**Revista Colombiana de Estadística**

Volumen 24 (2001) N° 2, páginas 111 *a* 120

# UNA GENERALIZACIÓN DE LA ESTADÍSTICA DE COOK

JOSÉ A. JIMÉNEZ M.*

**Resumen**

En este artículo se presenta una generalización de la estadística de Cook la cual permite identificar las observaciones más influyentes en la estimación vía mínimos cuadrados de los parámetros del modelo de regresión lineal múltiple.

**Palabras claves:** *Modelos Lineales, Mínimos cuadrados, Observaciones Influyentes, Estadística de Cook, Estadística $Q_k$, Estadística DF-Beta*

**Abstract**

This paper presents a generalization of the Cook's statistics useful in the identification of influential observations in the least squares estimation of the multiple regression parameters.

**Keywords:** *Linear Models, Least squares, Influential Observations, Cook's Statistics, $Q_k$ Statistics, DF-Beta Statistics*

*Profesor asistente, Universidad Nacional de Colombia, Departamento de matemáticas; e-mail: jjimenez@matematicas.unal.edu.co

# 1. Introducción

En Cook (1977) se introduce una estadística para indicar la influencia de una observación con respecto a un modelo particular. Para una única observación, Cook también mostró que la estadística proporcionaba información sobre si era también un outlier. Para el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \tag{1}$$

siendo $\vec{Y}$ un vector de respuestas de tamaño $n \times 1$, $X$ una matriz de constantes conocidas de tamaño $n \times r$ de rango completo, $\vec{\beta}$ el vector de parámetros de tamaño $r \times 1$ y $\vec{\epsilon}$ un vector de errores de tamaño $n \times 1$. Mediante el método de estimación mínimos cuadrados ordinarios (MCO) se obtiene para el modelo dado en (1) los siguientes estimadores:

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'\vec{Y} \\ \widehat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'\vec{Y} = H\vec{Y} \\ \widehat{\epsilon} &= \vec{Y} - \hat{Y} = \vec{Y} - H\vec{Y} = (I-H)\vec{Y} \\ SCE &= \widehat{\epsilon}'\widehat{\epsilon} = \left[(I-H)\vec{Y}\right]'(I-H)\vec{Y} = \vec{Y}'(I-H)\vec{Y} \end{aligned} \tag{2}$$

con $H = X(X'X)^{-1}X'$. Bajo el supuesto de normalidad en los residuales, se establece una región de $(1-\alpha)100\,\%$ de confianza para $\vec{\beta}$, mediante

$$\frac{\left(\vec{\beta} - \widehat{\beta}\right)'(X'X)\left(\vec{\beta} - \widehat{\beta}\right)}{rs^2} \leq F_{(r,n-r,\alpha)} \tag{3}$$

donde $s^2 = SCE/(n-r)$ es el estimador insesgado de $\sigma^2$ y $F_{(r,n-r,\alpha)}$ es el percentil $\alpha$-superior de una distribución $F$ con $r$ y $n-r$ grados de libertad. Esta desigualdad define una región elipsoidal centrada en $\widehat{\beta}$.

En el resultado estadístico propuesto por Cook, la influencia de una observación es medida por el cambio en el centro de la región elipsoidal dada en (3) cuando la $i$-ésima observación es eliminada. Para ello, define la siguiente medida (distancia):

$$D_i = \frac{\left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)'(X'X)\left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)}{rs^2} \tag{4}$$

donde $\widehat{\beta}_{(i)}$ es el estimador vía mínimos cuadrados (EMC) de $\vec{\beta}$ después de eliminar la $i$-ésima observación del modelo (1). En Cook (1980) se sugiere que

cada $D_i$ sea comparada con el percentil de una $F$ con $r$ y $n-r$ grados de libertad; en otras palabras, grandes valores de $D_i$ indican que la observación es influencial.

En este artículo se presenta una generalización de esta estadística que se denotará por $D_{(\vec{Y}_1)}$ la cual permitirá detectar si las observaciones del bloque $\vec{Y}_1$ son influyentes en la estimación de los parámetros del modelo de regresión lineal múltiple, se demostrará que $D_{(\vec{Y}_1)}$ no se distribuye $F_{(r,n-r)}$ y por lo tanto, se utilizará como criterio de decisión, que cuando $D_{(\vec{Y}_1)} > 0{,}5$, las observaciones del bloque $\vec{Y}_1$ sean consideradas influyentes.

