Revista Colombiana de Estadística
Volumen 27 N° 2. Págs. 109 a 121. Diciembre 2004

Un criterio para identificar datos atípicos

José Alfredo Jiménez Moscoso*

Resumen

En este artículo se presenta un método para determinar las observaciones que son atípicas en un modelo de regresión lineal múltiple; estos datos se estableceran de acuerdo al cambio que ejercen sobre la suma de los cuadrados de residuales del modelo.

Palabras Claves: *Modelos lineales, mínimos cuadrados, formas cuadráticas, observaciones atípicas, estadística* Q_k.

Abstract

This paper present a method to determine the observations that are outliers in a model of multiple linear regression; these data will be established according to the change that is presented on the sum of the squares of residual of the model.

Key words: *Linear models, Least squares, Quadratic forms, Outliers,* Q_k *Statistics.*

1. Introducción

Draper & John (1981) proponen una metodología para detectar un grupo de k observaciones atípicas, análoga a la propuesta de Bartlett (1937), citada en Little & Rubin (1987), para estimar los parámetros del modelo de regresión lineal cuando existen observaciones faltantes en la variable respuesta. En el

*Profesor asistente, Universidad Nacional de Colombia, Departamento de Matemáticas. E-mail: josajimenezm@unal.edu.co

planteamiento de Draper & John (1981) se considera el modelo de regresión lineal múltiple:

$$\underset{n\times 1}{Y} = \underset{n\times r}{X}\ \underset{r\times 1}{\beta} + \underset{n\times 1}{\epsilon}, \tag{1}$$

particionado de la siguiente manera:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & I \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \tag{2}$$

donde Y_1 es el bloque conformado por las observaciones consideradas atípicas. Para el modelo (2) establecen las estimaciones de β y γ mediante:

$$\hat{\beta} = (X_2'X_2)^{-1}X_2'Y_2\ , \qquad \hat{\gamma} = (I - H_{11})^{-1}\hat{\epsilon}_1,$$

donde $H_{ij} = X_i(X'X)^{-1}X_j'$ es una submatriz de la matriz

$$H = X(X'X)^{-1}X'\ , \quad \text{para } X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

La notación de H y el nombre de *matriz hat* fue introducido por Tukey (1977); por otra parte, el cambio en la suma de cuadrados de residuales lo calculan usando la estadística:

$$Q_k = \hat{\epsilon}_1'(I - H_{11})^{-1}\hat{\epsilon}_1, \qquad \text{con } k = \dim(Y_1). \tag{3}$$

En resumen, el método descrito permite detectar el grupo de observaciones atípicas en base al cambio en la suma de cuadrados de residuales, lo cual se cuantifica con la estadística Q_k, es decir, mediante este procedimiento se selecciona el bloque Y_1 que posee el Q_k más alto, como el bloque más atípico, y en muchos casos quedan datos atípicos dentro de un bloque y el método no los identifica. En este artículo se muestra un criterio para identificar el bloque Y_1 que contiene el grupo más grande de observaciones atípicas.

2. Resultados básicos del ajuste del modelo de regresión lineal múltiple

Mediante el método de estimación *mínimos cuadrados ordinarios* (MCO) se obtiene para el modelo dado en (1) los siguientes estimadores:

$$\begin{aligned} \widehat{\beta} &= (X^{'}X)^{-1}X'Y, \\ \widehat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'Y = HY, \\ \widehat{\epsilon} &= Y - \hat{Y} = Y - HY = (I-H)Y, \\ SCE &= \widehat{\epsilon}\,'\widehat{\epsilon} = [(I-H)Y]'\,(I-H)Y = Y'(I-H)Y. \end{aligned} \tag{4}$$

Obsérvese que la matriz H determina muchos de los resultados de las estimaciones por MCO; por ejemplo, cuando premultiplica al vector de respuestas Y se obtienen los valores predichos de la variable dependiente, por eso en la literatura estadística en algunos casos la denominan *matriz de predicción*, y a la matriz $I - H$ la llaman *matriz residual*, puesto que al anteponérsele a la variable dependiente Y se obtienen los respectivos residuales.

