PESQUIMAT, Revista de la Fac.CC.MM. de la
Universidad Nacional Mayor de San Marcos
Vol.III, $N°1$, págs.16-28,Lima - Perú. Julio 2000

MEDIDAS DE LA CLASE J_I^{DET} PARA DETECTAR CONJUNTOS DE OBSERVACIONES INFLUYENTES

M. Estela Ponce Aruneri
Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas

ABSTRACT Las medidas de la Clase J_I^{DET} [4] propuestas en este artículo, son la generalización de algunas medidas de influencia presentadas para el caso de Regresión Lineal Univariado (Jones y Ling, 1988).

Estas medidas nos permite detectar conjuntos de observaciones influyentes y medir la influencia que dichos conjuntos ejercen sobre los diversos resultados del Análisis de Regresión Lineal Multivariado.

Se muestra una Aplicación utilizando algunas Variables de la Encuesta de Seguimiento de Consumo en Hogares, de las principales ciudades del Perú [10].

1. INTRODUCCIÓN

Es importante para un analista de datos, estar en la capacidad de identificar observaciones o conjuntos de observaciones influyentes, así como evaluar sus efectos sobre los diversos aspectos del Análisis de Regresión.

Existe un gran número de medidas estadísticas propuestas para identificar y medir conjuntos de observaciones influyentes en el Modelo de Regresión Lineal Univariado pero poco se conoce de las medidas propuestas para el Modelo de Regresión Lineal Multivariado.

2. MODELO DE REGRESIÓN LINEAL MULTIVARIADO

El Análisis de Regresión Multivariado investiga y modela la relación entre un conjunto de variables de respuestas y un conjunto de variables regresoras, mediante el modelo de regresión lineal multivariado, el que es útil para evaluar los efectos de las variables regresoras sobre las variables de respuestas.

Sea

$$Vec\,Y = (I_r \otimes X)\,Vec\,\beta + Vec\,\epsilon$$

Con

$$E(Vec \quad \epsilon) = 0 \quad \text{y} \quad Cov(Vec \quad \epsilon) = \Sigma \otimes I_n$$

donde la :

1)Representación Vectorial de la matriz de observaciones de las "r" variables de respuesta en cada uno de los "n" individuos es:

$$\mathbf{Vec\,Y} = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \\ \cdot \\ \cdot \\ \cdot \\ Y_{(r)} \end{bmatrix}_{nr\times 1}$$

2)La representación vectorial de la parte determinística del modelo esta dada por:

a)

$$(\mathbf{I_r} \otimes \mathbf{X}) = \begin{bmatrix} X & 0 & 0 & \cdots & 0 \\ 0 & X & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdot & \cdot & X \end{bmatrix}_{n\times rp}$$

Es el producto de la matriz de Identidad de orden $r \times r$ con la matriz de variables regresoras de orden $n \times p$.

b) Representación vectorial de la matriz de parámetros del modelo:

$$\mathbf{Vec}\,\beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \beta_{(r)} \end{bmatrix}_{pr\times 1}$$

3) La representación vectorial de la matriz de perturbaciones aleatorias:

$$\mathbf{Vec}\,\epsilon = \begin{bmatrix} \epsilon_{(1)} \\ \epsilon_{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_{(r)} \end{bmatrix}_{nr\times 1}$$

Observaciones:

$$\mathbf{E}(\mathbf{Vec}\,\epsilon) = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = 0$$

y $$\mathbf{Cov}(\mathbf{Vec}\,\epsilon) = \begin{bmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1r}I_n \\ \sigma_{21}I_n & \sigma_{22}I_n & \cdots & \sigma_{2r}I_n \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{r1}I_n & \sigma_{r2}I_n & \cdots & \sigma_{rr}I_n \end{bmatrix} = \Sigma \otimes I_n$$

Estimación de los Parámetros del Modelo

Los métodos comúnmente utilizados para estimar los parámetros del modelo de regresión lineal multivariado son el de Mínimos Cuadrados y el de Máxima Verosimilitud.

El Método de los Mínimos Cuadrados permite encontrar el estimador para el vector de parámetros del modelo, minimizando la suma de cuadrados del vector de perturbaciones aleatorias con respecto al vector de parámetros de dicho modelo.