## 2. Deducción de la Estadística de Cook

En Jiménez (1999) se plantea el modelo

$$\begin{bmatrix} \vec{Y}_1^* \\ \vec{Y}_2^* \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta}^* + \begin{bmatrix} \vec{\epsilon}_1^* \\ \vec{\epsilon}_2^* \end{bmatrix} \tag{5}$$

siendo $\vec{Y}^* = \vec{Y} + \vec{\gamma}$ con $\vec{\gamma} \in \mathbb{R}^n$ un vector arbitrario; bajo este supuesto por MCO se obtienen los estimadores del modelo (5), en función de $\vec{\gamma}$, $\vec{Y}$ y de las expresiones dadas en (2). Los nuevos estimadores están dados por

$$\begin{aligned} \widehat{\beta}^* &= \widehat{\beta} + (X'X)^{-1}X'\vec{\gamma} \\ \hat{Y}^* &= \hat{Y} + H\vec{\gamma} \\ \hat{\epsilon}^* &= \hat{\epsilon} + (I-H)\vec{\gamma} \\ SCE^* &= SCE + 2\vec{\gamma}\,'\hat{\epsilon} + \vec{\gamma}\,'(I-H)\vec{\gamma} \end{aligned}$$

Considerando la generalización de la estadística $DFBeta(\vec{Y}_1)$ propuesta en Jiménez y Rincón (2000), dada por

$$DFBeta(\vec{Y}_1) = \widehat{\beta} - \widehat{\beta}^*_{\vec{Y}_1} = -(X'X)^{-1}X'\vec{\gamma} \tag{6}$$

se puede expresar la estadística de Cook dada en (4) de la siguiente manera:

$$\begin{aligned} D_{(\vec{Y}_1)} &= \frac{\left(\widehat{\beta} - \widehat{\beta}^*_{\vec{Y}_1}\right)'(X'X)\left(\widehat{\beta} - \widehat{\beta}^*_{\vec{Y}_1}\right)}{rs^2} \\ &= \frac{\left(-(X'X)^{-1}X'\vec{\gamma}\right)'(X'X)\left(-(X'X)^{-1}X'\vec{\gamma}\right)}{rs^2} \\ &= \frac{\vec{\gamma}'X(X'X)^{-1}(X'X)(X'X)^{-1}X'\vec{\gamma}}{rs^2} \end{aligned} \tag{7}$$

como asumimos que $X$ es una matriz de rango completo, se tiene que

$$(X'X)^{-1}(X'X)(X'X)^{-1} = (X'X)^{-1} \tag{8}$$

Por otra parte, para minimizar la suma de cuadrados de los residuales del modelo (5), se muestra en Jiménez (2001) que esto se logra cuando

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

lo cual equivale a la expresión

$$\widehat{\epsilon} + (I - H)\widehat{\gamma} = \vec{0}$$

donde $\widehat{\epsilon}$ es el EMC de $\vec{\epsilon}$ del modelo (1) y asumiendo $\hat{\gamma} = \begin{bmatrix} \hat{\gamma}_1 \\ \vec{0} \end{bmatrix}$ obtiene que

$$\hat{\gamma}_1 = -\vec{Y}_1 + X_1(X_2'X_2)^{-1}X_2'\vec{Y}_2$$

si se reescribe $\widehat{\gamma}$ se llega a

$$\widehat{\gamma} = \begin{bmatrix} \widehat{\gamma}_1 \\ \overrightarrow{0} \end{bmatrix} = \begin{bmatrix} -I_k & X_1(X_2'X_2)^{-1}X_2' \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \overrightarrow{Y_1} \\ \overrightarrow{Y_2} \end{bmatrix} \tag{9}$$

con $k$ la dimensión del bloque $\overrightarrow{Y_1}$; utilizar este valor de $\widehat{\gamma}$ corresponde a estimar los parámetros del modelo (1) después de eliminar el bloque $\vec{Y}_1$.

al reemplazarse (8) y (9) en (7), se obtiene

$$D_{(\vec{Y}_1)} = \frac{\widehat{\gamma}\,'H\widehat{\gamma}}{rs^2}$$

esta nueva expresión de la estadística de Cook tiene la ventaja de que esta en términos del $\widehat{\gamma}$.

