frac{n_2}{n_2 n_3 (n_2 + n_3)]^{1/2}} \right\} \text{ para máquina C}$$

donde n_1, n_2, n_3 son respectivamente, los números de observacio nes de máquinas A, B, C.

Asi de manera general cuando se tienen Y niveles, el número de variables falsas requeridas son Y - 1.

EJEMPLO DEL USO DE VARIABLES

EJEMPLO DE BLOQUES.

Tres plantas manufactureras hacen productos idénticos, pero difiere marcadamente en la cantidad de agua usada en cada planta,

La matriz de datos **X** es la siguiente

X_o	X_1	X_2	X_3	X_4	X_5	Y
1	1	0	$X_{3,11}$	$X_{4,11}$	$X_{5,11}$	Y_{11}
1	1	0	$X_{3,12}$	$X_{4,12}$	$X_{5,12}$	Y_{12}
1	1	0	$X_{3,13}$	$X_{4,13}$	$X_{5,13}$	Y_{13}
1	1	0	$X_{3,14}$	$X_{4,14}$	$X_{5,14}$	Y_{14}
1	1	0	$X_{3,15}$	$X_{4,15}$	$X_{5,15}$	Y_{15}
1	0	1	$X_{3,21}$	$X_{4,21}$	$X_{5,21}$	Y_{21}
1	0	1	$X_{3,22}$	$X_{4,22}$	$X_{5,22}$	Y_{22}
1	0	1	$X_{3,23}$	$X_{4,23}$	$X_{5,23}$	Y_{23}
1	0	1	$X_{3,24}$	$X_{4,24}$	$X_{5,24}$	Y_{24}
1	0	1	$X_{3,25}$	$X_{4,25}$	$X_{5,25}$	Y_{25}
1	0	0	$X_{3,31}$	$X_{4,31}$	$X_{5,31}$	Y_{31}
1	0	0	$X_{3,32}$	$X_{4,32}$	$X_{5,32}$	Y_{32}
1	0	0	$X_{3,33}$	$X_{4,33}$	$X_{5,33}$	Y_{33}
1	0	0	$X_{3,34}$	$X_{4,34}$	$X_{5,34}$	Y_{34}
1	0	0	$X_{3,35}$	$X_{4,35}$	$X_{5,35}$	Y_{35}

En esta tabla por ejemplo $X_{5,12}$ representa el valor observado de la variable X_5 en la primera planta cuando la segunda respuesta Y_{12} es registrada. Note que para valores dados,

X_{30}, X_{40} y X_{50} de X_3, X_4 y X_5 la respuesta estimada Y_{12} en la planta Nº 3 es

$$\hat{Y}_3 = b_o + b_3X_{30} + b_4X_{40} + b_5X_{50}$$

La respuesta estimada en la planta Nº 1 es

$$\hat{Y}_{o1} = \hat{Y}_{o3} + b_1$$

Y la respuesta estimada en la planta Nº 2 es

$$\hat{Y}_{o2} = \hat{Y}_{o3} + b_2$$

tq b_1 y b_2 estiman la diferencia en los niveles en la planta,

las cuales no dependen de X_3, X_4, X_5.

TENDENCIAS DE TIEMPO LINEALES

Hay muchas situaciones donde las predicciones son hechas para un período futuro de tiempo.

Cuando se hace un estudio a menudo se encuentra tendencias lineales a través del tiempo y puede removerse tales tendencias usando variables falsas. Se ilustrarán dos casos 1) datos en los cuales hay una tendencia lineal simple, 2) datos en los cuales hay dos tendencias lineales distintas.

CASO 1

Datos en los cuales hay una tendencia lineal simple

Los datos proporcionan los precios en centavos por libra de peso vivo de pollos en intervalos iguales de tiempo, dados en la tabla 5.1

Dos variables falsas son usadas, para remover la tendencia lineal de tiempo y son mostradas en la columna X y X' de la tabla 5. 1

TABLA 5.1

PRECIOS (¢) POR LIBRA DE POLLOS VIVOS

DATOS	PRECIOS POR LIBRA	X	X'
Enero 1955	29.1	1	-10
Mayo 1955	29.0	2	- 9
Sept. 1955	28.6	3	- 8
Enero 1956	28.1	4	- 7
Mayo 1956	28.6	5	- 6
Sept. 1956	28.7	6	- 5

Enero 1957	28.2	7	- 4
Mayo 1957	28.6	8	- 3
Sept. 1957	28.6	9	- 2
Enero 1958	28.1	10	- 1
Mayo 1958	28.7	11	0
Sept. 1958	28.6	12	1
Enero 1959	26.9	13	2
Mayo 1959	27.0	14	3
Sept. 1959	26.8	15	4
Enero 1960	25.7	16	5
Mayo 1960	25.9	17	6
Sept. 1960	25.6	18	7
Enero 1961	25.1	19	8
Mayo 1961	25 2	20	9
Sept. 1961	25.1	21	10

Aunque la columna centrada X' es la mejor puesto que es ortogonal a la columna de unos en la matriz **X**, los modelos apropiados en los dos casos son:

$$Y = \beta_o + \beta_1 X + (\text{otros términos con variables independiente}) + \varepsilon \qquad 5.3.2$$

Haciendo $X_i = X_i - \bar{X}$, $i = 1, \ldots, n$; de donde $X'_i = X_i + \bar{X}$ y sustituyendo en 5.3.2, obtenemos,

$$Y = \beta_o + \beta_1(X_i + \bar{X}) + (\text{otros términos}) + \varepsilon$$

$$Y = \beta_o + \beta_1 X'_i + \beta_1 \bar{X} + (\text{otros términos}) + \varepsilon$$

$$Y = \beta_o + \beta_1 \bar{X} + \beta_1 \bar{X}'_i + (\text{otros términos}) + \varepsilon$$

$$Y = \beta'_o + \beta_1 X'_i + (\text{otros términos}) + \varepsilon \quad , \quad 5.3.3 \text{, donde}$$

$$\beta'_o = \beta_o + \beta_1 \bar{X}$$

Nota. Aquí, puesto que n = 21 es impar, la cantidad $X'_i = X_i - \bar{X}$ son todos enteros. Cuando n es par podemos usar en lugar de ello $X'_i = 2(X_i - \bar{X})$ para evitar fracciones ejemplo

X_i	1	2	3	4	, $\bar{X} = 2\frac{1}{2}$
$X_i - \bar{X}$	$-1\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$1\frac{1}{2}$	
$2(X_i - \bar{X})$	-3	-1	1	3	

CASO 2

Datos en los cuales hay dos tendencias lineales distintas.

Los datos económicos a menudo aparecen tener dos ó más tendencias lineales. Supongamos, por ejemplo, que un gráfico de envíos de un producto particular aparece en la Figura 5.1 los datos aquí ploteados son imaginarios y sin error, y son elegidos solamente para el uso de las variables falsas.