2.1. Propiedades de las componentes de la matriz H

En Hoaglin & Welsch (1978) se establece para la matriz $H = [h_{ij}]$ de tamaño $n \times n$, las siguientes propiedades:

(a)	$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$	ya que H *es simétrica e idempotente.*
(b)	$0 < h_{ii} \leq 1,$	si $i = 1, 2, \ldots, n.$
(c)	$-0{,}5 \leq h_{ij} \leq 0{,}5,$	para $i \neq j.$
(d)	$(1 - h_{ii})(1 - h_{jj}) - h_{ij}^2 \geq 0.$	
(e)	$h_{ii}h_{jj} - h_{ij}^2 \geq 0.$	
(f)	Si $h_{ii} = 1,$ entonces $h_{ij} = 0,$	para todo $j \neq i.$

Si la matriz X de tamaño $n \times r$ es de rango r, entonces

$$(g) \quad \sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^2 = r = \mathrm{tr}(H),$$

$$(h) \quad \sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1,$$

donde $\mathrm{tr}(H)$ denota la traza de la matriz H.

Dado que $h_{ij} = x_i(X'X)^{-1}x_j'$, entonces h_{ii} está determinado por la locali-

zación de x_i en el espacio X, es decir, un valor pequeño (grande) de h_{ii} indica que x_i se encuentra cerca (lejos) de la masa de los otros puntos. Además, sugieren que x_i es un punto influyente si $h_{ii} > 2r/n$.

3. Cálculo de la estadística $\mathbf{Q_k}$

En Jiménez (2001*b*) se establece para la estadística dada en (3), la siguiente expresión:

$$Q_k = SCE - SCE^* = -2\gamma'\hat{\epsilon} - \gamma'(I-H)\gamma \ , \tag{5}$$

donde SCE es obtenida en términos algebraicos como en (4) y SCE^*, representa la estimación vía mínimos cuadrados (EMC) de SCE sin el bloque Y_1 de observaciones. Además, muestra que si el interés es minimizar la SCE^*, esto se logra haciendo:

$$\frac{\partial Q_k}{\partial \gamma} = 0,$$

lo cual equivalente a hacer:

$$\widehat{\epsilon} + (I-H)\widehat{\gamma} = 0, \tag{6}$$

donde $\widehat{\epsilon}$ es la estimación vía mínimos cuadrados (EMC) de ϵ del modelo (1).

Al remplazar (6) en (5) se tiene:

$$Q_k = \widehat{\gamma}'(I-H)\widehat{\gamma} = \widehat{\gamma}'\widehat{\gamma} - \widehat{\gamma}'H\widehat{\gamma}. \tag{7}$$

Esta nueva expresión de Q_k tiene la ventaja de que está en términos de la estimación del γ arbitrario, la cual para los objetivos de este trabajo es más atractiva, ya que se podrá establecer su distribución de probabilidad correspondiente.

4. Distribución de probabilidad de $\mathbf{Q_k}$

En Jiménez (2001*a*) al asumir la restricción $\hat{\gamma} = \begin{bmatrix} \hat{\gamma}_1 \\ 0 \end{bmatrix}$, se llega a:

$$\widehat{\gamma} = \begin{bmatrix} \widehat{\gamma}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} -I_k & X_1(X_2^{'}X_2)^{-1}X_2^{'} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \tag{8}$$

donde I_k es la matriz identidad de tamaño $k \times k$, con k igual a la dimensión del bloque Y_1 y $M_{ij} = X_i(X_2'X_2)^{-1}X_j'$.

