PESQUIMAT Revista de la Fac.CC.MM.de la
Universidad Nacional Mayor de San Marcos
Vol.III, $N°2$, Noviembre 2000, pag.51-64. Lima-Perú

MEDIDAS DE INFLUENCIA QUE SE BASAN EN LA CURVA DE INFLUENCIA

M. Estela Ponce Aruneri [1]

ABSTRACT. *Barret y Ling en 1992, presentaron las medidas multivariadas que se basan en la curva de influencia y a las que denominaron medidas de la clase J_I^{tr} [1]. En este arículo presentaremos tres de estas medidas: Distancia de Cook, DFFITS y Distancia de Welsch [2], las que permiten detectar conjuntos de observaciones influyentes en las estimaciones de los parámetros, variables de respuestas y matriz de covarianzas de los residuos del modelo de regresión lineal multivariado.*

Mostramos una aplicación en base a los datos de la Encuesta de Seguimiento del Consumo de Hogares (ENSECO 91) [6].

1. INTRODUCCIÓN

Las medidas de influencia son medidas estadísticas que permiten detectar e identificar "observaciones influyentes" [3]. Estas observaciones individualmente o colectivamente tienen influencia en el ajuste del modelo de regresión lineal multivariado (estimaciones de: los parámetros, las variables de respuestas, la matriz de covarianza de los residuos, etc.).

La caracterización de las medidas que se basan en la curva de influencia en la clase J_I^{tr}, permite descomponer la influencia total que ejerce el conjunto de observaciones, en dos componentes: leverage y residual, facilitando los cálculos numéricos así como la interpretación de las mismas.

La clase J_I^{tr} de medidas de influencia multivariadas que se basan en la curva de influencia [1], se representa como:

[1]Universidad Nacional Mayor de San Marcos.Facultad de Ciencias Matemáticas
e-mail: mepaunmsm@mixmail.com.pe

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

donde :

f : es una función basada en el orden de las matrices del Modelo de Regresión Lineal 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.

2. MODELO DE REGRESIÓN LINEAL MULTIVARIADO

Es útil para evaluar los efectos de las variables regresoras sobre las variables de respuestas[7]:

$$VecY = (I_r \otimes X)Vec\beta + Vec\varepsilon$$

con $E(Vec\varepsilon) = 0$ y $Cov(Vec\varepsilon) = \sum \otimes I_n$

donde:

$VecY$ es de orden $nr \times 1$ y es la representación Vectorial de la matriz de observaciones de las "r" variables de respuesta en cada uno de los "n" individuos.

X es la matriz de orden $n \times p$, de variables regresoras.

$Vec\beta$ es de orden $pr \times 1$, es la representación vectorial de la matriz de parámetros del modelo.

$Vec\varepsilon$ es de orden $nr \times 1$, es la representación vectorial de la matriz de perturbaciones aleatorias; con media 0 y matriz de covarianza $\sum \otimes I_n$.

2.1 Estimación de los Parámetros del Modelo

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 [7].

Se tiene : $Vec \quad \varepsilon = Vec \quad Y - (I_r \otimes X)Vec \quad \beta \qquad (1)$

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

$$S(Vec \quad \beta) = [Vec \quad \varepsilon]^T [Vec \quad \varepsilon] \qquad (2)$$

Reemplazando (1) en (2), se obtiene una 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 y se iguala al vector cero, encontrándose el siguiente estimador mínimo cuadrático [7]:

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

siempre que exista : $[(I_r \otimes X)^T (I_r \otimes X)]^{-1}$

3. MEDIDAS DE INFLUENCIA QUE SE BASAN EN LA CURVA DE INFLUENCIA

Una clase importante de medidas de la influencia de un conjunto de observaciones sobre los resultados del análisis de regresión lineal multivariado se basa en el concepto de la curva de influencia o función de influencia presentado por HAMPEL(1968,1974).