Se tiene :

$$Vec\,\epsilon = VecY - (I_r \otimes X)Vec\,\beta \tag{2.1}$$

Para obtener el estimador mínimo cuadrático $VecB$, minimizamos la suma de cuadrados :

$$S(Vec\,\beta) = [Vec\epsilon]^T \quad [Vec\epsilon] \tag{2.2}$$

Reemplazando (2.1) en (2.2), se obtiene una nueva expresión para la suma de cuadrados del vector de perturbaciones aleatorias; de ella se obtienen las derivadas parciales con respecto al vector de parámetros

del modelo e igualando al vector cero, encontramos el siguiente estimador mínimo cuadrático :

$$VecB = [I_r \otimes (X^T X)^{-1} X^T] \, VecY$$

siempre que:

$$[(I_r \otimes X)^T (I_r \otimes X)]^{-1} \qquad \text{exista.}$$

Matriz Leverage H. Se le conoce así debido a que al examinar los elementos de su diagonal denominados "leverage", permite detectar observaciones que pueden ser consideradas como observaciones "high-leverage" en el espacio de las variables regresoras.

Descomposición de la matriz Leverage. :Sea $I = \{i_1, i_2, \ldots, i_m\}$ un conjunto que contiene los índices de las "m" posibles observaciones influyentes.
Sin pérdida de generalidad; podemos considerar las "m" observaciones, como las últimas filas de la matriz X, luego dicha matriz puede particionarse de la siguiente forma:

$$\mathbf{X} = \begin{bmatrix} X_{(I)} \\ - - - \\ X_I \end{bmatrix}$$

donde :

$X_{(I)}$ de orden $(n-m) \times p$, es la matriz que contiene las "$n-m$" observaciones.

X_I de orden $(m \times p)$ es la matriz que contiene las "m" observaciones retiradas para realizar el análisis de influencia.

Como $H = X(X^T X)^{-1} X^T$, la partición de la matriz de predicción se puede representar como:

$$\mathbf{H} = \begin{bmatrix} X_{(I)}(X^T X)^{-1} X_{(I)}^T & X_{(I)}(X^T X)^{-1} X_I^T \\ \\ X_I(X^T X)^{-1} X_{(I)}^T & X_I(X^T X)^{-1} X_I^T \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} H_{(I)} & H_{(I),I} \\ \\ H_{I,(I)} & H_I \end{bmatrix}_{n \times n}$$

donde:

$H_{(I)}$ es de orden $(n-m) \times (n-m)$ $\quad H_I$ es de orden $(m \times m)$

$H_{(I),I}$ es de orden $(n-m) \times m$ $\quad H_{I,(I)}$ es de orden $m \times (n-m)$

Residuos Multivariados

El análisis de los residuos permite validar los supuestos del modelo de regresión (normalidad, no autocorrelación, varianza constante, etc.), es un método efectivo para detectar deficiencias en el modelo, utilizando diversos tipos de gráficos. La validez de los resultados obtenidos en el análisis de regresión, son verificados luego de realizar un examen detallado y cuidadoso de los residuos.

Los residuos ayudan a detectar observaciones que pueden ser consideradas como *"outliers"*.

Con el mismo objetivo que la matriz leverage H fue particionada, es necesario descomponer la matriz de residuos E:

$$\mathbf{E} = \begin{bmatrix} E_{(I)} \\ - - - \\ E_I \end{bmatrix} .$$

donde: $E = Y$ - $\hat{Y}$

$E_{(I)}$ de orden $(n-m) \times r$.

E_I de orden $m \times r$, contienen las "m" observaciones a ser evaluadas.

Se tiene además que una partición para $Q = E(E^T E)^{-1} E^T$, similar a las de H y E, es descrita por Barret y Ling [4].

$$\mathbf{Q} = \begin{bmatrix} E_{(I)}(E^T E)^{-1} E_{(I)}^T & E_{(I)}(E^T E)^{-1} E_I^T \\ E_I(E^T E)^{-1} E_{(I)}^T & E_I(E^T E)^{-1} E_I^T \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} Q_{(I),(I)} & Q_{(I),I} \\ Q_{I,(I)} & Q_{I,I} \end{bmatrix}$$

donde:

$Q_{(I)}$ es de orden $(n-m) \times (n-m)$ Q_I es de orden $(m \times m)$

$Q_{(I),I}$ es de orden $(n-m) \times n$ $Q_{I,(I)}$ es de orden $m \times (n-m)$

3. MEDIDAS DE LA CLASE J_I^{DET}

Barret y Ling (1992),presentaron dos clases de medidas de influencia para la regresión multivariada: J_I^{tr} y J_I^{det}. Esta caracterización de las medidas de influencia permite descomponer la influencia total de un

conjunto de observaciones, en dos componentes: la componente leverage y la componente residual; facilitando además el cálculo numérico de dichas componentes cuando se dispone de varios posibles conjuntos de observaciones influyentes.