Si se reemplaza (8) en el primer término de la expresión (7) se obtiene

$$\begin{aligned}\widehat{\gamma}^{'}\widehat{\gamma} &= Y^{'}\begin{bmatrix} -I_k & 0 \\ X_2(X_2^{'}X_2)^{-1}X_1^{'} & 0 \end{bmatrix}\begin{bmatrix} -I_k & X_1(X_2^{'}X_2)^{-1}X_2^{'} \\ 0 & 0 \end{bmatrix} Y \\ &= Y^{'}\begin{bmatrix} I_k & -M_{12} \\ -M_{21} & M_{21}M_{12} \end{bmatrix} Y. \end{aligned} \tag{9}$$

Por otra parte, si se sustituye (8) en el segundo término de la expresión (7) y se emplean los resultados dados en Jiménez (2001*a*), se tiene que:

$$\begin{aligned}\widehat{\gamma}^{'}H\widehat{\gamma} =& Y^{'}\begin{bmatrix} H_{11} & H_{12}-M_{12} \\ H_{21}-M_{21} & H_{22}+M_{21}M_{12}-M_{22} \end{bmatrix} Y \\ =& Y^{'}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} Y - Y^{'}\begin{bmatrix} 0 & M_{12} \\ M_{21} & M_{22}-M_{21}M_{12} \end{bmatrix} Y. \end{aligned} \tag{10}$$

Finalmente, al sustituir (9) y (10) en la ecuación (7), se obtiene que:

$$\begin{aligned} Q_k =& \widehat{\gamma}^{'}(I-H)\widehat{\gamma} \\ =& Y^{'}\begin{bmatrix} I_k & -M_{12} \\ -M_{21} & M_{21}M_{12} \end{bmatrix} Y - Y^{'}\begin{bmatrix} H_{11} & H_{12}-M_{12} \\ H_{21}-M_{21} & H_{22}+M_{21}M_{12}-M_{22} \end{bmatrix} Y \\ =& Y^{'}\begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix} Y - Y^{'}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} Y \\ =& Y^{'}MY - Y^{'}HY = Y^{'}(M-H)Y. \end{aligned} \tag{11}$$

Nótese que la matriz $(M-H)$ es simétrica; además, es idempotente. Esto se puede verificar de la siguiente manera:

$$(M-H)(M-H) = M^2 - MH - HM + H^2 \ ,$$

pero $M^2 = M$, ya que:

$$\begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix}\begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & M_{22}M_{22} \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix}.$$

Esto se tiene, ya que para $i, j = 1, 2$:

$$M_{i2}M_{2j} = [X_i(X_2^{'}X_2)^{-1}X_2^{'}][X_2(X_2^{'}X_2)^{-1}X_j^{'}] = X_i(X_2^{'}X_2)^{-1}X_j^{'} = M_{ij} \ ;$$

por otra parte, $HM = H$ lo cual se puede verificar como sigue:

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}\begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12}M_{22} \\ H_{21} & H_{22}M_{22} \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}.$$

Aquí cabe notar que cuando $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ es de rango completo, entonces:

$$H_{i2}M_{2j} = [X_i(X^{'}X)^{-1}X_2^{'}][X_2(X_2^{'}X_2)^{-1}X_j^{'}] = X_i(X^{'}X)^{-1}X_j^{'} = H_{ij},$$

para $i, j = 1, 2$; además, como las matrices H y M son simétricas se tiene que $H = (MH)^t = HM$. En consecuencia,

$$(M - H)(M - H) = M - H.$$

Para establecer la distribución de Q_k, se presentan, sin demostración, los teoremas 1 y 2, mencionados en Searle (1971).

Teorema 1. Si Y es un vector aleatorio de tamaño $n \times 1$, distribuido $N(\mu, V)$, donde μ es en si mismo un vector entonces:

$$\mathrm{E}[Y'AY] = \mathrm{tr}(AV) + \mu' A\mu \quad \text{y} \quad \mathrm{Var}[Y'AY] = 2\ \mathrm{tr}(AV)^2 + 4\mu' AVA\mu.$$

Teorema 2. Si $Y \sim N(\mu, V)$, entonces $Y'AY \sim \chi^{2^{'}}_{(\nu,\lambda)}$, con grados de libertad $\nu = \rho(A)$ y parámetro de no centralidad $\lambda = \frac{1}{2}\mu' A\mu$, si y sólo si AV es idempotente.