3.1 Definición de la curva de influencia:

Para estimar algún parámetro de interés, supongamos que tenemos una estadística T construída en base a una muestra aleatoria suficientemente grande, la que proviene de una población con función de distribución F. Si agregamos una o más observaciones adicionales a esta muestra, observamos como cambia T y las conclusiones que se deriven de ella. La curva de influencia se define [4]como:

$$IF(x, T, F) = \lim_{t \to 0} \frac{T((l-t)F + t\triangle x)) - T(F)}{t}, \text{ si el límite existe.}$$

donde:

$$\triangle x = \begin{cases} 1 & \text{en} \quad x \\ \\ 0 & \text{en otro caso.} \end{cases}$$

Esta función es de gran utilidad para estudiar las propiedades asintóticas de un estimador, comparar estimadores, entre otros aspectos.

Las medidas basadas en la curva de influencia [7] evalúan el efecto que produce una o un conjunto de observaciones en la estimación del vector de parámetros.

Para muestras finitas la función de influencia tiene varias aproximaciones como son: la curva de influencia empírica basada en n observaciones, la curva de influencia muestral, la curva de sensibilidad entre otras [5].

3.2 Medidas Basadas en la curva de Influencia

Una expresión para la curva de influencia es:

$$Sup_{X_I^T,Y} \frac{[IF^T(X_I^T, Y, F, \hat{\beta}(F))]M[IF(X_I^T, Y, F, \hat{\beta}(F))]}{c}$$

para valores apropiados de M y c.

Nuestro interés es encontrar la máxima influencia posible que pueda ejercer un conjunto de observaciones sobre el vector de parámetros del modelo de regresión, más aún estamos interesados en ordenar estos conjuntos de observaciones de acuerdo a su nivel de influencia, por lo que resulta conveniente utilizar:

$$D_I(M, c) = \frac{[IF^T(X_I^T, Y_I, F, \hat{\beta}(F))]M[IF(X_I^T, Y_I, F, \hat{\beta}(F))]}{c}$$

Valores grandes de esta distancia, indica que el I-ésimo conjunto ejerce una fuerte influencia sobre el vector de parámetros estimados.

COOK Y WEISBERG (1982), y CHATTERJEE Y HADI (1988), analizaron esta distancia para diferentes expresiones de M y c. BARRET y LING [1], sugirieron la siguiente expresión para la distancia D_I :

$$D_I(M,V) = \frac{Vec(B - B_{(I)})]^T [V^{-1} \otimes M][Vec(B - B_{(I)})]}{p}$$

con: $M = X^T X$ y $V = S$ matrices definidas positivas.

En el presente trabajo consideramos los siguientes casos particulares de la distancia $D_I(M,V)$: Distancia de Cook, Medida DFFITS y la distancia de Welsch:

Tabla 1. Medidas Multivariadas que se basan en la Curva de Influencia

Medida	Matrices	$D_I(M,V)$
Distancia de Cook	$M = X^T X$ $V = pS$	$D_I(M,V) = \left(\frac{n-p}{p}\right) tr\left[(I - H_I)^{-2} H_I Q_I\right]$ $D_I(M,V) = J_I^{tr}\left(\frac{n-p}{p}; 0, 2\right)$
DFFITS	$M = X^T X$ $V = pS_{(I)}$	$D_I(M,V) = \left(\frac{n-p-m}{p}\right) tr\left[(I - H_I)^{-1} Q_I (I - H_I - Q_I)^{-1}\right]$ $D_I(M,V) = J_I^{tr}\left(\frac{n-p-m}{p}; -1, -1\right)$
Distancia de Welsch	$M = X^T X - X_I^T (I - H_I)^{-1} X_I$ $V = \frac{m^2}{n-m} S_{(I)}$	$D_I(M,V) = \left(\frac{(n-m)(n-p-m)}{m^2}\right) tr\left[(I - H_I)^{-2} Q_I (I - H_I - Q_I)^{-1}\right]$ $D_I(M,V) = w_I = J_I^{tr}\left(\frac{(n-m)(n-p-m)}{m^2}; -1, -2\right)$

donde:

$X_{n \times p}$ matriz de valores fijos de las "k" variables regresoras del modelo.

$X_{(I)}$ matriz de orden $(n-m) \times p$ de valores fijos de las "k" variables regresoras.

S matriz de covarianzas de los residuos calculada en base a "n"

observaciones.