La generalización de las medidas de influencia multivariada J_I^{det}, se representan de la siguiente forma:

$J_I^{det}\ (\ f\ ;\ a, b) = f(n, p, r, m)\ det[(I - H_I - Q_I)^a (I - H_I)^b]$

según Barret y Ling [4].

donde:

f : es una función basada en el orden de las matrices del Modelo de Regresión Multivariado, (n, p, r) y del conjunto "m" de observaciones que han de ser retiradas.

I : es la matriz de identidad de orden "m".

a y b : son valores enteros asociados con la componente residual y leverage.

Se mostrará que las medidas de influencia multivariadas: Andrews y Pregibon, Covratio y Fvaratio pertenecen a la Clase J_I^{det}. Estas medidas muestran el cambio del volumen del elipsoide confidencial cuando el I-ésimo conjunto de observaciones es retirado; es decir miden la razón, del volumen cuando se han retirado m observaciones en relación al volumen con todas las n observaciones.

1° **ESTADISTICA DE ANDREWS Y PREGIBON.** Mide la influencia de un conjunto de observaciones sobre las variables de regresoras y de respuesta; permite detectar observaciones que se encuentran lejos del resto y que pueden ser consideradas como outliers. Andrews y Pregibon sugirieron la siguiente razón:

$$A\,P_I = \frac{det[Z_{(I)}^T Z_{(I)}]}{det[Z^T Z]} \tag{3.1}$$

donde $z = [X \quad Y]$ es la matriz que considera, la matriz X de variables regresoras y la matriz Y de variables de respuesta.

Realizando el producto de las matrices $(Z^T Z)$ y hallando su determinante se obtiene:$det(Z^T Z) = det(X^T X)\ det(E^T E)$ similarmente $det(Z_{(I)}^T Z_{(I)}) = det(X_{(I)}^T X_{(I)})\ det(E_{(I)}^T E_{(I)})$; reemplazando estas expresiones en (3.1) y utilizando propiedades del álgebra matricial se tiene:

$$A\,P_I = det(I - H_I)\ det[(I - H_I - Q_I)(I - H_I)^{-1}]$$

$$AP_I = det(I - H_I - Q_I) = J_I^{det}(1; 1, 0) \tag{3.2}$$

Lo que muestra que la estadística de Andrews y Pregibon es una medida de influencia que pertenece a la clase J_I^{det}.

2° **COVRATIO.** mide la influencia de un conjunto de observaciones, en la matriz de covarianzas de los parámetros estimados cuando m observaciones son retiradas.

$$COVRATIO_I = \frac{Cov(Vec\, B_{(I)})}{Cov(Vec\, B)} = \frac{det[S_{(I)} \otimes (X_{(I)}^T X_{(I)})^{-1}]}{det[S \otimes (X^T X)^{-1}]}$$

con:

$$S_{(I)} = \frac{E_{(I)}^T E_{(I)}}{n - p - m} \quad \text{y} \quad S = \frac{E^T E}{n - p} \tag{3.3}$$

Reemplazando las expresiones dadas en (3.3), en $COVRATIO_I$ y haciendo uso de las propiedades de determinantes llegamos a:

$$COVRATIO_I = \left(\frac{n-p}{n-p-m}\right)^{rp} [det(I - H_I - Q_I)]^m det(I - H_I)^{-(r+p)}$$

$$COVRATIO_I = J_I^{det}\left(\left(\frac{n-p}{n-p-m}\right)^{rp}; p, -(r+p)\right) \tag{3.4}$$

La medida $COVRATIO_I$ pertenece a la clase de medidas de influencia J_I^{det}.