Puesto que, bajo el supuesto de normalidad en los residuales se tiene que

$$Y \sim N(X\beta, \sigma^2 I_n). \tag{12}$$

Como la expresión dada en (11) es una forma cuadrática se establecerá a continuación la respectiva distribución asociada. Por el teorema 1, se tiene que

$$\mathrm{E}\left[\frac{Y^{'}(M-H)Y}{\sigma^2}\right] = \left\{k - r + tr\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]\right\},$$

$$\mathrm{Var}\left[\frac{Y^{'}(M-H)Y}{\sigma^2}\right] = 2\left\{k - r + tr\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]\right\},$$

donde r es el rango de la matriz X definida en el modelo (1). Cuando esta matriz es de rango completo se tiene que $\mathrm{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right] = r$.

Utilizando el teorema 2, también se concluye que Q_k/σ^2 tiene distribución ji-cuadrado central:

$$\frac{Q_k}{\sigma^2} \sim \chi^2_{(\nu)}, \tag{13}$$

donde $\nu = k - r + \ \mathrm{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]$. Aquí el teorema 2 es aplicable ya que $\dfrac{1}{\sigma^2}(M - H)\sigma^2 I_n$ es una matriz idempotente.

5. Metodología para establecer datos atípicos

Dado que la estadística Q_k se puede obtener de la forma cuadrática:

$$Q_k = \widehat{\gamma}^{'}(I - H)\widehat{\gamma}, \tag{14}$$

al expresarla en términos del vector de respuestas Y, queda como:

$$Q_k = Y^{'}\begin{bmatrix} I_k & 0 \\ 0 & M_{22} \end{bmatrix} Y - Y^{'}\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} Y. \tag{15}$$

Si se considera que en la partición $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$, el bloque Y_1 está conformado por las observaciones atípicas, dicho bloque afectará todas las EMC del modelo dado en (1). Por otra parte, si se reescribe la expresión (5), se tiene que:

$$SCE = SCE^* + Q_k,$$

y dado que SCE^* puede expresarse en forma matricial como sigue

$$SCE^* = Y^{'}\begin{bmatrix} 0 & 0 \\ 0 & I_{n-k} - M_{22} \end{bmatrix} Y = Y^{'}\left[I_n - M\right] Y; \tag{16}$$

usando (12), se puede establecer que las expresiones,

$$\frac{SCE}{\sigma^2} \qquad \text{y} \qquad \frac{SCE^*}{\sigma^2}\ , \tag{17}$$

tienen distribución ji-cuadrado central. Luego, si se divide la ecuación (13) por cualquiera de las expresiones dadas en (17), se elimina el término σ^2 y queda el cociente entre dos formas cuadráticas que se distribuyen ji-cuadrado.

Por la teoría estadística se sabe que cuando se realiza el cociente entre dos variables aleatorias independientes con distribución ji-cuadrado y cada una se divide por sus respectivos grados de libertad, se obtiene una nueva variable con distribución F.

Para llevar a cabo el cociente mencionado anteriormente se debe verificar con cuál de las distribuciones asociadas a las expresiones dadas en (17) la distribución de probabilidad expresada en (13) es independiente; para ello, se enuncia sin demostración el teorema 3, citado en Searle (1971).

Teorema 3. Cuando $Y \sim N(\mu, V)$, las formas cuadráticas $Y'AY$ y $Y'BY$, están distribuidas independientemente si y sólo si $AVB = 0$.

Veamos si las distribuciones asociadas a Q_k y SCE son independientes. Si se retoman las ecuaciones dadas en (11) y (4), se tiene por el teorema 3 que Q_k y SCE no son independientes, pues,

$$\begin{aligned}(M-H)(\sigma^2 I_n)(I_n-H) &= \sigma^2(M-H)(I_n-H)\\ &= \sigma^2[M-MH-H+H^2] = \sigma^2(M-H)\neq 0;\end{aligned}$$

en la última ecuación se tuvo en cuenta que H es idempotente y que $MH=H$.