$S_{(I)}$ matriz de covarianzas de los residuos calculada en base a "$n-m$" observaciones.

H_I matriz de orden $m \times m$, cuyos elementos de la diagonal se les denomina leverage.

Q_I matriz de orden $m \times m$, es una función de los residuos estudentizados [11].

n número de observaciones muestrales.

p número de parámetros en el modelo de regresión.

m número de observaciones retiradas del análisis.

Tabla 2.

Medida	Influye sobre
Distancia de Cook	Los parámetros estimados del modelo de regresión lineal multivariado.
DFFITS	El vector ajustado de las variables de respuesta.
Distancia de Welsch	Las estimaciones del vector de parámetros y las estimaciones de la matriz de covarianza de los residuos.

3.3 Descomposición de las Medidas de Influencia

La caracterización de las medidas de influencia que se basan en la curva de influencia en la clase J_I^{tr}, facilita la descomposición de la influencia total en dos componentes, leverage y residual [1]:

$$J_I^{tr}(f; a, b) = f(.)tr(L_I * R_I)$$

donde:

L_I es la matriz leverage, una función de H_I.

R_I se le denomina matriz de residuos.

Esta separación de las componentes permite evaluar con mayor objetividad el comportamiento de los conjuntos de observaciones, que pueden estar ejerciendo una fuerte influencia sobre los resultados del modelo de regresión lineal multivariado.

Utilizando los resultados de la Tabla 1, se obtiene:

Tabla 3. Componentes Leverage y Residual

Medida	Componente Leverage	Componente Residual
Distancia de Cook	$L_I = H_I(I-H_I)^{-1}$	$R_I = (I-H_I)^{-1/2}Q_I(I-H_I)^{-1/2}$
DFFITS	$L_I = H_I(I-H_I)^{-1}$	$R_I = (I-H_I)^{-1/2}Q_I(I-H_I-Q_I)^{-1}(I-H_I)^{1/2}$
Distancia de Welsch	$L_I = H_I(I-H_I)^{-2}$	$R_I = (I-H_I)^{-1/2}Q_I(I-H_I-Q_I)^{-1}(I-H_I)^{1/2}$

3.4 Gráficos para las componentes

Los gráficos facilitan el análisis de la contribución de las componentes leverage y residual a la influencia total, para cada una de las medidas presentadas.

Las contribuciones relativas de las componentes reescaladas a la influencia total está dada por:

$$tr(L_I * R_I) = \mathcal{L}_I^* * \mathcal{R}^*$$

donde:

$$\mathcal{L}_I^* = \mathcal{L}_I(\cos\theta_I)^{1/2} \quad \text{con} \quad \mathcal{L}_I = \|Vec(L_I)\| \quad \text{y}$$
$$\mathcal{R}_I^* = \mathcal{R}_I(\cos\theta_I)^{1/2} \quad \text{con} \quad \mathcal{R}_I = \|Vec(R_I)\| \tag{3}$$

varios gráficos se pueden trazar para analizar la contribución de las componentes, pero preferimos utilizar los logaritmos de las contribuciones relativas, por facilitar la ubicación de los conjuntos de observaciones que más contribuyen a las componentes.

En este caso la contribución a la influencia total se define como:

$$\textbf{Influencia Total} = Ln(\mathcal{L}_I^*) + Ln(\mathcal{R}_I^*)$$

4. APLICACIONES

4.1 Modelo propuesto para la aplicación

Para estimar el modelo:

$$\hat{Y}_{1i} = \hat{\beta}_{01} + \hat{\beta}_{11}X_{1i} + \hat{\beta}_{21}X_{2i}$$
$$\hat{Y}_{2i} = \hat{\beta}_{02} + \hat{\beta}_{12}X_{1i} + \hat{\beta}_{22}X_{2i}$$

Se utilizaron los datos de las siguientes variables:

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

Estos datos se obtuvieron en base a una muestra de hogares en cada una de las 24 principales ciudades del Perú: 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. La Encuesta de Seguimiento del Consumo de Hogares (ENSECO 91) se realizó en Junio de 1991.