3° **FVARATIO.** Es una medida de influencia propuesta por Belsley,1980 [2] que compara la varianza generalizada reducida de $Vec(Y)$ cuando "m" observaciones son retiradas, con la varianza generalizada del total de observaciones. Mide la influencia sobre las estimaciones de la matriz de covarianzas del vector de respuestas ajustado.

$$FVARATIO_I = \frac{det[Cov[Vec(X_I\, B_{(I)})X_I\, B]]}{det[Cov[Vec(X_I\, B)]]}$$

trabajando separadamente el numerador y el denominador se tiene:

$$FVARATIO_I = \frac{det[S_{(I)} \otimes (I - H_I)^{-1} H_I]}{det[S \otimes H_I]}$$

reemplazando los valores de $S_{(I)}$ y S se obtiene:

$$FVARATIO_I = \left(\frac{n-p}{n-p-m}\right)^{rm} det(I - H_I - Q_I)^m$$

$$det(I - H_I)^{-(r+m)}$$

aplicando propiedades de determinantes:

$$FVARATIO_I = J_I^{det}\left(\left(\frac{n-p}{n-p-m}\right)^{rm}; m, -(r+m)\right) \tag{3.5}$$

4. DESCOMPOSICIÓN DE LAS MEDIDAS DE INFLUENCIA

Facilita la identificación de los conjuntos de observaciones influyentes.

$$Medida\ de\ Influencia = f(.)det(L_I R_I)$$

donde :

L_I es la matriz leverage, una función de H_I, que no considera los residuos.

R_I considera al resto de elementos de la matriz producto y se le llama matriz de residuos.

La elección para construir L_I y R_I depende de la preferencia que tenga el investigador.

En este caso utilizaremos la versión multivariada de los residuos estudentizados por las consideraciones señaladas por Chatterjee y Hadi [8].

a) Para la medida de Andrews-Pregibon; tomando el recíproco de la medida dada en (3.2) y utilizando propiedades de determinantes se tiene:

$$L_I = (I - H_I)^{-1} \qquad R_I = (I - H_I - Q_I)^{-1}(I - H_I)$$

las componentes leverage y residual para el recíproco de la medida de Andrews-Pregibon.

b) Para la medida COVRATIO, tomando el recíproco de la medida dada en (3.4) y utilizando propiedades de determinantes se tiene:

$$L_I = (I - H_I)^r \qquad R_I = (I - H_I - Q_I)^{-1}(I - H_I)^p$$

las componentes leverage y residual del recíproco de la razón de Covarianzas.

c) Para la medida FVARATIO, procediendo de forma similar en (3.5) se obtiene:

$$L_I = (I - H_I)^r \qquad R_I = (I - H_I - Q_I)^{-1} (I - H_I)^m$$

Estas tres medidas de influencia proporcionan información, de la contribución de la componente leverage y residual a la influencia total.

5. GRÁFICOS PARA LA COMPONENTE LEVARAGE Y RESIDUAL

El análisis se complica cuando se tiene un conjunto de "m" observaciones influyentes, debido a que la información de las componentes son matrices, las cuales deberán ser reducidas a un solo escalar. Para evaluar la contribución relativa de las componentes a la influencia total se tiene:

$$det(L_I\, R_I) = det(L_I)\, det(R_I)$$

Los gráficos son de gran ayuda para evaluar la contribución de la componente leverage y residual, ya que nos permiten identificar conjuntos de observaciones cuyos efectos simultáneos neutralizan o eliminan el efecto de otras observaciones.

Se pueden trazar los siguientes gráficos para realizar el análisis de la contribución de las componentes:

1° Gráfico del $det(L_I)$ versus $det(R_I)$, permiten examinar los roles de la componente leverage y residual.

2° $Log[det(L_I)]$ versus $Log[det(R_I)]$; son los logaritmos de las contribuciones relativas de las componentes. Muestran las observaciones ampliamente dispersas, facilitando la ubicación de los conjuntos que más contribuyen a la componente leverage como a la componente residual.

OBSERVACIÓN:

La contribucin a la influencia total se obtiene :

$$Influencia\,Total = Log[det(L_I)] + Log[det(R_I)]$$

6. APLICACIONES

DESCRIPCIÓN DE LOS DATOS. Se recopiló datos de los resultados que se obtuvieron en la Encuesta de Seguimiento del Consumo de Hogares (ENSECO) en las principales ciudades del Perú; encuesta realizada en Junio de 1991 y cuyos resultados fueron publicados por el INEI en Junio de 1992.