De manera análoga, se verifica si son independientes las distribuciones de probabilidad de Q_k y SCE^*; de las ecuaciones (11) y (16) utilizando el teorema 3, se concluye que son independientes, ya que:

$$\begin{aligned}(M-H)(\sigma^2 I_n)\,(I_n-M) &= \sigma^2(M-H)\,(I_n-M)\\ &= \sigma^2\left[M-M^2-H+HM\right]=0.\end{aligned}$$

En esta última expresión se utilizaron los resultados: $MH=H$ y $M^2=M$.

La media y varianza de la SCE^* se obtienen por el teorema 1, como sigue:

$$\begin{aligned}E\left[\frac{Y^{'}(I_n-M)\,Y}{\sigma^2}\right] &= \left\{n-k-\operatorname{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]\right\},\\ Var\left[\frac{Y^{'}(I_n-M)\,Y}{\sigma^2}\right] &= 2\left\{n-k-\operatorname{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]\right\}.\end{aligned}$$

Como la media y la varianza de la distribución χ^2_η son η y 2η respectivamente, se deduce que $\left[Y^{'}(I_n-M)\,Y\right]/\sigma^2$ tiene distribución ji-cuadrado central. Se llega a la misma conclusión, ya que $\dfrac{1}{\sigma^2}(I_n-M)\,\sigma^2 I_n$ es idempotente, utilizando el teorema 2. Así pues,

$$\frac{Y^{'}(I_n-M)\,Y}{\sigma^2}\sim\chi^2_\eta, \tag{18}$$

con $\eta=n-k-\operatorname{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]$. Cuando la matriz X es de rango completo se tiene que $\operatorname{tr}\left[(X_2^{'}X_2)^{-1}(X_2^{'}X_2)\right]=r$.

Como las distribuciones de probabilidad asociadas a las expresiones (15) y (16) son independientes, al hacer el cociente entre las relaciones (13) y (18),

dividiendo cada una por sus correspondientes grados de libertad, se llega a:

$$\frac{\frac{Q_k}{k\sigma^2}}{\frac{SCE^*}{(n-r-k)\sigma^2}} = \frac{n-r-k}{k}\frac{\widehat{\gamma}(I-H)\widehat{\gamma}}{SCE^*},$$

$$\left(\frac{n-r-k}{k}\right)\frac{Q_k}{SCE^*} \sim F_{(k,n-r-k)}.$$

Estos resultados se pueden resumir en los siguientes teoremas.

Teorema 4. Si en un modelo de regresión lineal múltiple particionado como:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix},$$

se elimina el bloque Y_1 de dimensión k, entonces el cambio que se presenta en la SCE se calcula mediante la expresión:

$$\Delta_{(Y_1)} = \frac{1}{k}\ \frac{\widehat{\gamma}'\left[I_n - H\right]\widehat{\gamma}}{S^2_{(Y_1)}}\ , \tag{19}$$

donde $S^2_{(Y_1)} = \widehat{\sigma}^2 = \dfrac{SCE^*}{n-k-r}$ es la estimación usual de σ^2, después de eliminar las observaciones del bloque Y_1, y $\widehat{\gamma} = \begin{bmatrix} \widehat{\gamma}_1 \\ 0 \end{bmatrix}$, con $\widehat{\gamma}_1 = -Y_1 + X_1(X_2'X_2)^{-1}X_2'Y_2$.

Teorema 5. En un modelo de regresión lineal múltiple $Y = X\beta + \epsilon$, bajo el supuesto de que $\epsilon \sim N(0, \sigma^2 I_n)$, se tiene que:

$$\Delta_{(Y_1)} \sim F_{(k,n-r-k)}, \qquad \text{con} \qquad \begin{array}{l} k = \text{dimensión del bloque } Y_1, \\ r = \text{rango de la matriz } X. \end{array}$$

En este caso, se clasifica como atípico al bloque Y_1 de observaciones, si con un nivel de significancia α se satisface que:

$$\Delta_{(Y_1)} > F_{(k,n-r-k,\alpha/2)}. \tag{20}$$

6. Ejemplo

En la Tabla 1, se considera el conjunto de 21 observaciones (x, y), dado por Mickey, Dunn & Clark (1967).