El modelo estimado es :

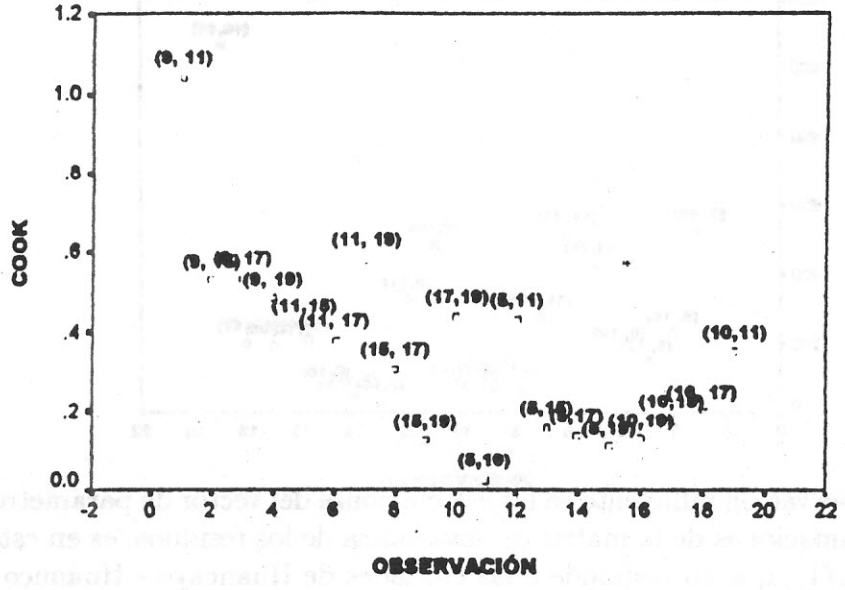
$$\hat{Y}_{1i} = -122,79 + 179,66X_{1i} - 16,95X_{2i}$$
$$\hat{Y}_{2i} = -85,29 + 53,13X_{1i} + 1,22X_{2i}$$

Este modelo permite estimar los gastos: en alimentos y bebidas, así como en transportes y comunicaciones en base al promedio de preceptores y promedio de cuartos por hogar; además de la matriz de residuos y la matriz de covarianzas de los residuos que se requieren para obtener las medidas de influencia multivariadas, y las componentes leverage y residual.

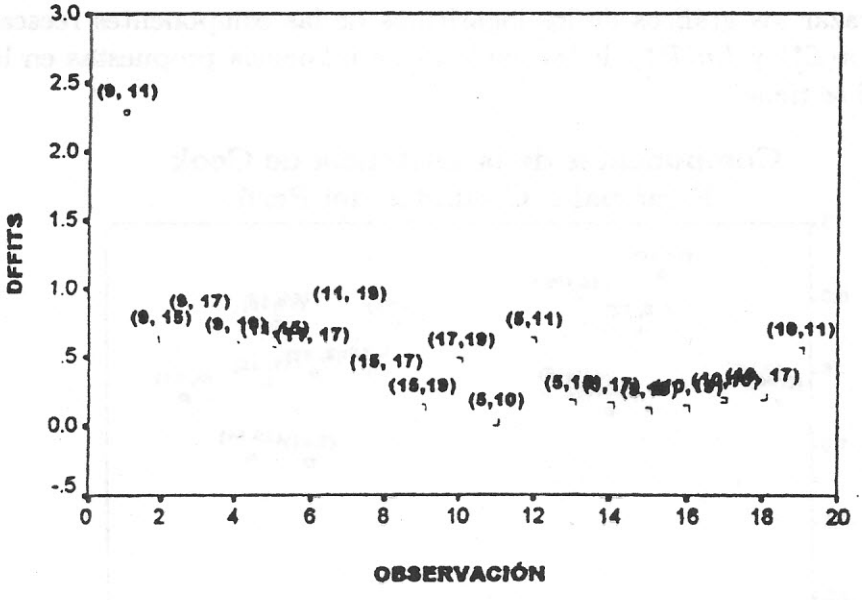
4.2 Medidas de Influencia Multivariadas

Los siguientes gráficos nos muestran los resultados obtenidos al realizar los cálculos [8] para las medidas Distancia de Cook y Medida DFFITS :

