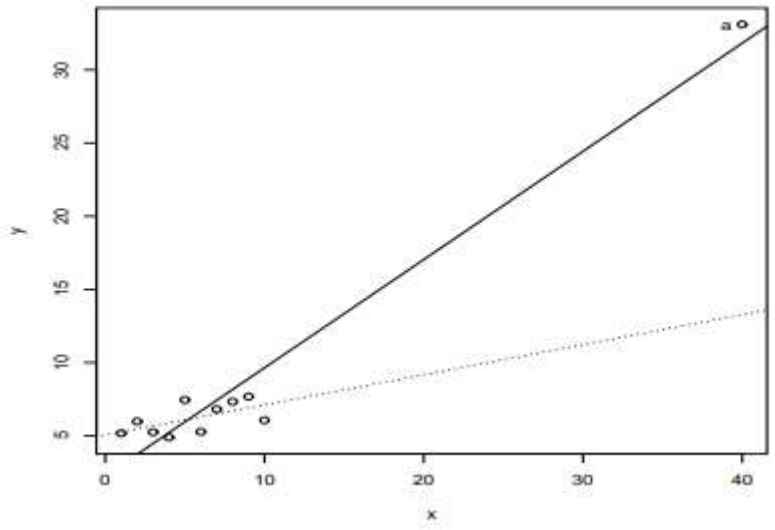
ANALISIS DE REGRESION

ANALISIS DE INFLUENCIA

I.- Introducción

Las medidas de influencia son estadísticas que nos permiten detectar e identificar observaciones influyentes sobre los resultados del modelo de regresión lineal múltiple.

Es importante para el analista de los datos tratar de identificar observaciones influyentes e investigar el efecto que ejercen sobre uno o varios aspectos del análisis de regresión (estimaciones de los parámetros, varianza estimada de las estimaciones de los parámetros, los valores ajustados de la variable de respuesta entre otros.



R

r=0.42

r=-0.83 y= b

II.- Observaciones atípicas, Puntos de Balanceo y observaciones Influyentes

Tres conceptos que se encuentran relacionados:

i) Observación atípica:

Es denominada también como observación discordante, aberrante, inusual (outlier), presentan un comportamiento diferente a las demás observaciones en el campo de las variables regresoras o es aquella observación con mayor residuo estudentizado en el espacio de la variable de respuesta. Este tipo de observaciones pueden afectar los resultados del modelo de regresión estimado mediante el método de los Mínimos Cuadrados Ordinarios.

ii) Punto de Balanceo;

Es la observación para la cual su vector xi se encuentra lejos del resto de las observaciones, también se dice que es la observación con mayor valor en la diagonal de la matriz sombrero H cuyos elementos de la diagonal se expresan como:

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

Que contiene i-ésimo elemento de la diagonal de la matriz sombrero.

Se le considera una observación discordante o atípica en el espacio de las variables regresoras.

iii) Observación influyente

Es aquella que individualmente o en conjunto ejercen una influencia excesiva en el ajuste de la ecuación de un modelo de regresión. Estas observaciones se ubican en el espacio de las variables regresoras y de respuesta.

A una observación que provoca que los estimadores de la regresión sean sustancialmente diferentes de lo que serían si se eliminara a tal observación del conjunto de datos se le llama observación influyente. Las observaciones que son aberrantes (outliers) o tienen una palanca grande no necesariamente son influyentes, mientras que las observaciones influyentes son aberrantes y tienen palanca grande.