El anterior resultado se puede resumir en el siguiente teorema.

## Teorema 1.

Si un modelo de regresión lineal múltiple se particiona como:

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon}_1 \\ \vec{\epsilon}_2 \end{bmatrix} \tag{10}$$

entonces el cambio en la EMC de los parámetros del modelo (10) al eliminar el bloque $\vec{Y}_1$, se calcula mediante la expresión:

$$D_{(\vec{Y}_1)} = \frac{\widehat{\gamma}' H \widehat{\gamma}}{r s^2} \tag{11}$$

donde $s^2 = \dfrac{SCE}{n-r}$ y $\widehat{\gamma} = \begin{bmatrix} \widehat{\gamma}_1 \\ \vec{0} \end{bmatrix}$ con $\widehat{\gamma}_1 = -\vec{Y}_1 + X_1(X_2'X_2)^{-1}X_2'\vec{Y}_2$.

## 3. Distribución de probabilidad de la Estadística de Cook

Si se reemplaza (9) en el numerador de la expresión (11) se obtiene

$$\begin{aligned} \widehat{\gamma}\,' H \widehat{\gamma} =& \overrightarrow{Y}' \begin{bmatrix} -I_k & 0 \\ X_2(X_2'X_2)^{-1}X_1' & 0 \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} -I_k & X_1(X_2'X_2)^{-1}X_2' \\ 0 & 0 \end{bmatrix} \overrightarrow{Y} \\ =& \overrightarrow{Y}' \begin{bmatrix} H_{11} & -H_{11}M_{12} \\ -M_{21}H_{11} & M_{21}H_{11}M_{12} \end{bmatrix} \overrightarrow{Y} \end{aligned} \tag{12}$$

donde $M_{ij} = X_i(X_2'X_2)^{-1}X_j'$ y $H_{ij} = X_i(X'X)^{-1}X_j'$, es una submatriz de la matriz $H$; por otra parte, como $HX = X$, se puede verificar fácilmente que

$$H_{11}X_1 = X_1 - H_{12}X_2 \tag{13}$$
$$H_{21}X_1 = X_2 - H_{22}X_2 \tag{14}$$

reemplazando (13) en las submatrices que aparecen en (12) se tiene que

$$\begin{aligned} H_{11}M_{12} =& H_{11}X_1(X_2'X_2)^{-1}X_2' = [X_1 - H_{12}X_2]\,(X_2'X_2)^{-1}X_2' \\ =& X_1(X_2'X_2)^{-1}X_2' - H_{12} \end{aligned} \tag{15}$$

al sustituir (15) y (13) en la última submatriz de (12) se obtiene

$$\begin{aligned} M_{21}H_{11}M_{12} =& X_2(X_2'X_2)^{-1}X_1'H_{11}X_1(X_2'X_2)^{-1}X_2' \\ =& \left[H_{11}X_1(X_2'X_2)^{-1}X_2'\right]' X_1(X_2'X_2)^{-1}X_2' \\ =& \left[X_1(X_2'X_2)^{-1}X_2' - H_{12}\right]' X_1(X_2'X_2)^{-1}X_2' \\ =& H_{22} + X_2(X_2'X_2)^{-1}(X_1'X_1)(X_2'X_2)^{-1}X_2' - X_2(X_2'X_2)^{-1}X_2' \\ =& H_{22} + M_{21}M_{12} - M_{22} \end{aligned} \tag{16}$$

reemplazando (15) y (16) en (12) se llega a

$$\begin{aligned}\widehat{\gamma}\,'H\widehat{\gamma} =&\overrightarrow{Y}'\begin{bmatrix} H_{11} & H_{12}-M_{12} \\ H_{21}-M_{21} & H_{22}+M_{21}M_{12}-M_{22}\end{bmatrix}\overrightarrow{Y} \\ =&\overrightarrow{Y}'\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22}\end{bmatrix}\overrightarrow{Y}-\overrightarrow{Y}'\begin{bmatrix} 0 & M_{12} \\ M_{21} & M_{22}-M_{21}M_{12}\end{bmatrix}\overrightarrow{Y} \\ =&\overrightarrow{Y}'H\overrightarrow{Y}-\overrightarrow{Y}'M\overrightarrow{Y} \end{aligned} \tag{17}$$

para establecer la distribución de $\widehat{\gamma}\,'H\widehat{\gamma}$ se enuncian sin demostración los teoremas 2 y 3, citados en Searle (1971).