Distancia de Cook para conjuntos de tamaño 2
Principales Ciudades del Perú

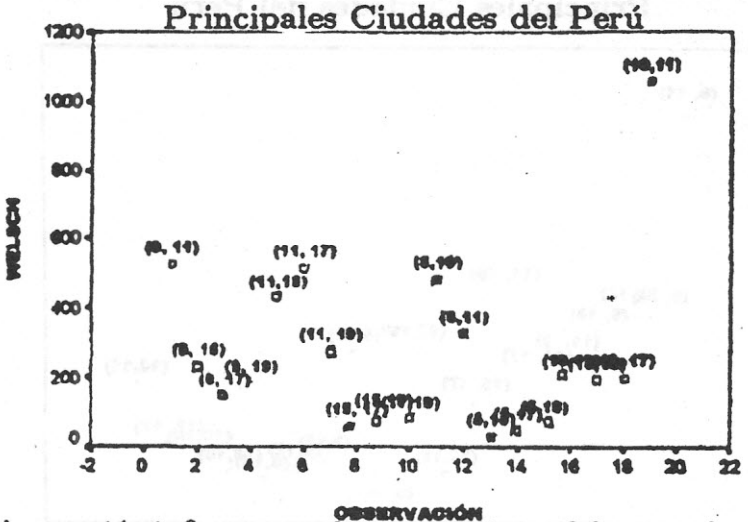


Medida DFFITS para conjuntos de tamaño 2.
Principales Ciudades del Perú



Ambos gráficos, muestran como observación influyente en la estimación de los parámetros del modelo y en el vector ajustado de las variables de respuestas al par (9,11) que corresponde a las ciudades de Huancavelica - Huánuco. Los cálculos de las medidas se obtuvieron en base a las expresiones dadas en la Tabla 1.

Medidas de Welsch para conjuntos de tamaño 2
Principales Ciudades del Perú

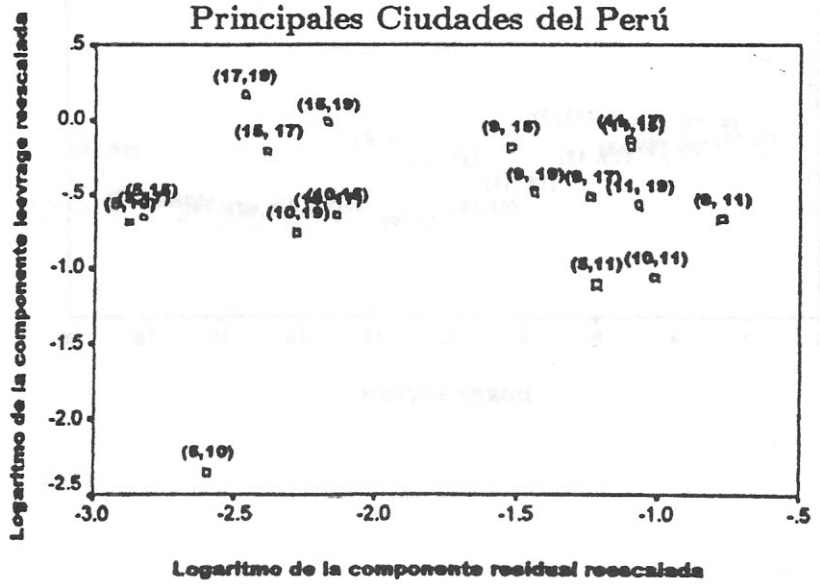


La observación influyente en las estimaciones del vector de parámetros y las estimaciones de la matriz de covarianza de los residuos, es en este caso (10,11), que corresponde a las ciudades de Huancayo - Huánuco.