El objetivo del INEI al realizar dicha encuesta fue elaborar la estructura de los gastos e ingresos en los hogares peruanos. La información la obtuvieron de un muestra de hogares en cada una de las siguientes ciudades: Abancay, Arequipa, Ayacucho, Cajamarca, Cerro de Pasco, Cuzco, Chachapoyas, Chiclayo, Chimbote, Huancavelica, Huancayo, Huánuco, Huaraz, Ica, Iquitos, Lima, Moquegua, Moyobamba, Piura, Puerto Maldonado, Puno, Tacna, Tumbes, Trujillo y Pucallpa. Por motivos no conocidos no se publicaron los resultados de la Ciudad de Cerro de Pasco.

MODELO DE REGRESIÓN PROPUESTO PARA LA APLICACIÓN

Variables utilizadas:

Y_1: Gastos en alimentos y bebidas, Y_2: Gastos en transporte y comunicaciones.
X_1: Promedio de perceptores por hogar, X_2: Promedio de cuartos por hogar.

Luego, el modelo estimado es :

$$\begin{aligned} \hat{Y}_1 &= -122.79 + 174.66X_1 - 16.95X_2 \\ \hat{Y}_2 &= -85.29 + 53.13X_1 + 1.22X_2 \end{aligned}$$

Realizado el ajuste del modelo es necesario realizar un análisis de influencia que permita detectar la existencia o no de conjuntos de observaciones influyentes.
Cuando el I-ésimo conjunto de tamaño dos es retirado se obtiene el siguiente cuadro, que muestra las ciudades detectadas como influyentes $(*)$:

MEDIDAS DE INFLUENCIA DE LA CLASE J_I^{det}

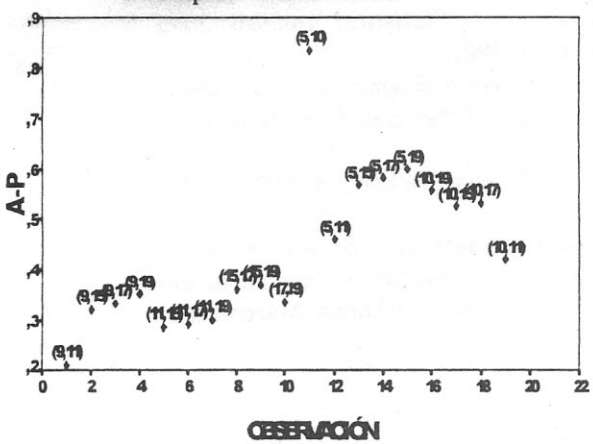
CIUDAD	Andrews-Pregibon	COVRATIO	FVARATIO
Cuzco-Huancayo	*		
Lima-Pto.Maldonado		*	*
Lima-Moyobamba		*	*
Moyobamba-Pto.Maldonado		*	*
	Outlier en el espacio de X e Y	Influye en Cov (B)	Influye en Cov($\hat{Y}$) (gastos)

Logaritmos de las Componentes Reescaladas de las Medidas de Influencia. Conjuntos de Observaciones que más contribuyen a las Componentes Leverage, Residual e Influencia Total

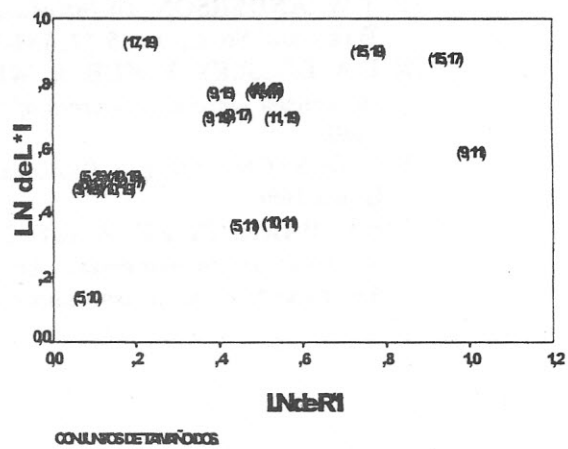
OBSERVACIÓN	CIUDAD	A-P			COVRATIO			FVARATIO		
		L	R	I.T	L	R	I.T	L	R	I.T
(9-11)	Huancavelica/ Huánuco		X			X	X		X	X
(15-17)	Lima/Moyobamba			X						
(17-19)	Moyobamba / Pto.Maldonado	X								
(5-10)	Cuzco / Huancayo				X			X		

Los siguientes gráficos corresponden a los datos de las dos tablas presentadas, pero para la medida de influencia de Andrews-Pregibon:

Medida de Adrews - Pregibon para componentes de tamaño 2. Principales Ciudades del Perú



Logaritmo de la reescalada Leverage y Residual para la medida de Andrews -Pregibon Principales ciudades del Perú



Además dichos hogares presentan menor, promedio de perceptores y promedio de cuartos por hogar.

* Huánuco, tiene el mayor número promedio de perceptores por hogar.

* Lima, es la ciudad cuyos hogares tienen el mayor promedio de cuartos por hogar, el mayor número promedio de perceptores por hogar y dichos hogares tienen los mayores gastos en alimentos y bebidas así como en transporte y comunicaciones.

* Moyobamba y Puerto Maldonado, son ciudades cuyo promedio de perceptores y gastos en alimentos y bebidas son relativamente altos.

6. CONCLUSIONES

1. Se muestra que las medidas multivariadas que pertenecen a la clase J_I^{det}: Andrews-Pergibon, Covratio y Fvaratio miden la influencia sobre la precisión de los estimadores.
2. Las medidas de influencia multivariadas identifican conjunto de observaciones influyentes, pero no permiten distinguir entre una observación outlier y una observación high - leverage; por lo que es necesario descomponer dichas medidas en sus componentes leverage y residual y luego realizar un reescalamiento, para evaluar la influencia real de cada una de dichas componentes en la influencia total.
3. La aplicación muestra:
   (a) Que es más conveniente y revelador utilizar la descomposición de las medidas multivariadas (a utilizar las medidas de influencia) para detectar conjuntos de observaciones que puedan ser influyentes en el modelo de regresión lineal multivariado.
   (b) El conjunto de observaciones que proporciona mayor contribución a la influencia total no necesariamente aporta mayor contribución a la componente Leverage y/o componente Residual; de allí la importancia de analizar separadamente la contribución de las componentes.

BIBLIOGRAFIA

[1] T.W. ANDERSON *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons., Inc. 5-57, 154-175. New York.,1966.

[2] D.A. BELSLEY; E. KUH; R. WELSCH *E.Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons. New York. 1980.

[3] O. BUSTOS, *Outliers y Robustez.* Informe de Matemática Serie B.$N°$044, 1-14. Brasil.1988.

[4] B.E. BARRETT RUCE; R.F. LING *General Classes of Influence Measures for Multivariate Regression.* American Statistical Association. Journal of the American Statistical Association, Vol.87 $N°$417. Theory Methods. March 1992.

[5] R.D. COOK. *Influential Observations in Linear Regression.* Journal of the American Statistical Association. Vol.74, $N°$365,169-174. March 1979.

[6] R. D. COOK *Assessment of Local Influence.* JR Statistical Soc. B., Vol.48,$N°$2,133-169. 1986.

[7] R. D. COOK AND S. WEISBERG *Residuals and Influence in Regression.* Chapman and Hall. London. 1982.

[8] S. CHATTERJEE;A. HADI *Influential Observation, High Leverage Points and Outliers in Linear Regression.* Statistical Science. Vol.1 $N°$3,379-416. 1986.

[9] S. CHATTERJEE;A. HADI *Sensitivity Analysis in Linear Regression.* John Wiley. New York. 1988.

[10] INEI. *Estructura de Ingresos y Gastos de los Hogares.* Tomo 1 a 24."ENSECO 91". Junio de 1992.

[11] M. E. PONCE *Medidas de Influencia en Regresión Multivariada.* Tesis para Maestría en Estadística. UNMSM.FCM. Lima-Perú. 1999.

[12] SAS/ IML. *User's Guide: Statistics. Version 6.03.* Edition 1985.

[13] S. R. SEARLE *Matrix Algebra Useful for Statistics.* John Wiley and Sons. New York. 1982.