## Teorema 2.

Si $\vec{Y}$ es un vector aleatorio de tamaño $n \times 1$, distribuido $N(\vec{\mu}, V)$; entonces

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

## Teorema 3.

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

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

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

como la expresión dada en (17) es la diferencia de dos formas cuadráticas se establecerá para cada una por aparte la distribución asociada.

Para $\overrightarrow{Y}'H\overrightarrow{Y}$, utilizando el teorema 2 se obtiene que

$$E\left[\frac{\overrightarrow{Y}'H\overrightarrow{Y}}{\sigma^2}\right] = r + 2\eta \qquad Var\left[\frac{\overrightarrow{Y}'H\overrightarrow{Y}}{\sigma^2}\right] = 2r + 8\eta$$

donde $\eta = \dfrac{1}{2\sigma^2}\vec{\beta}'(X'X)\vec{\beta}$, y por el teorema 3 se concluye que $\overrightarrow{Y}'H\overrightarrow{Y}$ tiene distribución Ji-cuadrado no-central, es decir

$$\frac{\overrightarrow{Y}'H\overrightarrow{Y}}{\sigma^2} \sim \chi^{2\,'}_{(\nu,\lambda)} \quad \text{con} \quad \begin{array}{l}\nu = r = \text{rango}(H) \\ \lambda = \dfrac{1}{2\sigma^2}\vec{\beta}'(X'X)\vec{\beta}\end{array}$$

ya que $\frac{1}{\sigma^2} H\sigma^2 I_n$ es idempotente.

Para la expresión $\overrightarrow{Y}' M \overrightarrow{Y}$, se tiene que

$$E\left[\frac{\overrightarrow{Y}' M \overrightarrow{Y}}{\sigma^2}\right] = \left\{r - tr\left[(X_2' X_2)^{-1}(X_1' X_1)\right]\right\} + 2\left\{\frac{1}{2\sigma^2}\vec{\beta}\,'(X'X)\,\vec{\beta}\right\}$$

$$Var\left[\frac{\overrightarrow{Y}' M \overrightarrow{Y}}{\sigma^2}\right] = 2\left\{r + tr\left[(X_2' X_2)^{-1}(X_1' X_1)\right]^2\right\} + 8\left\{\frac{1}{2\sigma^2}\vec{\beta}\,'(X'X)\,\vec{\beta}\right\}$$

puesto que la media y la varianza de la distribución $\chi^{2\,'}_{(\nu,\lambda)}$ son $\nu + 2\lambda$ y $2\nu + 8\lambda$ respectivamente, se deduce que $\overrightarrow{Y}' M \overrightarrow{Y}$ no tiene distribución Ji-cuadrado no-central; y utilizando el teorema 3, se llega a la misma conclusión ya que $\frac{1}{\sigma^2} M\sigma^2 I_n$ no es una matriz idempotente.

Luego,

$$\frac{\widehat{\gamma}\,' H \widehat{\gamma}}{\sigma^2} \nsim \chi^2_{(r)}$$

y por consiguiente, la comparación que hace Cook con la $F_{(r,n-r,\alpha)}$ no es válida, ya que

$$D_{(\vec{Y}_1)} = \frac{\vec{\gamma}\,' H \vec{\gamma}}{r s^2} \nsim F_{(r,n-r)}$$

en esta última expresión se debe tener en cuenta que

$$(n-r)\frac{s^2}{\sigma^2} \sim \chi^2_{(n-r)}$$

## 4. Ejemplo

Para el conjunto de 21 observaciones $(x, y)$ dados por Mickey, Dunn, and Clark (1967) tabla 1, se presentan los siguientes resultados

1. La estimación del modelo de regresión lineal, con las 21 observaciones
2. Los valores $h_{ii}$, las estimaciones de los $\gamma_i$ y la distancia de Cook al eliminar el $i$-ésimo dato.
3. La estimación del modelo de regresión lineal, después de eliminar la observación influyente determinada mediante distancia de Cook

**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.

| 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 | | | |

| | Coeficientes | Error típico | Estadístico $t$ |
|---|---|---|---|
| Intercepto | 109.8738 | 5.0678 | 21.6808 |
| Variable $X$ | -1.1270 | 0.3102 | -3.6334 |