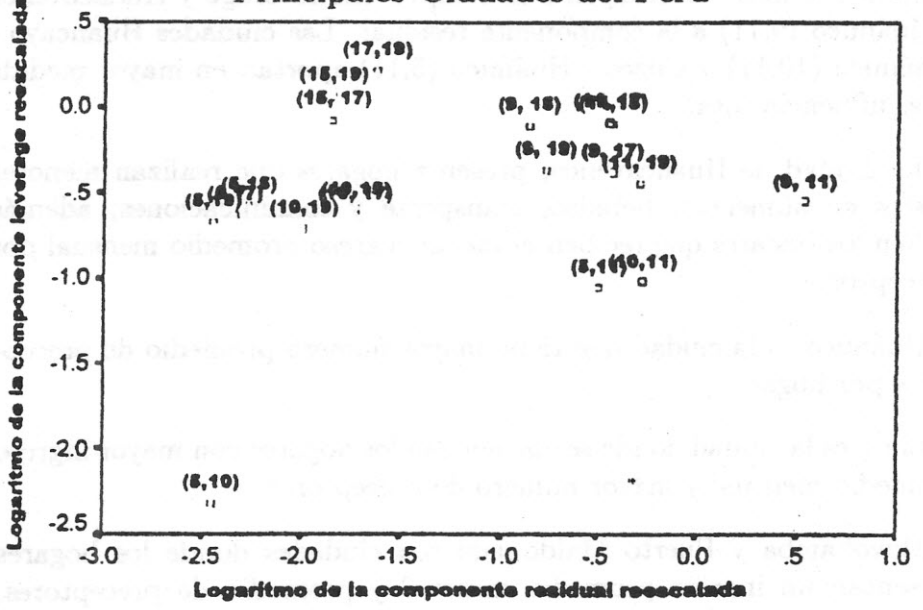
4.3 Descomposición de las Medidas de Influencia

Al trazar los gráficos de los logaritmos de las componentes reescaladas $Ln(\mathcal{L}_I^*)$ y $Ln(\mathcal{R}_I^*)$ de las medidas de influencia propuestas en la Tabla 3 se tiene:

Componentes de la Distancia de Cook
Principales Ciudades del Perú

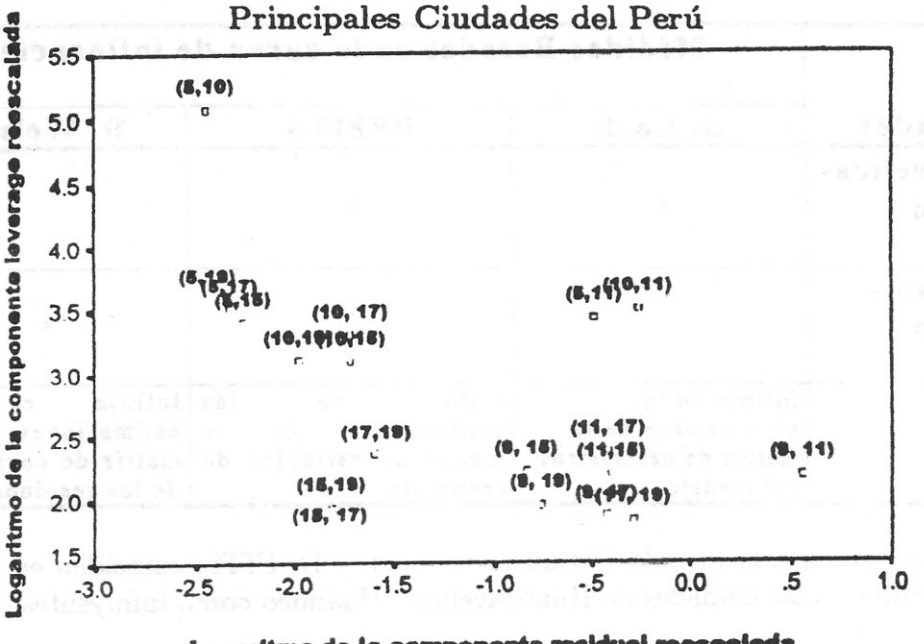


Componentes de la Medida DFFITS
Principales Ciudades del Perú



En las medidas distancia de Cook y DFFITS las ciudades con mayor contribución a la componente residual son Huancavelica - Huánuco (9,11); a la componente leverage: Moyobamba - Puerto Maldonado (17,19), Lima - Puerto maldonado (15,19) y Lima - Moyobamba (11,17); a la influencia total las ciudades de Huánuco - Moyobamba (11,17) y Huánuco - Lima (11,15).