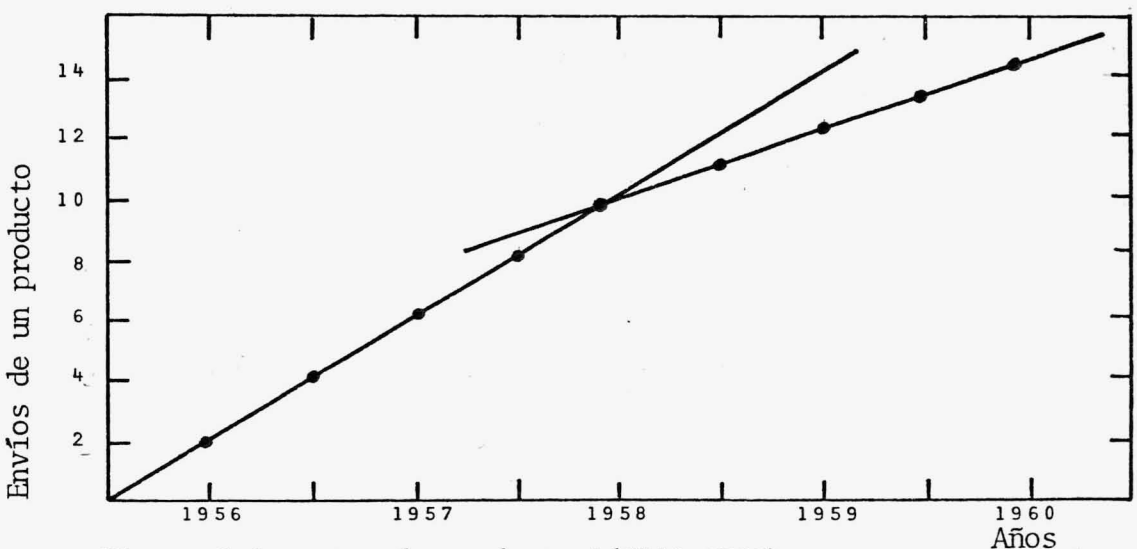


Figura 5.1 ventas de producto A(1956-1960)

Figura 5.1

Deseamos determinar la pendiente, de las dos rectas y el punto de intersección de las dos tendencias lineales, cuando este es

desconocido.

EJEMPLO 1. EL PUNTO DE INTERSECCION ES CONOCIDO, PERO LAS DOS - PENDIENTES SON DESCONOCIDAS.

Los valores de Y ploteados son sucesivamente 2, 4, 6, 8, 10, 11, 12, 13, 14, y es conocido el punto de intersección siendo Y=10.

Para tomar en cuenta las dos pendientes, serán necesarias dos - variables falsas X_1 y X_2, X_1 tomará valores igualmente espaciados para todos los puntos en la primera línea, siendo constante en todos los puntos de la segunda línea. Hagamos X_2 en forma -- opuesta. Entonces la matriz que toma en cuenta las dos tenden-- cias lineales está dada por

NUMERO DE OBSERVACION	X_o	X_1	X_2	Y
1	1	-4	0	2
2	1	-3	0	4
3	1	-2	0	6
4	1	-1	0	8
5	1	0	0	10
6	1	0	1	11
7	1	0	2	12
8	1	0	3	13
9	1	0	4	14

Note que en la 5 observación en la columnas de X_1 y X_2, $(X_1, X_2) = (0, 0)$ que es el punto de intersección conocido. Consideremos el modelo

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad 5.3.4$$

donde b_1 representa la pendiente de la primera línea de tendencia y b_2 la de la segunda, la cual comienza en la observación -

cinco y termina en la nueve. El estimador de b_o es, $\hat{Y}$ cuando $X_1 = X_2 = 0$, esto es, $\hat{Y}$ en el punto de intersección. Las ecuaciones normales para la Ec. 5.3.4 son

$$\begin{aligned} 9b_o - 10b_1 + 10b_2 &= 80 \\ -10b_o + 30b_1 &= -40 \\ 10b_o + 30b_2 &= 130 \end{aligned}$$

resolviendo esas ecuaciones, tenemos

1. El valor de $\hat{Y}$ en los puntos de intersección es dada por $b_o = 10$

2. La pendiente en la primera línea es dada por $b_1 = 2$

3. La pendiente en la segunda línea es dada por $b_2 = 1$

EJEMPLO 2. UN METODO ALTERNATIVO CUANDO EL PUNTO DE INTERSECCION ES CONOCIDO, PERO NO LAS PENDIENTES.

Usando los mismos datos como el ejemplo anterior, la matriz de datos puede ser construida como sigue:

Nº DE OBSERVACIONES	X_o	X_1	X_2	Y
1	1	1	0	2
2	1	2	0	4
3	1	3	0	6
4	1	4	0	8
5	1	5	0	10
6	1	5	1	11
7	1	5	2	12
8	1	5	3	13
9	1	5	4	14

Esto proporciona estimadores de las pendientes como antes, pero el término constante b_o será el valor de Y cuando $X_1 = X_2 = 0$, -

esto es el intersecto de la primera línea de tendencia.

EJEMPLO 3

EL PUNTO DE INTERSECCION Y LAS PENDIENTES SON DESCONOCIDAS.

Otra variable falsa X es necesaria, tomar para conocer el punto desconocido de intersección.

(Se debe observar que se estan usando nuevos datos).

Nº DE OBSERVACIONES	X_o	X_1	X_2	X_3	Y
1	1	1	0	0	2
2	1	2	0	0	4
3	1	3	0	0	6
4	1	4	0	0	8
5	1	5	0	1	9.5
6	1	5	1	1	10.5
7	1	5	2	1	11.5
8	1	5	3	1	12.5
9	1	5	4	1	13.5

Consideremos el modelo

$$\gamma = \beta_o X_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad 5.3.5$$

El valor de β_3 representa un cambio de paso el cual viene, a partir del gráfico de la quinta observación y es la distancia vertical de la segunda línea a la primera en ese punto, cuando la segunda línea está por encima (si está situada debajo β_3 es negativo).

Las ecuaciones normales para el modelo de la ecuación 5.3.5 son

$$9b_o + 35b_1 + 10b_2 + 5b_3 = 77.5$$

$$35b_o + 155b_1 + 50b_2 + 25b_3 = 347.5$$

$$10b_o + 50b_1 + 30b_2 + 10b_3 = 12.5$$

$$5b_o + 25b_1 + 10b_2 + 5b_3 = 57.5$$

Resolviendo esas ecuaciones tenemos.

$b_o = 0$ intersecto de línea Nº 1

$b_1 = 2$ pendiente de línea Nº 1

$b_2 = 1$ pendiente de línea Nº 2

$b_3 = -\frac{1}{2}$ distancia vertical entre línea 1 y 2 en el quinto punto de observación y segunda línea debajo de la primera.

Lo cual se muestra en la figura 5.2 el signo de b_3 (negativo) y el hecho $b_2 > b_1$ indica que el punto de intersección está a la izquierda del quinto punto de observación. Esto ocurre cuando $X = 4\frac{1}{2}$