NOTA:

1º *Una observación discordante puede ser una observación extrema (pero es extrema en relación a su valor de residuo).*

2º *Una observación extrema no es necesariamente una observación discordante.*

3º *Una observación influyente no es necesariamente una observación discordante.*

4º *Observaciones discordantes no necesariamente son observaciones influyentes.*

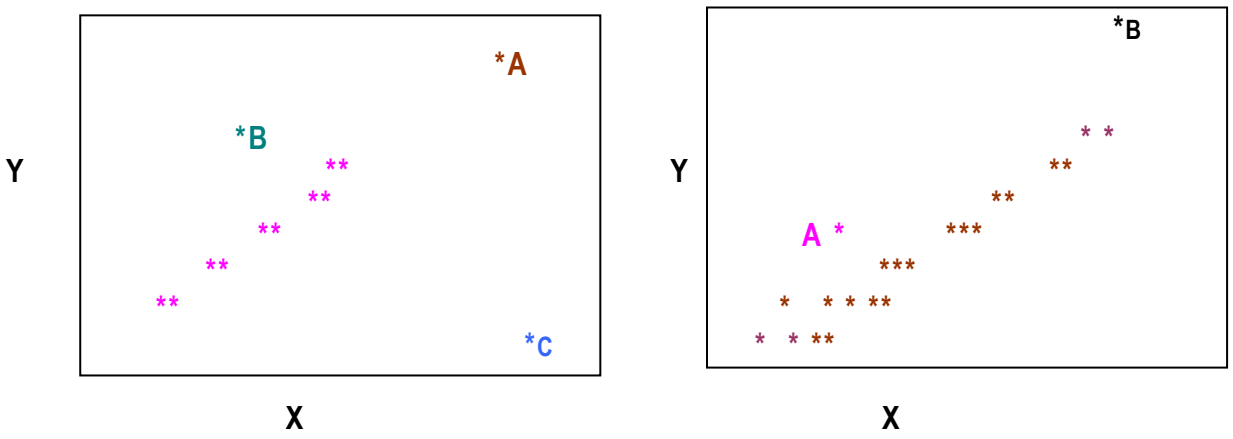
5º *Los puntos de balanceo por lo general tienen residuos con valores pequeños.*

6º *Las observaciones influyentes no son necesariamente son puntos de balanceo.*

7º Los puntos de balanceo no son necesariamente observaciones influyentes.

Sin embargo un punto de balanceo es una observación potencialmente influyente.

Se tiene:



III.- MEDIDAS DE INFLUENCIA

Existen diversas medidas para identificar observaciones influyentes, como por ejemplo las medidas que se basan en :

a. ***Los residuos***: gráficos mostrados en el análisis de los residuos.

b. ***Las observaciones que se encuentran distantes en el espacio de las variables regresoras y de respuesta***: Las estadísticas que se utilizan en este caso son elementos de la diagonal de la matriz sombrero (puntos de balanceo),

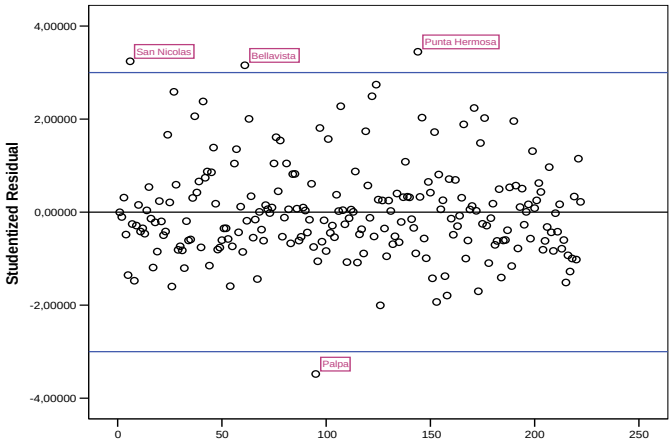
c. ***La curva de influencia:*** distancia de Cook, distancia de Welsch, Welsch- Kuh denominada DFFITS por Belsley en 1980), etc.

d. ***El volumen de los elipsoides confidenciales***: Covratio,

e. Un solo Coeficiente de regresión: DFBETAS.

3.1 Medidas basadas en los residuos:

Gráficos de los Residuos escalados.