Componentes de Distancia de Welsch
Principales Ciudades del Perú



En este gráfico se observa que las ciudades Cuzco - Huancayo (5,10) son las que más contribuyen a la componente leverage y Huancavelica - Huánuco (9,11) a la componente residual. Las ciudades Huancayo - Huánuco (10,11) y Cuzco - Huánuco (5,11) aportan en mayor medida a la influencia total.

La ciudad de Huancavelica, presenta hogares que realizan menores gastos en alimentos, bebidas, transporte y comunicaciones; además tienen los hogares que reciben el menor ingreso promedio mensual por preceptor.

Huánuco es la ciudad que tiene mayor número promedio de preceptores por hogar.

Lima es la ciudad donde se encuentran los hogares con mayor ingreso promedio mensual y mayor número de preceptores.

Moyobamba y Puerto Maldonado son ciudades donde los hogares presentan un ingreso promedio mensual y promedio de preceptores, relativamente altos.

5. CONCLUSIONES

La aplicación nos muestra:

1° En relación a las medidas de influencia:

Tabla 5.1: Medidas de Influencia Multivariada de la clase J_I^{tr}

	Medidas Basadas en la curva de influencia		
Ciudades	**D. Cook**	**DFFITS**	**D. Welsch**
Huancavelica-Huánuco	*	*	
Huancayo – Huánuco			*
	Influye en las estimaciones de la matriz de parámetros del modelo.	Influye en las estimaciones de la matriz de variables de respuesta.	Influye en las estimaciones de la matriz de covarianza de los residuos.

Se observa que la medida Distancia de Cook y DFFITS, coinciden en identificar a las ciudades de Huancavelica - Huánuco como influyentes.

2° En relación a las componentes residual y leverage:

Tabla 5.2: **Logaritmos de las Componentes reescaladas de las medidas de Influencia**

Ciudades	Medidas Basadas en la curva de influencia								
	D. Cook			DFFITS			D. Welsch		
	L.	R.	I.T.	L.	R.	I.T.	L.	R.	I.T.
Moyobamba-P. Maldonado	*			*					
Lima- P. Maldonado	*			*					
Lima-Moyobamba	*			*					
Huancavelica-Huánuco		*			*			*	
Huánuco-Moyobamba			*			*			
Huánuco-Lima			*			*			
Cuzco-Huancayo							*		
Huancayo-Huánuco									*
Cuzco-Huánuco									*

a) Las componentes de las medidas distancia de Cook y DFFITS, proporcionan en este caso los mismos resultados, es decir identifican como conjuntos de observaciones high-leverage (mayor contribución a la componente leverage), outliers (mayor contribución a la componente residual) así como las que más aportan a la influencia total, a las mismas ciudades.

b) Las ciudades que proporcionan mayor contribución a la influencia total, no necesariamente son las que más contribuyen a la componente leverage y/o residual.

3° La descomposición de las medidas: distancia de Cook y Welsch, DFFITS, en sus componentes leverage y residual (Tabla 5.2); muestran

mayor información que sus medidas de influencia multivariadas (Tabla 5.1).

BIBLIOGRAFÍA

[1] BARRETT BRUCE E. AND LING ROBERT F. *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).

[2] COOK R. DENNIS. *Influential Observations in Linear Regression.*, Journal of the American Statistical Association. Vol. 74, N. 365, (March 1979)pp. 169 - 174.

[3] COOK DENNIS AND WEISBERG, SANFORD. *Residuals and Influence in Regression.* Chapman and Hall. London. (1982).

[4] CHATTERJEE, SAMPRIT and HADI, ALI. *Influential Observation, High Leverage Points and Outliers in Linear Regression.* Statistical Science. Vol.1 N 3,(1986)pp. 379 - 416.

[5] CHATTERJEE, SAMPRIT and HADI, ALI. *Sensitivity Analysis in Linear Regression.* John Wiley. New York. (1988).

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

[7] PONCE ARUNERI, M. ESTELA. *Medidas de Influencia en Regresión Multivariada. Tesis para Maestría en Estadística.* UNMSM. FCM. Lima - Perú.(1999).

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