Coeficiente de determinación $R^2 = 0,409971261$

Error típico $\widehat{\sigma} = 11{,}0229086$

2. Puesto que $s^2 = 121{,}504515$, se tiene que

| **Obs Elim** | $h_{ii}$ | $\widehat{\gamma}_i$ | $D_i$ $(100{*}D_i)$ | **Obs Elim** | $h_{ii}$ | $\widehat{\gamma}_i$ | $D_i$ $(100{*}D_i)$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.0479 | -2.1332 | 0.09 | 12 | 0.0705 | 4.0141 | 0.47 |
| 2 | 0.1545 | 11.3214 | 8.15 | 13 | 0.0628 | 16.6498 | 7.17 |
| 3 | 0.0628 | 16.6498 | 7.17 | 14 | 0.0567 | 14.2866 | 4.76 |
| 4 | 0.0705 | 9.3936 | 2.56 | 15 | 0.0567 | -4.7948 | 0.54 |
| 5 | 0.0479 | -9.4856 | 1.77 | 16 | 0.0628 | -1.4896 | 0.06 |
| 6 | 0.0726 | 0.3602 | 0.00 | 17 | 0.0521 | -9.1255 | 1.79 |
| 7 | 0.0580 | -3.6220 | 0.31 | 18 | 0.6516 | 15.9026 | 67.81 |
| 8 | 0.0567 | -2.6746 | 0.17 | 19 | 0.0531 | -31.9816 | 22.33 |
| 9 | 0.0799 | -3.4148 | 0.38 | 20 | 0.0567 | 12.1664 | 3.45 |
| 10 | 0.0726 | -7.1879 | 1.54 | 21 | 0.0628 | -1.4896 | 0.06 |
| 11 | 0.0908 | -12.1145 | 5.48 | | | | |

En los resultados anteriores se verifica que los valores de la distancia de Cook, corresponden a la expresión

$$D_i = \frac{h_{ii}}{2}\left(\frac{\widehat{\gamma}_i}{s}\right)^2$$

Según los cálculos realizados, la observación que puede ser considerada como influyente sobre la estimación de los parámetros es la observación 18, pues nótese que es la única que cumple que $D_i > 0{,}5$.

3. Cuando se elimina la observación 18 y ajustamos los datos a un nuevo modelo, se obtiene la siguiente tabla de análisis de varianza

| 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 | | | |

| | Coeficientes | Error típico | Estadístico $t$ |
|---|---|---|---|
| Intercepto | 105.62987 | 7.1619276 | 14.7488045 |
| Variable $X$ | -0.77922 | 0.516733 | -1.5079754 |

Coeficiente de determinación $R^2 = 0,112162$

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

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 inicial.

## 5. Conclusiones

En este artículo se obtuvo la generalización de una de las medidas más utilizadas para el estudio de las observaciones influenciales. La generalización aquí presentada detecta la influencia de un grupo de observaciones sobre el cambio en el centro de la región elipsoidal de confianza para $\vec{\beta}$, de manera análoga a como lo hace la distancia de Cook.

## Referencias

[1] COOK, R.D. (1977) *Detection of Influential Observations in Linear Regression.* Technometrics, vol. 19, pag. 15-18.

[2] COOK, R.D., and WEISBERG, S. (1980) *Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression.* Technometrics, vol. 22, pag. 495-508.

[3] JIMÉNEZ, J.A. (1999) *Propuesta Metodológica para Imputar Valores no Influyentes en Modelos de Regresión Lineal Múltiple con Información Incompleta.* Universidad Nacional de Colombia, Tesis de Maestría

[4] JIMÉNEZ, J.A. y RINCÓN, L.F. (2000) *Una generalización de la Estadística DF-Beta.* En: Revista Colombiana de Estadística, vol. 23, $N^o 1$

[5] JIMÉNEZ, J.A. (2001) *Una Maximización de la Estadística* $Q_k$. En: Revista Colombiana de Estadística, vol. 24, No. 1

[6] MICKEY, M. R., DUNN, O. J., and CLARK, V. (1967) *Note on the use of stepwise regression in detecting outliers.* Computers and Biomedical Research, 1, pag 105-111

[7] SEARLE, S. R. (1971) *Linear Models.* John Wiley & Sons, New York