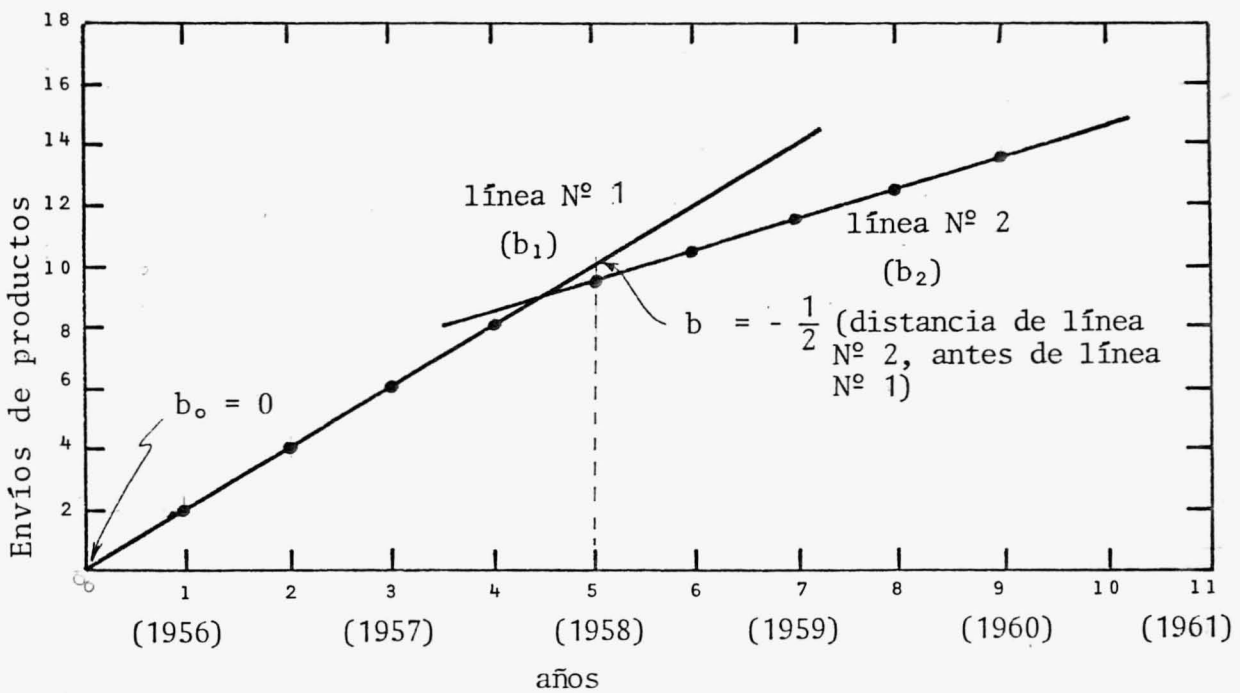


Figura 5.2 Uso de variables falsas; dos líneas, punto de intersección desconocida.

C A P Í T U L O V I

SELECCIONANDO LA MEJOR ECUACIÓN DE REGRESIÓN

6.0 INTRODUCCION

En este capítulo se utilizarán solo procedimientos específicos para seleccionar variables que forman parte de una ecuación de regresión.

Supóngase que deseamos establecer una ecuación lineal para una respuesta particular Y, en términos de las variables predictoras X_1, X_2, ..., X_k. Siendo estas un conjunto de variables de las cuales la ecuación va a ser elegida, incluyendo algunas funciones tales como cuadrados y productos de ellas, siempre que fuese necesario. Dos criterios opuestos pueden usarse al seleccionar una ecuación, los cuales son:

1. Para seleccionar una ecuación útil, con propósitos predictivos, sería necesario que nuestro modelo incluyera tantas equis como sea posible, para que los valores ajustados puedan ser bien determinados.
2. Para obtener información para un número grande de equis, implicaría un gran costo, por lo que se buscaría una ecuación con pocas equis.

En ambos casos se trata de seleccionar la mejor ecuación de regresión, para lo cual no hay procedimiento único. Lo que se hará en este capítulo es describir algunos procedimientos, aunque para algunos puede obtenerse la misma respuesta. Los procedimientos son: 1) todas las regresiones posibles; 2) eliminación ha--

cia atrás; 3) selección hacia adelante; 4) regresión paso a paso.

6.1 TODAS LAS REGRESIONES POSIBLES

Este procedimiento es demasiado largo y no es posible realizarlo sin utilizar una computadora de alta velocidad. El procedimiento requiere ajustar todas las posibles regresiones, el cual involucra X_0, más algún número de variables $X_1, X_2, \ldots, X_k$. (Donde al conjunto usual se le ha aumentado $X_0 = 1$). Lo que nos da un gran número de ecuaciones, ya que cada variable tiene la oportunidad de estar o no en una ecuación, ya que para un número K de variables, en total sería 2^k ecuaciones, así para K = 4, se tendría $2^4 = 16$ ecuaciones las que tendrían que ser examinadas. Pero este total de ecuaciones se ordena de acuerdo a algún criterio, por ejemplo el valor de R^2 obtenido por mínimos cuadrados, para poder tomar una decisión acerca de la mejor ecuación, debemos comparar los valores de R^2, por ejmplo, para aclarar esta situación usaremos los datos de cuatro variables (K=4). Los datos y el ajuste de las diferentes ecuaciones, estan dados en el apéndice al final de este trabajo.

Las variables independientes son X_1, X_2, X_3, X_4 y la variable respuesta X_5 (Y), incluyendose siempre a β_0.

PROCEDIMIENTO

1. Dividir las corridas en cinco conjuntos

   Conjunto A, consiste de la ecuación con solamente el valor medio (modelo $E(Y) = \beta_0$)

   Conjunto B, consiste de cuatro ecuaciones con una variable

(Modelo $E(Y) = \beta_o + \beta_i X_i$)

Conjunto C, consiste de seis ecuaciones con dos variables -- (Modelo $E(Y) = \beta_o + \beta_i X_i + \beta_j X_j$)

Conjunto D, consiste de cuatro ecuaciones con tres variables (Modelo $E(Y) = \beta_o + \beta_i X_i + \beta_j X_j + \beta_k X_k$)

Conjunto E, consiste de la ecuación con todas las variables (Modelo $E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$)

2. Ordenar las ecuaciones de cada conjunto por el valor del cuadrado del coeficiente de correlación múltiple, R^2

3. Al examinar los valores y ver si hay un patrón consistente - de las variables en las ecuaciones de cada conjunto, escogiendo la que tenga, el mayor valor de R^2, del apéndice obtendremos:

CONJUNTO	VARIABLES EN LA ECUACION	R^2
B	$\hat{Y} = f(X_2)$	39.8%
C	$\hat{Y} = f(X_2, X_4)$	57.42%
D	$\hat{Y} = f(X_1, X_2, X_4)$	63.19%
E	$\hat{Y} = f(X_1, X_2, X_3, X_4)$	76.70%

Puede suceder que un conjunto de los cinco, puede haber dos ecuaciones con prácticamente el mismo valor de R^2. Hay necesidad de examinar los resultados, luego que las variables de ese conjunto han sido introducidas y verificar cual R^2 proporciona más - ventajas, pudiendose para ello examinar la matriz de correla--- ción, para ver cuales son las más altamente correlacionadas y - aún así, podría haber una inconsistencia de seleccionar una de las dos ecuaciones, que dependerá para seleccionarla del esta--

dístico que efectúa el trabajo.

El analisis de todas las ecuaciones posibles no porporciona una respuesta clara del problema. Si todas las regresiones son da-- das en un problema grande; una revisión de los cuadrados medios residuales, como el incremento del número de variables en la re gresión, indica el mejor punto de corte para el número de varia bles en la regresión.

Por ejemplo, el problema de los cuadrados medios residuales, pa ra todos los conjuntos de " r " variables, puede ser calculado y ploteado contra r. Así tenemos

r	CUADRADOS MEDIOS RESIDUALES	PROMEDIO $S^2(r)$
1	195460, 128166.67, 211160, 176493.33	177820
2	116650, 196521.43, 165642.85, 131742.86, 97100, 186728.57.	149064.28
3	99984.62, 90400, 160238.46, 91661.54	110571.15
4	61983.1	61983.1

Al plotearlo queda así

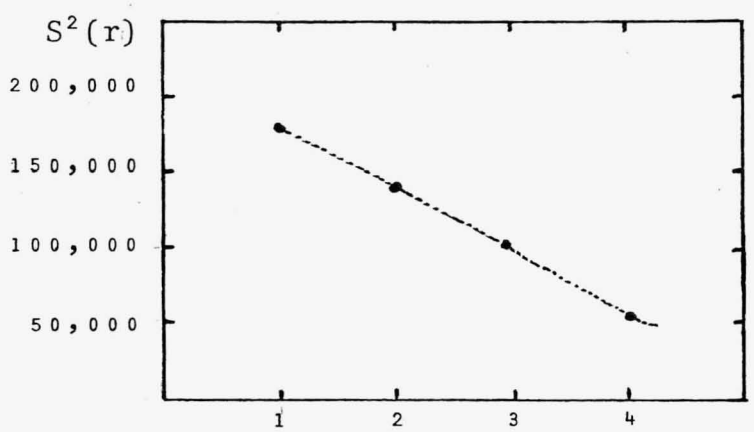


Figura 6.1

Para este caso las cuatro variables deberían ser incluidas, sin embargo este procedimiento admite solo una guía a seleccionar - variables a entrar en la ecuación de regresión, no eligiendo al conjunto específico de variables y el procedimiento no garantiza, que no hay un mejor conjunto de k variables ($k < r$).