Para este conjunto de datos, se presentan los siguientes resultados:

Tabla 1: Datos de Mickey, Dunn, and Clark (1967)

Obs.	**x**	**y**	**Obs.**	**x**	**y**	**Obs.**	**x**	**y**
1	15	95	8	11	100	15	11	102
2	26	71	9	8	104	16	10	100
3	10	83	10	20	94	17	12	105
4	9	91	11	7	113	18	42	57
5	15	102	12	9	96	19	17	121
6	20	87	13	10	83	20	11	86
7	18	93	14	11	84	21	10	100

1. La estimación del modelo de regresión lineal, con las 21 observaciones.

2. Los elementos de la diagonal de la matriz H, las estimaciones de los γ_i y al eliminar el i-ésimo dato se establecen la estadística Q_1, la distancia de Cook y la estadística $\Delta_{(i)}$ con su p-valor correspondiente.

3. La estimación del modelo de regresión lineal, después de eliminar la observación influyente determinada mediante distancia de Cook.

4. La estimación del modelo de regresión lineal, sin la observación que se considera influyente por la estadística $\Delta_{(i)}$.

1. Análisis de varianza para el conjunto completo de datos:

Fuente de variación	**Grados libertad**	**Suma de cuadrados**	**Cuadrados Medios**	***F***	**Valor crítico de *F***
Regresión	1	1604,0809	1604,0809	13,2018	0,00177
Residuos	19	2308,5858	121,5045		
Total	20	3912,6667			

Coeficiente de determinación $R^2 = 0{,}409971261$:

	Coeficientes	**Error típico**	**Estadístico *t***
Intercepto	109,8738	5,0678	21,6808
Variable X	-1,1270	0,3102	-3,6334

2. Compendio de estadísticas:

Obs. Elim.	$\mathbf{h_{ii}}$	$\widehat{\gamma}_{\mathbf{i}}$	$\mathbf{Q_k}$ (k=1)	$\mathbf{D_i^*}$ ($100*D_i$)	$\mathbf{\Delta_{(i)}}$	**p−valor**
1	0,0479	-2,1332	4,333	0,09	0,0338	0,8561
2	0,1545	11,3214	108,370	8,15	0,8866	0,3589
3	0,0628	16,6498	259,803	7,17	2,2826	0,1482
4	0,0705	9,3936	82,015	2,56	0,6630	0,4261
5	0,0479	-9,4856	85,664	1,77	0,6937	0,4158
6	0,0726	0,3602	0,120	0,00	0,0009	0,9759
7	0,0580	-3,6220	12,358	0,31	0,0969	0,7592
8	0,0567	-2,6746	6,748	0,17	0,0528	0,8209
9	0,0799	-3,4148	10,729	0,38	0,0840	0,7752
10	0,0726	-7,1879	47,914	1,54	0,3815	0,5445
11	0,0908	-12,1145	133,443	5,48	1,1043	0,3072
12	0,0705	4,0141	14,976	0,47	0,1175	0,7357
13	0,0628	16,6498	259,803	7,17	2,2826	0,1482
14	0,0567	14,2866	192,540	4,76	1,6378	0,2169
15	0,0567	-4,7948	21,687	0,54	0,1707	0,6844
16	0,0628	-1,4896	2,080	0,06	0,0162	0,9000
17	0,0521	-9,1255	78,936	1,79	0,6373	0,4351
18	0,6516	15,9026	88,105	67,81	0,7142	0,4091
19	0,0531	-31,9816	968,562	22,33	13,0103	0,0020
20	0,0567	12,1664	139,634	3,45	1,1588	0,2959
21	0,0628	-1,4896	2,080	0,06	0,0162	0,9000
Valor Inusual	$h_{ii} > 4/21$	$\lvert\gamma_i\rvert \geq \gamma_j$ para todo$j$	*el más grande*	$D_i > 0,5$		$p < \alpha$ (α=5 %)

De los resultados anteriores se tiene que:

a) La observación que se clasifica como influyente, usando la estadística propuesta por Cook, coincide con la que se detecta con el criterio para el elemento h_{ii}.

b) Los otros métodos detectan la misma observación como atípica cuando se elimina una sola observación, pero cuando se eliminan dos o más observaciones el procedimiento más formal es el del p-valor asociado a la estadística $\Delta_{(Y_1)}$.

3. Cuando se elimina la observación 18, se obtiene:

Fuente de variación	**Grados libertad**	**Suma de cuadrados**	**Cuadrados medios**	***F***	**Valor crítico de *F***
Regresión	1	280,5195	280,5195	2,27399	0,1489
Residuos	18	2220,4805	123,3600		
Total	19	2501			

Coeficiente de determinación $R^2 = 0{,}112162$.

Cambio en la suma de los residuales $Q_k = 88{,}10525836$.

	Coeficientes	**Error típico**	**Estadístico *t***
Intercepto	105,62987	7,1619276	14,7488045
Variable X	-0,77922	0,516733	-1,5079754

La distancia de Cook nos indicó que la pareja $(42, 57)$ era la que más afectaba la EMC de los parámetros, pero al eliminarla el modelo obtenido fue más deficiente que el modelo completo. Por lo tanto, la observación es solamente influencial pero no es atípica.

4. Eliminando la observación 19 que detectó $\Delta_{(i)}$ como atípica, se tiene:

Fuente de variación	**Grados libertad**	**Suma de cuadrados**	**Cuadrados medios**	***F***	**Valor crítico de *F***
Regresión	1	1788,17619	1788,17619	24,01985	0,0001151
Residuos	18	1340,02381	74,44577		
Total	19	3128,2			

Coeficiente de determinación $R^2 = 0, 57163103$.

Cambio en la suma de los residuales $Q_k = 968, 5619674$.

	Coeficientes	**Error típico**	**Estadístico *t***
Intercepto	109,30468	3,96996	27,5329
Variable X	-1,19331	0,24348	-4,9010

El modelo que se obtiene al eliminar la pareja $(17, 121)$ es mejor que el modelo completo, pues el nuevo coeficiente de determinación es superior al del modelo inicial. El valor crítico de la F es también inferior al valor crítico que se determinó en el análisis de varianza del modelo inicial y, además, el cuadrado medio del error (CME) fue menor que el CME del modelo completo. Aunque dicha observación es atípica, no es influyente en la estimación de los parámetros del modelo.

7. Conclusiones

La metodología aquí presentada permite detectar en un grupo de observaciones la observación más atípica, es decir, el dato más influyente sobre el cambio en la suma de cuadrados de los residuales. Además, este procedimiento proporciona una manera de cuantificar el impacto de cada observación sobre la suma de cuadrados de los residuales, pues empleando la distribución F-central este método permite asignarle un p-valor a cada influencia; de esta manera se obtiene un criterio más exacto que el usado tradicionalmente.

Bibliografía

Bartlett, M. S. (1937), 'Some examples of statistical methods of research in agriculture and applied botany', *Journal of the Royal Statistical Society* **B4**, 137–170.

Draper, N. R. & John, J. A. (1981), 'Influential observations and outliers in regression', *Technometrics* **23**(1), 21–26.

Hoaglin, D. C. & Welsch, R. E. (1978), 'The hat matrix in regression and anova', *The American Statistician* **32**(1), 17–22.

Jiménez, J. A. (2001*a*), 'Una generalización de la estadística de Cook', *Revista Colombiana de Estadística* **24**(2), 111–120.

Jiménez, J. A. (2001*b*), 'Una maximización de la estadística Q_k', *Revista Colombiana de Estadística* **24**(1), 45–57.

Little, R. J. & Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, John Wiley & Sons.

Mickey, M. R., Dunn, O. J. & Clark, V. (1967), 'Note on the use of stepwise regression in detecting outliers', *Computers and Biomedical Research,* **1**, 105–111.

Searle, S. (1971), *Linear Models*, John Wiley & Sons.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison Wesley.