Medidas que se basan en las observaciones que se encuentran distantes en el espacio de las variables regresoras y de respuesta

3.2.1. El i-ésimo elemento de la Diagonal de la matriz H "BALANCEO"

Se tiene que: $\mathbf{H = X'(X'X)^{-1}X'}$

Los elementos de esta matriz, se pueden denotar por:

$h_{ij} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j$ donde i,j =1,2,...,n,

por lo tanto el i-ésimo elemento de la matriz H es:

$h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ con i=1,2,...,n,

Estos elementos juegan un rol importante en la determinación de los valores ajustados ($\hat{y}$), en las magnitudes de los residuos y en la estructura de la matriz de varianza-covarianza.

Existen diversas propuestas acerca del punto de corte, utilizaremos la de Hoaglin y Welsch (1978):

Se tiene la siguiente propiedad de H:

$$tr(\mathbf{H}) = rango(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p$$

El promedio de los elementos de la diagonal de H es p/n.
En un diseño balanceado todos los h_{ii} son iguales a su valor promedio p/n, por lo tanto si:

$h_{ii} > \frac{2p}{n}$ será clasificado como un punto de balanceo.

Observaciones:

Siempre es conveniente diferenciar las fuentes o causas por la que se presentan observaciones influyentes. Una observación puede ser influyente sobre algunos o todos los resultados del análisis del modelo de regresión lineal múltiple, debido a que:

- Es discordante en la variable de respuesta.
- Es una observación discordante y punto de balanceo.
- En un punto de balanceo, en el espacio de las variables regresoras.

3.3. Medidas Basadas en la curva de influencia,

3.3.1 Distancia de Cook,

La medida de distancia de Cook es un diagnóstico de eliminación; es decir, mide la influencia de la i-ésima observación si se eliminara de la muestra. Está medida fue sugerida por Cook y se define como:

$$C_i = D_i(M, c) = \frac{\left\{\hat{\beta}_{(i)} - \hat{\beta}\right\}^T M \left\{\hat{\beta}_{(i)} - \hat{\beta}\right\}}{c}$$

con $M = X^T X$ y c = pCMRes

La distancia de Cook puede también ser escrita de la siguiente forma:

$$D_i(\mathbf{X}'\mathbf{X}, pMS_{\text{Res}}) \equiv D_i = \frac{\left(\hat{\beta}_{(i)} - \hat{\beta}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}_{(i)} - \hat{\beta}\right)}{pMS_{\text{Res}}}, \quad i = 1, 2, \ldots, n$$

esta medida es la distancia al cuadrado entre el estimador obtenido cuando se ha retirado la i-ésima observación y el estimador calculado en base a todas las observaciones.
La magnitud de esta distancia; es decir, Di puede evaluarse comparando con Fα,p,n-p.
Si Di = $F_{0.5,p,n-p}$ entonces al eliminar el punto i se movería $\hat{\beta}_{(i)}$ hacia la frontera de una región de confianza aproximada del 50% para β, para todos los datos del modelo.

se tiene que $F_{0.5,n,n-p} \approx 1$. Entonces si:

Di > 1 la i-ésima observación es influyente; es decir, influye considerablemente sobre los estimadores mínimos cuadrados de **β** del modelo.
La estadística Di puede expresarse también como sigue:

$$D_i = \frac{r_i^2}{p}\frac{\text{Var}(\hat{y}_i)}{\text{Var}(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}, \quad i = 1, 2, \ldots, n$$

Así, Di está formada por un componente que refleja lo bien que se ajusta el modelo a la i-ésima observación Yi' y un componente que mide lo alejado que el punto está, del resto de los datos

3.3.3. Estadística DFFITS:

También se puede investigar la influencia de la eliminación de la i-ésima observación sobre el valor predicho o ajustado. Esta es una propuesta hecha por Belsley en 1980:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}} \quad i = 1,2, \ldots, n$$

Donde $\hat{y}_{(i)}$ es el valor ajustado de y_i, obtenido sin usar la i-ésima observación.

Así, DFFITSi es el número de desviaciones estándar en que el valor ajustado de la variable de respuesta cambia si la i-ésima observación es retirada.