Para ajustar la ecuación de regresión el número de variables potenciales y éste número es grande, es decir, más grande que -- $k = 10$ y el número de datos, más grande que k, digamos 5k ó 10k, el ploteo de $S^2(r)$ es muy importante cuando se hacen todas las regresiones.

La ecuación de regresión que involucra más variables independientes de las que son necesarias para ajustar satisfactoriamente - los datos, es llamada sobreajustada. Cuando más y más variables vayan siendo adicionadas a la ecuación y ajustadas, los cuadrados medios de los residuos tendrán a aproximar y estabilizar el verdadero valor de σ^2, cuando el número de variables es incre-- mentado, provocando que todas las variables importantes han sido incluidas.

Esta situación es ilustrada en la figura 6.2

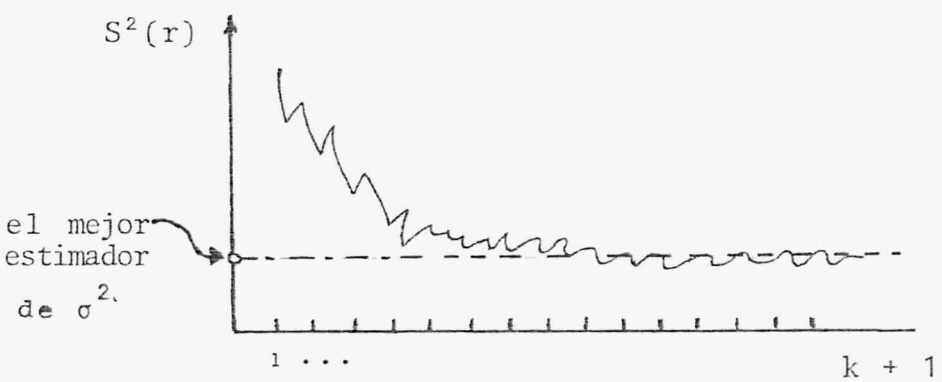


Figura 6.2

Nota. En general, el análisis de todas las regresiones, es quizá incierto. El estadístico ha visto todas las probabilidades y significa que ha examinado muchas ecuaciones de regresiones, -- consumiendo mucho tiempo, siendo preferible, procedimientos que llevan menos tiempo.

6.2 PROCEDIMIENTO DE ELIMINACION HACIA ATRAS.

Este procedimiento es una mejora del método " todas las regre-- siones ". Este método no permite el análisis de todas las regresiones, solamente de la " mejor " regresión, conteniendo un - cierto número de variables.

Los pasos básicos en el procedimiento son estos:

1. La ecuación de regresión conteniendo todas las variables es obtenida.
2. El valor de la prueba F parcial es calculado, para cada va-- riable, tratada como si fuera la última variable a entrar en la ecuación de regresión.
3. El valor de la prueba F parcial más bajo, digamos F_1, es comparado con un valor preseleccionado, de nivel de significa-- ción F_o

a) Si $F_1 < F_o$, removeríamos la variable X_1, la cual sale debido F_1, y se recalcula la ecuación de regresión con las variables restantes.

b) Si $F_1 > F_o$, adoptamos la ecuación de regresión calculada.

Se utilizarán los mismos datos que el procedimiento anterior.

Primero, encontramos la regresión completa con todas las varia-

bles independientes. La ecuación encontrada es

$$\hat{Y} = f(X_1, X_2, X_3, X_4) = 6360.4 + 13.07X_1 + 0.21X_2 - 126.69X_3 - 21.82X_4$$

Este procedimiento forza un ordenamiento de las variables hacia la regresión. El modelo así obtenido ajustado primero a X_1 , entonces X_2, entonces X_3, y finalmente X_4. Para poder eliminar variables debe tener en cuenta, la contribución de cada una de las variables X_1, X_2, X_3 y X_4, a la suma de cuadrados en la regresión, como si cada una estuvo en la última posición. Los valores F parciales mostrados en el apéndice proporcionan medidas de esas contribuciones.

Usando la prueba F parcial, elegimos el más pequeño valor y lo comparamos a algunos valores críticos de F, basados en un nivel predeterminado, α. En este caso el valor crítico para $\alpha = 0.05$, es $F(1, 12, 0.95) = 4.75$. El más pequeño F parcial, es para la variable X_3, siendo su valor calculado $F = 6.96$, siendo su valor mayor que el valor crítico 4.75, por lo que la ecuación dada se adopta, es decir

$$\hat{Y} = 6360.4 + 13.87X_1 + 0.21X_2 - 126.69X_3 - 21.82X_4$$

Nota. Este es un procedimiento satisfactorio, en general, especialmente para quienes les gustaría ver todas las variables en la ecuación. Es más corto el tiempo para hacerlo y más potente que el método " todas las regresiones ". Sin embargo, si los datos provienen de una matriz $\mathbf{X'X}$, la cual está mal condicionada, es decir, es casi singular, entonces este procedimiento puede darnos errores de redondeo. Creemos que este procedimiento es ligeramente inferior, que el que va ser dado, aunque es un exce

lente procedimiento. Pudiendose utilizar t^2 en lugar de F

6.3 PROCEDIMIENTO DE SELECCION HACIA ADELANTE.

Este procedimiento de selección, se llega a una conclusión similar, trabajando en otra dirección, esto es, introduciendo variables, hasta que la ecuación es satisfactoria. El orden de inserción es determinado, por el uso del coeficiente de correlación parcial, como una medida de importancia de las variables, que no estan todavía en la ecuación.

El procedimiento básico es como sigue:

Primero, seleccionamos la X más correlacionada con Y (suponga-- mos que es X_1) y encontramos la ecuación lineal de primer órden $\hat{Y} = f(X_1)$. Encontramos el próximo coeficiente de correlación -- parcial de $X_j (j \neq 1)$ y Y (después de la asignación de X_1). Matemáticamente esto es equivalente a encontrar la correlación entre:

a) Los residuos de la regresión $\hat{Y} = f(X_1)$ y

b) Los residuos de otra regresión $\hat{X}_j = f_j(X_1)$ (la cual actual-- mente no está calculada). El X_j con el más alto coeficiente de correlación parcial con Y, es ahora seleccionado, digamos X_2, y si $\hat{Y} = f(X_1, X_2)$ es ajustada. Este proceso continúa así, si $X_1, X_2, \ldots, X_q$ están ya en la ecuación de regresión los coeficientes de correlación parcial son entre:

   a) los residuos de la regresión $\hat{Y} = f(X_1, X_2, \ldots, X_q)$ y

   b) los residuos de otra regresión $X_j = f_j(X_1, X_2, \ldots, X_q)$ $(j > q)$. Para la entrada de cada variable en la regresión -- los siguientes valores son examinados:

1. R^2, el coeficiente de correlación múltiple.
2. El valor de la prueba F parcial, para la variable más recientemente ingresada, la cual muestra, si suma una cantidad significante de variación, sobre las variables removidas previamente en la regresión.

Tan pronto como el valor F parcial, relacionado con la variable que ha ingresado recientemente, parece no ser significante el proceso finaliza.

Usaremos los datos de los procedimientos anteriores.

El análisis es como sigue:

1. Calcular la correlación de todas las variables independientes, con la respuesta, seleccionar como primera variable de entrada a la regresión, la más alta correlacionada con la respuesta. Al examinar la matriz de correlación en el apendice, demuestra que X_2 es la más alta correlacionada con Y (X_5)

   $r_{25} = 0.63074956$ ó $r_{25} = 0.631$

2. Hacer la regresión Y con X_2 y obtenemos la ecuación de mínimos cuadrados, cuya ecuación es

   $$\hat{Y} = 2273.1 + 0.0799X_2$$

   El valor del recubrimiento de la prueba F es significante para $\alpha = 0.05$ ya que, el valor calculado de F = 9.91 es mayor que $F(1,15,0.95) = 4.54$.

3. Calcular el coeficiente de correlación parcial, de todas las variables que no están en la ecuación, con la respuesta.

Elegimos como la próxima variable a entrar en la ecuación, - la que tenga el más alto coeficiente de correlación parcial. Esta es la variable X_4, ver apéndice donde el cuadrado de -- coeficiente de correlación parcial de X_4 con $Y(X_5)$, dado que X_2, está en la regresión, su valor es

$$r^2_{4.5 \cdot 2} = 0.2928$$

4. Con X_4, así como X_2, ya en la ecuación de regresión de mínimos cuadrados $\hat{Y} = f(X_2, X_4)$, $= Y = 4600.8 + 1.6061X_2 - 1.062X_4$

   Esta ecuación tiene un porcentaje de R^2 de 57.42% (ver apéndice) el cual es significante, además tenemos un valor F=9.44, el cual excede a F(2, 14, 0.95) = 3.74 (excediendo también a F(2, 14, 0.99) = 6.51), por lo que el proceso continúa, es - decir X_4 queda en la ecuación de regresión.

5. Ahora se encontrará el cuadrado de los coeficientes de correlación parcial de todas las variables, que no estan en la regresión con la variable respuesta (ver apéndice). La variable a ingresar es X_1, cuyo valor de correlación parcial, cuando X_2 y X_4 estan ya en la regresión es

$$r^2_{15 \cdot 24} = 0.0522897$$

6. La nueva ecuación es $\hat{Y} = f(X_1, X_2, X_4) = 3865.9 - 0.2431X_1 + 1.5265X_2 - 0.9688X_4$, el porcentaje del coeficiente de correlación múltiple, R^2, 57.42%, se incrementa a 63.19%, siendo la adición de X_1, estadisticamente poco significante, el valor F parcial de X_1 es 2.04, (ver apéndice, la ecuación - que tiene esas variables), el cual no excede a F(1, 13, 0.95) = 4.67, ni a F(1, 13, 0.99) = 9.07, entonces

X_1 sería rechazado y el proceso terminado.

El análisis completo de la tabla de varianza es así:

FUENTE	gl	sc	cm
Total	12	3192631	
Debido a la regresión	4	2448834	
Debido a X_2	1	1270172	1270172
Debido a X_4/X_2	1	563099	563099
Debido a X_1/X_2, X_4	1	1454341	1454341

La tabla llega, hasta aquí, pudiendose observar que cuando una variable, pasa a formar parte de la ecuación la suma de cuadrados disminuye, mientras que si queda en la ecuación, la suma - aumenta, por lo que la ecuación es

$$\hat{Y} = 4600.8 + 1.6061X_2 - 1.062X_4$$

Nota. Este procedimiento es básicamente una buena idea, lleva - menos tiempo de computadora, que los procedimientos anteriores. Una desventaja es que si hacemos un esfuerzo de explorar el -- efecto de una nueva variable, es dificil hacerlo, esta dificultad es superada por el procedimiento paso a paso.

6.4 PROCEDIMIENTO DE REGRESION PASO A PASO

Este procedimiento es una versión mejorada del procedimiento - " hacia adelante ". Mejora en el re-análisis, en cada etapa de la regresión de las variables incorporadas en el modelo en etapas previas.

Al revisar, el criterio F parcial, para cada variable en la regresión es calculado y comparado, con un valor apropiado de la

distribución F.

Esto proporciona un juicio de la contribución hecha por cada variable, como si ha sido la variable recientemente ingresada, independiente de la posición de ingreso en el modelo.

Alguna variable, que no da una contribución significante, es removida del modelo. Este proceso continúa, hasta que no haya más variables que ingresen y no más sean rechazadas.

Paso 1. Comienza con la matriz de correlación simple y entra en la ecuación, la más altamente correlacionada con la respuesta.

Aquí X_2 como en el procedimiento de selección " hacia adelante".

Paso 2. Usando el coeficiente de correlación parcial como antes, ahora se selecciona como la próxima variable a ingresar en la regresión, la variable X, cuya correlación parcial con la respuesta es más alto. Para este caso X_4, como en el procedimiento " hacia adelante ",

Paso 3. Dada la ecuación de regresión $\hat{Y} = f(X_4, X_2)$, el método examina ahora la contribución que X_4 tendría, si X_4 ha ingresado 1º y de 2da. a X_2. (El procedimiento " hacia adelante " no hace esto). Puesto que el valor de F parcial es 13.27, el cual es significante para $\alpha = 0.05$, puesto que $F(1, 14, 0.95) = 4.60$ por lo que X_4 es retenida, los valores de F pueden verse en el apéndice, siendo el valor de F, como F secuencial.

El método selecciona la próxima variable a ingresar, la más altamente correlacionada parcialmente con la respuesta. Esta es la variable X_1, donde el cuadrado de su coeficiente de correlación parcial es 0.0522. Puede verse al final en el apéndice.

Paso 4. Una ecuación de regresión de la forma $\hat{Y} = f(X_2, X_4, X_1)$ - es ahora determinada por mínimos cuadrados, la variable X_1 in-- gresa con un valor F secuencial de 2.04 (coincide F parcial) el cual no excede a $F(1, 13, 0.95) = 4.67$. (Ver apéndice.)

Por lo que X_1 es rechazado y termina el proceso, quedando la -- ecuación.

$$\hat{Y} = 4600.81 + 0.203431X_2 - 21.56737X_4$$

OPINION. Hasta ahora es el mayor procedimiento estudiado, pero puede ser mal utilizado fácilmente por un estadístico princi--- piante, requiriendo una opinión sensata en la selección inicial de las variables y un análisis crítico del modelo a través de - la observación de los residuos.

APENDICE

Para el uso de este apéndice, al inicio se encuentran la tabla de datos, la matriz de correlación [donde cada columna corresponde, al órden que llevan las variables en la tabla de datos, así el valor 0.4362975 corresponde a r_{13} (1 de la fila y 3 subindice de la variable en esa columna)], los coeficientes de correlación parcial.

Luego las tablas de cada una de las ecuaciones, donde encontramos:

a) variables independiente(s) y dependiente, el número indica el subindice de la variable en la ecuación.
b) BETA: son los coeficientes de la ecuación de regresión
c) T-DIST: es la distribución t, para el uso en este trabajo, se utiliza como $t^2 \equiv F$, llamándosele F parcial o F secuencial.
d) Y INTERCEPT: es el intercepto de la ecuación
e) MULTIPLE-R: es la raíz cuadrada de R^2 (coeficiente de correlación múltiple)
f) SSR: regresión
g) SSE: residuos
h) F: valor F
i) S: desviación estandar

La suma total se obtiene sumando f) con e)

En una industria se tuvierón las siguientes variables:

X_1 = promedio mensual de temperatura (°F)

X_2 = costo total de producción (M libras)

X_3 = número de días operados en planta por mes

X_4 = número de personas mensualmente pagadas

$Y = X_5$ = es el uso del agua mensualmente (en galones)

Datos originales

T A B L A

	X_1	X_2	X_3	X_4	$X_5 = Y$
1	58.8	7107	21	129	3067
2	65.2	6373	22	141	2828
3	70.9	6796	22	153	2891
4	77.4	9208	20	166	2994
5	79.3	14792	25	193	3082
6	81.0	14564	23	189	3898
7	71.9	11964	20	175	3502
8	63.9	13526	23	186	3060
9	54.5	12656	20	190	3211
10	39.5	14119	20	187	3286
11	44.5	16691	22	195	3542
12	43.6	14571	19	206	3125
13	56.0	13619	22	198	3022
14	64.7	14575	22	192	2922
15	73.0	14556	21	191	3950
16	78.9	18573	21	200	4488
17	79.4	15618	22	200	3295

MATRIZ DE CORRELACION

1	1.00000000	-0.02410741	0.43762975	-0.08205777	0.28575758
2	-0.02410741	1.00000000	0.10573055	0.91847987	0.63074956
3	0.43762975	0.10573055	1.00000000	0.03188120	-0.08882581
4	-0.08205777	0.91847987	0.03188120	1.00000000	0.41324613
5	0.28575758	0.63074956	-0.08882581	0.41324613	1.00000000

COEFICIENTES DE CORRELACION PARCIAL

1. Coeficientes de correlación parcial al cuadrado para la ecuación $\hat{Y} = f(X_2)$, para las variables no en la ecuación de regresión.

VARIABLE Nº	COEFICIENTE (r^2)
1	0.1505109
3	0.0406178
4	0.2928

2. Coeficientes de correlación parcial al cuadrado para la ecuación $\hat{Y} = f(X_2, X_4)$, para variables no en la ecuación de regresión.

VARIABLE Nº	COEFICIENTE (r^2)
1	0.0522897
3	2.7547×10^{-4}

WHICH PROCEDURE DO YOU WISH TO USE?
#REGRESSION
SPECIFY THE DEPENDENT VARIABLE(S)
>
 5
SPECIFY THE INDEPENDENT VARIABLE(S)
>1

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
1	9.4483	.28576	8.1811	1.1549

Y INTERCEPT: 2691.0
MULTIPLE R: .28576
SSR: .26070E+06
SSE: .29319E+07
S: 442.11
F: 1.3338
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3246.5	-179.52	-.40604
2	2928.0	3307.0	-478.98	-1.0834
3	2891.0	3360.8	-469.84	-1.0627
4	2994.0	3422.3	-428.25	-.96866
5	3082.0	3440.2	-358.21	-.81022
6	3898.0	3456.3	441.73	.99914
7	3502.0	3370.3	131.71	.29792
8	3060.0	3294.7	-234.70	-.53087
9	3211.0	3205.9	5.1119	.11563E-01
10	3286.0	3064.2	221.84	.50177
11	3542.0	3111.4	430.59	.97395
12	3125.0	3107.9	22.098	.49984E-01
13	3022.0	3220.1	-198.06	-.44799
14	2922.0	3302.3	-380.26	-.86010
15	3950.0	3380.7	569.32	1.2877
16	4488.0	3436.4	1051.6	2.3785
17	3295.0	3441.2	-146.15	-.33057

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>2

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
2	.79890E-01	.63075	.25377E-01	3.1481

Y INTERCEPT: 2273.1
MULTIPLE R: .63075
SSR: .12702E+07
SSE: .19225E+07
S: 358.00
F: 9.9105
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2840.9	226.13	.63166
2	2828.0	2782.2	45.773	.12786
3	2891.0	2816.0	74.980	.20944
4	2994.0	3008.7	-14.715	-.41103E-01
5	3082.0	3454.8	-372.82	-1.0414
6	3898.0	3436.6	461.39	1.2888
7	3502.0	3228.9	273.11	.76287
8	3040.0	3353.7	-293.68	-.82033
9	3211.0	3284.2	-73.175	-.20440
10	3286.0	3401.1	-115.05	-.32138
11	3542.0	3606.5	-64.531	-.18025
12	3124.0	3437.2	-312.16	-.87197
13	3023.0	3361.1	-339.11	-.94723
14	2922.0	3437.5	-515.48	-1.4399
15	3950.0	3436.0	514.03	1.4358
16	4488.0	3756.9	731.12	2.0422
17	3295.0	3520.8	-225.81	-.63075

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>3

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
3	-27.125	-.88824E-01	78.587	-.34539

Y INTERCEPT: 3886.1
MULTIPLE R: .88824E-01
SSR: 25190.
SSE: .31674E+07
S: 459.52
F: .11929
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3316.5	-249.47	-.54289
2	2828.0	3289.3	-461.35	-1.0040
3	2891.0	3289.3	-398.35	-.86686
4	2994.0	3343.6	-349.60	-.76078
5	3082.0	3208.0	-125.97	-.27413
6	3898.0	3262.2	635.78	1.3836
7	3502.0	3343.6	158.40	.34471
8	3060.0	3362.2	-302.22	-.44006
9	3211.0	3343.6	-132.60	-.28855
10	3286.0	3343.6	-57.596	-.12534
11	3542.0	3289.3	252.65	.54982
12	3124.0	3370.7	-245.72	-.53473
13	3023.0	3284.3	-267.35	-.58179
14	2992.0	3289.3	-367.35	-.79940
15	3950.0	3316.5	633.53	1.3787
16	4488.0	3316.5	1171.5	2.5494
17	3295.0	3289.3	5.6547	.12305E-01

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
!SPECIFY THE INDEPENDENT VARIABLE(S)
>4

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-STAT.
4	8.3927	.41325	4.7751	1.7574

Y INTERCEPT: 1777.7
MULTIPLE R: .41325
SSR: .54521E+06
SSE: .26474E+07
S: 420.11
F: 3.0891
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y- OBSERVED	Y- ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2860.4	206.62	.49183
2	2828.0	2961.1	-133.09	-.31679
3	2891.0	3061.8	-170.80	-.40656
4	2994.0	3170.9	-176.90	-.42109
5	3082.0	3397.5	-315.51	-.75100
6	3898.0	3363.9	534.06	1.2712
7	3502.0	3246.4	255.56	.60832
8	3060.0	3338.8	-278.76	-.66353
9	3211.0	3372.3	-161.33	-.38401
10	3286.0	3347.1	-61.150	-.14556
11	3542.0	3414.3	127.71	.30399
12	3125.0	3506.6	-381.61	-.90835
13	3022.0	3439.5	-417.47	-.99371
14	2922.0	3389.1	-467.11	-1.1119
15	3950.0	3380.7	569.28	1.3551
16	4488.0	3456.3	1031.7	2.4559
17	3295.0	3456.3	-161.25	-.38384

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>1,2

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
1	9.9569	.30114	6.3220	1.5750
2	.80809E-01	.63801	.24218E-01	3.3368

Y INTERCEPT: 1615.5
MULTIPLE R: .69891
SSR: .15595E+07
SSE: .16331E+07
S: 341.54
F: 6.6846
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2775.3	291.73	.85416
2	2828.0	2779.7	48.320	.14148
3	2891.0	2870.6	20.383	.59681E-01
4	2994.0	3130.2	-136.25	-.39892
5	3082.0	3600.4	-518.41	-1.5178
6	3898.0	3598.9	299.09	.87571
7	3502.0	3298.2	203.80	.59672
8	3060.0	3344.8	-284.77	-.83377
9	3211.0	3180.9	30.132	.88225E-01
10	3286.0	3149.7	136.26	.39896
11	3542.0	3407.4	134.64	.39420
12	3125.0	3227.1	-102.09	-.29890
13	3022.0	3273.6	-251.62	-.73673
14	2922.0	3437.5	-515.50	-1.5093
15	3950.0	3518.6	431.39	1.2631
16	4488.0	3902.0	586.04	1.7159
17	3295.0	3668.2	-373.15	-1.0925

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5

SPECIFY THE INDEPENDENT VARIABLE(S)
>1,3

REGRESSION OUTPUT Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
1	13.276	.40153	9.1233	1.4552
3	-80.787	-.26455	84.262	-.95876

Y INTERCEPT: 4177.2
MULTIPLE R: .37181
SSR: .44135E+06
SSE: .27513E+07
S: 443.31
F: 1.1229
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y- OBSERVED	Y- ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3261.4	-194.36	-.43844
2	2828.0	3265.5	-437.54	-.98700
3	2891.0	3341.2	-450.22	-1.0156
4	2994.0	3589.1	-595.09	-1.3424
5	3082.0	3210.4	-128.38	-.28959
6	3898.0	3394.5	503.48	1.1357
7	3502.0	3516.1	-14.068	-.31735E-01
8	3060.0	3167.5	-107.50	-.24249
9	3211.0	3285.1	-74.061	-.16707
10	3286.0	3085.9	200.08	.45134
11	3542.0	2990.7	551.28	1.2436
12	3125.0	3221.1	-96.137	-.21686
13	3022.0	3143.4	-121.40	-.27385
14	2922.0	3258.9	-336.90	-.75998
15	3950.0	3449.9	500.11	1.1281
16	4488.0	3528.2	959.79	2.1651
17	3295.0	3454.1	-159.07	-.35882

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>1,4

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.	F
1	10.641	.32183	7.5567	1.4082	1.98
4	8.9290	.43966	4.6416	1.9237	3.7

Y INTERCEPT: 990.09
MULTIPLE R: .52312
SSR: .87367E+06
SSE: .23190E+07
S: 406.99
F: 2.6373
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2767.6	299.37	.73556
2	2828.0	2942.9	-114.89	-.28228
3	2891.0	3110.7	-219.69	-.53979
4	2994.0	3295.9	-301.93	-.74187
5	3082.0	3557.2	-475.23	-1.1677
6	3898.0	3539.6	358.39	.88059
7	3502.0	3317.8	184.23	.45267
8	3060.0	3330.9	-270.86	-.66551
9	3211.0	3266.5	-55.546	-.13648
10	3286.0	3080.1	205.86	.50581
11	3542.0	3204.8	337.22	.82857
12	3125.0	3293.4	-168.42	-.41382
13	3022.0	3353.9	-331.94	-.81560
14	2922.0	3392.9	-470.94	-1.1571
15	3950.0	3472.3	477.66	1.1737
16	4488.0	3615.5	872.52	2.1438
17	3295.0	3620.8	-325.80	-.80052

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>2,3

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
2	.81996E-01	.64738	.25874E-01	3.1691
3	-48.028	-.15727	62.383	-.76989

Y INTERCEPT: 3277.1
MULTIPLE R: .64985
SSR: .13483E+07
SSE: .18444E+07
S: 362.96
F: 5.1171
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2851.3	215.74	.59437
2	2828.0	2743.1	84.948	.23404
3	2891.0	2777.7	113.26	.31205
4	2994.0	3071.6	-77.567	-.21370
5	3082.0	3289.3	-207.29	-.57112
6	3898.0	3366.7	531.35	1.4639
7	3502.0	3297.5	204.45	.56329
8	3060.0	3281.5	-221.54	-.61037
9	3211.0	3354.3	-143.29	-.39478
10	3286.0	3474.2	-188.25	-.51865
11	3542.0	3589.1	-47.087	-.12973
12	3125.0	3559.3	-434.34	-1.1967
13	3022.0	3337.2	-315.20	-.86840
14	2922.0	3415.6	-493.58	-1.3599
15	3950.0	3462.1	487.95	1.3443
16	4488.0	3791.4	696.57	1.9191
17	3295.0	3501.1	-206.11	-.56784

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>2,4

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
2	.20343	1.6061	.55854E-01	3.6422
4	-21.567	-1.0620	8.9559	-2.4082

Y INTERCEPT: 4600.8
MULTIPLE R: .75777
SSR: .18333E+07
SSE: .13594E+07
S: 311.60
F: 9.4404
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3264.4	-197.40	-.63349
2	2828.0	2856.3	-28.273	-.90733E-01
3	2891.0	2683.5	207.48	.66586
4	2994.0	2893.8	100.18	.32151
5	3082.0	3447.5	-365.45	-1.1728
6	3898.0	3487.3	410.66	1.3179
7	3502.0	3260.4	241.64	.77546
8	3060.0	3340.9	-280.88	-.90141
9	3211.0	3077.6	133.37	.42802
10	3286.0	3439.9	-153.95	-.49405
11	3542.0	3790.6	-248.63	-.79792
12	3125.0	3122.1	2.8800	.92425E-02
13	3022.0	3101.0	-78.993	-.25350
14	2922.0	3424.9	-502.88	-1.6138
15	3950.0	3442.6	507.42	1.6284
16	4488.0	4065.7	422.35	1.3554
17	3295.0	3464.5	-169.52	-.54401

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>3,4

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
3	-31.180	-.10210	73.890	-.42198
4	8.4588	.41650	4.9141	1.7213

Y INTERCEPT: 2435.2
MULTIPLE R: .42566
SSR: .57846E+06
SSE: .26142E+07
S: 432.12
F: 1.5490
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2871.6	195.44	.45229
2	2828.0	2941.9	-113.88	-.26354
3	2891.0	3043.4	-152.39	-.35265
4	2994.0	3215.7	-221.71	-.51308
5	3082.0	3288.2	-206.20	-.47718
6	3898.0	3316.7	581.28	1.3452
7	3502.0	3291.8	210.16	.48635
8	3060.0	3291.3	-231.35	-.53538
9	3211.0	3418.7	-207.72	-.48071
10	3286.0	3393.3	-107.35	-.24842
11	3542.0	3398.7	143.34	.33173
12	3125.0	3585.2	-460.24	-1.0651
13	3022.0	3424.0	-402.03	-.93037
14	2972.0	3373.3	-451.28	-1.0443
15	3950.0	3396.0	554.00	1.2821
16	4488.0	3472.1	1015.9	2.3509
17	3295.0	3440.9	-145.95	-.33775

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5

SPECIFY THE INDEPENDENT VARIABLE (S)
>1,2,3
REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
1	15.234	.46074	4.5278	2.3837
2	.86150E-01	.68017	.22611E-01	3.8100
3	-110.66	-.36237	60.618	-1.8257

Y INTERCEPT: 3580.3
MULTIPLE R: .7699
SSR: .18928E+07
SSE: .12998E+07
S: 316.21
F: 6.3102
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	2764.5	302.53	.95676
2	2828.0	2688.1	139.93	.44253
3	2891.0	2811.3	79.657	.25191
4	2994.0	3339.5	-345.48	-1.0926
5	3082.0	3296.2	-214.18	-.67733
6	3898.0	3523.8	374.25	1.1835
7	3502.0	3493.1	8.8796	.28082E-01
8	3060.0	3173.8	-113.83	-.35999
9	3211.0	3287.7	-76.665	-.24245
10	3286.0	3185.2	100.81	.31880
11	3542.0	3261.6	280.38	.88671
12	3125.0	3397.3	-272.25	-.86099
13	3022.0	3172.2	-150.16	-.47487
14	2922.0	3387.0	-465.05	-1.4707
15	3950.0	3622.5	327.48	1.0357
16	4488.0	4058.5	429.54	1.3584
17	3295.0	3700.8	-405.84	-1.2835

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>1,2,4

REGRESSION OUTPUT. Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T-DIST.
1	8.0364	.24306	5.6304	1.4273
2	.19334	1.5265	.54355E-01	3.5570
4	-19.676	-.96884	8.7425	-2.2507

Y INTERCEPT: 3865.9
MULTIPLE R: .79492
SSR: .20174E+07
SSE: .11752E+07
S: 300.66
F: 7.4390
DO YOU WANT TO PRINT RESIDUALS?
>YES
RESIDUAL TABLE. Y = VARIABLE # 5

OBS. NUMBER	Y-OBSERVED	Y-ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3174.3	-107.32	-.35694
2	2828.0	2847.7	-19.724	-.65602E-01
3	2891.0	2739.2	151.80	.50488
4	2994.0	3002.0	-7.9820	-.26548E-01
5	3082.0	3565.6	-483.61	-1.6085
6	3898.0	3613.9	284.11	.94493
7	3502.0	3313.5	188.46	.62680
8	3060.0	3334.8	-274.81	-.91401
9	3211.0	3012.4	198.64	.66068
10	3286.0	3233.7	52.304	.17396
11	3542.0	3613.7	-71.741	-.23861
12	3125.0	2980.2	144.81	.48165
13	3022.0	3053.2	-31.188	-.10373
14	2922.0	3426.0	-504.00	-1.6763
15	3950.0	3508.7	441.30	1.4677
16	4488.0	4155.7	332.32	1.1053
17	3295.0	3588.4	-293.38	-.97575

```
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>1,3,4

REGRESSION OUTPUT. Y = VARIABLE #  5

  IND.          B              BETA          STD.          T-
VARIABLE                                    ERROR         DIST.

    1       15.084           .45620         8.2856        1.8205
    3      -92.575          -.30315         76.306       -1.2132
    4       9.3492           .46035         4.5784        2.0420

Y INTERCEPT:     2613.2
MULTIPLE R:      .58951
SSR:             .11095E+07
SSE:             .20831E+07
S:               400.30
F:               2.3081
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE #  5

  OBS.          Y-             Y-           RESIDUAL       STANDARD
NUMBER       OBSERVED       ESTIMATED                      RESIDUAL

    1         3067.0         2762.1          304.89         .76165
    2         2828.0         2878.3         -50.263        -.12556
    3         2891.0         3076.4         -185.43        -.46323
    4         2994.0         3481.2         -487.16        -1.2170
    5         3082.0         3299.4         -217.38        -.54304
    6         3898.0         3472.8          425.23         1.0623
    7         3502.0         3482.3          19.653        .49096E-01
    8         3060.0         3186.8         -126.79        -.31675
    9         3211.0         3360.1         -149.13        -.37254
   10         3286.0         3105.8          180.18         .45011
   11         3542.0         3070.9          471.11         1.1769
   12         3125.0         3437.9         -312.88        -.78160
   13         3022.0         3272.4         -250.40        -.62552
   14         2922.0         3347.5         -425.53        -1.0630
   15         3950.0         3556.0          394.05         .98439
   16         4488.0         3729.1          758.91         1.8959
   17         3295.0         3644.1         -349.06        -.87199
```

```
DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES
SPECIFY THE DEPENDENT VARIABLE(S)
>5
SPECIFY THE INDEPENDENT VARIABLE(S)
>2,3,4

REGRESSION OUTPUT. Y = VARIABLE #  5

  IND.          B              BETA          STD.          T-
VARIABLE                                    ERROR         DIST.

    2        .21790          1.7204       .55312E-01     3.9395
    3      -71.385          -.23376         52.766       -1.3529
    4      -23.547          -1.1595         8.8238       -2.6686

Y INTERCEPT:    6306.8
MULTIPLE R:      .79169
SSR:            .20010E+07
SSE:            .11916E+07
S:              302.76
F:              7.2769
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE. Y = VARIABLE #  5

 OBS.          Y-             Y-           RESIDUAL       STANDARD
NUMBER      OBSERVED       ESTIMATED                      RESIDUAL

    1        3067.0         3318.7        -251.74        -.83149
    2        2828.0         2804.8         23.157         .76487E-01
    3        2891.0         2614.4         276.55         .91345
    4        2994.0         2976.7         17.316         .57194E-01
    5        3082.0         3200.8        -118.75        -.39223
    6        3898.0         3388.0         509.97         1.6844
    7        3502.0         3365.3         136.70         .45152
    8        3060.0         3232.5        -172.49        -.56972
    9        3211.0         3162.9         48.123         .15895
   10        3286.0         3552.3        -266.31        -.87963
   11        3542.0         3781.6        -239.61        -.79143
   12        3125.0         3274.8        -149.79        -.49475
   13        3022.0         3041.6        -19.569        -.64635E-01
   14        2922.0         3391.2        -469.17        -1.5497
   15        3950.0         3482.0         468.04         1.5459
   16        4488.0         4145.4         342.65         1.1318
   17        3295.0         3430.1        -135.06        -.44611
```

DO YOU WANT TO PERFORM MORE REGRESSION NOW?
>YES

SPCIFY THE INDEPENDENT VARIABLE(S)
>1,2,3,4

SPECIFY THE DEPENDENT VARIABLE(S)
>5

REGRESSION OUTPUT, Y = VARIABLE # 5

IND. VARIABLE	B	BETA	STD. ERROR	T- DIST.
1	13.869	.41945	5.1598	2.6879
2	.21170	1.6714	.45543E-01	4.6484
3	-126.69	-.41486	48.022	-2.6382
4	-21.818	-1.0743	7.2845	-2.9951

Y INTERCEPT: 6360.3
MULTIPLE R: .87580
SSR: .24488E+07
SSE: .74380E+06
S: 248.96
F: 9.8770
DO YOU WANT TO PRINT RESIDUALS?
>YES

RESIDUAL TABLE, Y = VARIABLE # 5

OBS. NUMBER	Y- OBSERVED	Y- ESTIMATED	RESIDUAL	STANDARD RESIDUAL
1	3067.0	3205.4	-138.38	-.55584
2	2828.0	2750.2	77.751	.31230
3	2891.0	2657.0	233.96	.93975
4	2994.0	3227.6	-233.56	-.93812
5	3082.0	3213.5	-131.52	-.52827
6	3898.0	3529.5	368.52	1.4802
7	3502.0	3538.4	-36.371	-.14609
8	3060.0	3138.0	-78.031	-.31342
9	3211.0	3116.3	94.718	.38045
10	3286.0	3283.4	2.5757	.10346E-01
11	3542.0	3469.3	72.656	.29183
12	3125.0	3148.1	-23.125	-.92885E-01
13	3022.0	2913.0	108.97	.43769
14	2922.0	3367.0	-444.99	-1.7873
15	3950.0	3626.6	323.42	1.2990
16	4488.0	4362.5	125.54	.50425
17	3295.0	3617.1	-322.12	-1.2938

B I B L I O G R A F Í A

1. APPLIED REGRESSION ANALYSIS
   Norman Draper y Harry Smith

2. STADISTICAL ANALYSIS A COMPUTER ORIENTED APPROACH.
   A.A. Fifi y S.P. Azen

3. INTRODUCCION A LA PROBABILIDAD Y ESTADISTICA
   Beaver y Mendenhall

4. PROBABILIDAD Y ESTADISTICA PARA INGENIEROS
   Irwin Miller y John Freund