Para los cálculos se puede usar

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{e_i}{S_{(i)}(1-h_{ii})^{1/2}}$$

$$= \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i$$

Donde:
t_i es el residual R de Student y $[h_{ii}/(1-h_{ii})]^{½}$ es el balanceo de la i-ésima observación.

Así, una observación para el cual:

$$|DFFITS_i| > 2\sqrt{\frac{p}{n}}$$

significa, que la i-ésima observación influye sobre el valor ajustado de la variable de respuesta y merece ser investigada.

Resumiendo: Las medidas que se basan en la curva de influencia son medidas que miden el cambio en el centro del elipsoide confidencial, cuando la i-ésima observación es retirada.

3.4. Medidas Basadas en el Volumen de Elipsoides Confidenciales

Miden el cambio en el volumen del elipsoide confidencial, cuando la i-ésima observación es retirada.

3.4.1 COVRATIO:

Se puede evaluar la influencia de la i-ésima observación, comparando la varianza estimada de los parámetros estimados del modelo considerando todas las observaciones, con la varianza estimada de los parámetros estimados cuando se ha retirado la i-ésima observación. Si el rango de la segunda matriz es “p” las

matrices son definidas positivas; existen varios métodos para comparar matrices definidas positivas, como la razón de sus trazas o la razón de sus determinantes.
Podemos utilizar el determinante de la matriz de covarianza como una medida de precisión, la varianza estimada be ^β:

$$COV(\hat{\beta}) = |Var(\hat{\beta})| = | \sigma^2 (X'X)^{-1}$$

Para evaluar la importancia de la i-ésima observación sobre la precisión de la estimación se puede definir:

$$COVRATIO_i = \frac{|(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} S^2_{(i)}|}{|(\mathbf{X}'\mathbf{X})^{-1} MS_{\text{Res}}|}, \quad i = 1, 2, \ldots, n$$

COVRATIO>1, la i-ésima observación mejora la precisión de la estimación de los parámetros.

COVRATIO <1, la i-ésima observación disminuye o degrada la precisión de la estimación de los parámetros.

También puede escribirse de la siguiente manera:

$$COVRATIO_i = \frac{\left(S^2_{(i)}\right)^p}{MS^p_{\text{Res}}} \left(\frac{1}{1 - h_{ii}}\right)$$

Si los valores de hii son altos entonces $COVRATIO_i$ será grande
Si ambos, el cuadrado del residuo estudentizado y el valor de la diagonal de la matriz sombrero son grandes o pequeños COVRATIO se aproxima a uno.

Se sugiere que si:

COVRATIOi > 1+(3p/n)
COVRATIOi < 1-(3p/n), entonces la i-ésima observación se considera influyente en la covarianza estimada de los parámetros estimados del modelo.

3.5. Medidas Basadas en un solo Coeficiente de regresión: DFBETAS

Es una estadística que permita medir cuanto cambia la estimación del j-ésimo parámetro del modelo, en unidades estándares, si la i-ésima observación es retirada del análisis; la estadística está dada por:

$$\mathbf{DFBETAS}_{(ji)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{Var(\hat{\beta}_j)}}$$

donde:

$\hat{\beta}_{j(i)}$ es la estimación del j-ésimo parámetro del modelo, calculado sin la i-ésima observación.

Para su cálculo puede escribirse en función de los residuales como:

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{e_i}{S_{(i)}(1 - h_{ii})}$$

$$= \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{t_i}{\sqrt{1 - h_{ii}}}$$

donde ti es el residual de R de Student.

Los $DFBETAS_{j,i}$ mide tanto el balanceo ($r_{j,i} / \sqrt{r'_j r_j}$ que es una medida del impacto de la i-ésima observación sobre $\hat{\beta}_j$) como el efecto de un residual grande.

Se sugiere que si:

$$\left|\mathbf{DFBETAS}_{(ji)}\right| > \frac{2}{\sqrt{n}}$$

Entonces la i-ésima observación tiene influencia apreciable en la estimación del j-ésimo parámetro del modelo y debe ser examinada.

TAREAS: