**Revista Colombiana de Estadística**
Volumen **23** (2000) No. 1, páginas 27–34

# UNA GENERALIZACIÓN DE LA ESTADÍSTICA DFBETA

JAJ MOSCOSO[1]
LFR SUARÉZ[2]

**Resumen.** En este artículo se presenta una generalización de la estadística *DFBeta* con la cual se logra cuantificar el impacto que ejercen un grupo de obsevaciones seleccionadas, en la estimación vía mínimos cuadrados del modelo de regresión lineal múltiple.

*Palabras claves:* Modelos lineales, mínimos cuadrados, estadística DFBeta.

## 1. Introducción

En la estimación mínimos cuadrados par detectar específicamente la influencia que una observación seleccionada ejerce en la estimación de los parámetros del modelo de regresión lineal $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$. La estadística *DFBeta(i)* presentada en Belsley y colaboradores (1980) es la más reconocida al respecto. Para la $i$–ésima observación el valor de la estadística se obtiene a partir de la expresión:

$$DFBeta(i) = \frac{\hat{\epsilon}_i}{1 - h_{ii}} c_i, \qquad 1 \leq i \leq n \tag{1}$$

con $c_i$ la $i$–ésima fila de la matriz $C = (X'X)^{-1}X'$, $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ y $h_{ii}$ el $i$–ésimo elemento en la diagonal de la matriz $H = (X(X'X)^{-1}X')$ siendo este valor DFBeta(i) la diferencia entre los parámetros estimados al eliminar la $i$–ésima observación.

(1) Magister en Estadística. Universidad Nacional de Colombia; e-mail: @matematicas.unal.edu.co
(2) Profesor Asociado, Departamento de Estadística, Universidad Nacional de Colombia; e-mail: frincon@matematicas.unal.edu.co.

Por su importancia se logró la generalización de esta estadística en modelos de regresión lineal simple que permite detectar la influencia que un grupo de observaciones ejerce en la estimación de los parámetros, Rincón y López (1997); e interesa en este artículo presentar una generalización que notaremos $DFBeta(Y_1)$ y que permite medir los cambios que ejercen las observaciones contenidas en el bloque $Y_1$ sobre los parámetros asociados a un modelo de regresión lineal múltiple.

## 2. Derivación de la estadística DFBeta($\vec{Y}_1$)

Para el modelo de regresión lineal múltiple

$$\vec{Y} = X\vec{\beta} + \vec{\epsilon} \tag{2}$$

mediante la estimación vía mínimos cuadrados se obtiene el estimador $\hat{\beta}$ de los parámetros $\vec{\beta}$, los valores estimados de $\vec{Y}$, los residuales $\hat{\epsilon}$ y la suma de cuadrados de los residuales SCE de acuerdo con las siguientes expresiones:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'\vec{Y} \\
\hat{Y} &= X\hat{\beta} = X(X'X)^{-1}X'\vec{Y} = H\vec{Y} \text{ con } H = X(X'X)^{-1}X' \\
\hat{\epsilon} &= \vec{Y} - \hat{Y} = \vec{Y} - H\vec{Y} = (I-H)\vec{Y} \\
SCE &= \hat{\epsilon}'\hat{\epsilon} = [(I-H)\vec{Y}]'(I-H)\vec{Y} = \vec{Y}'(I-H)\vec{Y},
\end{aligned}$$

Se considera el modelo expresado en (2), particionado como

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

Para medir la influencia que ejerce el bloque $Y_1$ de dimendión $k \times 1,\ k < n$, en la estimación de los parámetros vía mínimos cuadrados; se modifica cada una de las componentes del bloque $Y_1$ con constantes arbitrarias $\gamma_i,\ i = 1, 2, ..., k$ y se plantea el modelo

$$\vec{Y}^* = X\vec{\beta}^* + \vec{\epsilon}^* \tag{3}$$

siendo $\vec{Y}^* = \vec{Y} + \vec{\gamma}$, donde $\vec{\gamma} = \begin{bmatrix} \vec{\gamma}_1 \\ 0 \end{bmatrix}$, es decir el modelo (3) particionado se puede escribir ahora en la forma

$$\begin{bmatrix} \vec{Y_1} \\ \vec{Y_2} \end{bmatrix} + \begin{bmatrix} \vec{\gamma}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta}^* + \begin{bmatrix} \vec{\epsilon_1^*} \\ \vec{\epsilon_2^*} \end{bmatrix} \tag{4}$$

interesa establecer las expresiones de los "nuevos" estimadores, en función de $\vec{\gamma}$, $\vec{Y}$ y de los estimadores obtenidos para el modelo (2). El estimador del vector

$\vec{\beta}^*$, obtenido por el método de mínimos cuadrados, es dado por la siguiente expresión

$$\hat{\beta}^* = (X'X)^{-1}X'\vec{Y}^*$$

reemplazando $\vec{Y}^*$ se tiene

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(\vec{Y} - \vec{\gamma}) \\ &= (X'X)^{-1}X'\vec{Y} + (X'X)^{-1}X'\vec{\gamma} \\ &= \hat{\beta} + (X'X)^{-1}X'\vec{\gamma} \\ &= \hat{\beta} + (X'X)^{-1}\begin{bmatrix} X_1' & X_2' \end{bmatrix}\begin{bmatrix} \vec{\gamma}_1 \\ 0 \end{bmatrix}\end{aligned}$$

donde se concluye que

$$\hat{\beta} - \hat{\beta}^* = -(X'X)^{-1}X_1'\vec{\gamma}_1 \tag{5}$$

De la misma forma, el "nuevo" vector de predicciones $\hat{Y}^*$ se obtiene de acuerdo con

$$\begin{aligned}\hat{Y}^* &= X\hat{\beta}^* \\ &= X(\hat{\beta} + (X'X)^{-1}X'\vec{\gamma}) \\ &= \hat{Y} + X(X'X)^{-1}X'\vec{\gamma} \\ &= \hat{Y} + X(X'X)^{-1}X'\vec{\gamma} \\ &= \hat{Y} + H\vec{\gamma} \qquad (6)\end{aligned}$$

Bajo la partición dada en (4) esta ecuación es equivalente a

$$\begin{bmatrix} \vec{Y}_1^* \\ \vec{Y}_2^* \end{bmatrix} = \begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} + \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}\begin{bmatrix} \vec{\gamma}_1 \\ 0 \end{bmatrix}$$

con $H_{ij} = X_i(X'X)^{-1}X_j'$.

Con la misma metodología se obtiene el vector de errores estimado para el modelo (3) según

$$\begin{aligned}\hat{\epsilon}^* &= \vec{Y}^* - \hat{Y}^* \\ &= (\vec{Y} + \vec{\gamma}) - (\hat{Y} + H\vec{\gamma}) \\ &= (\vec{Y} - \hat{Y}) - (\vec{\gamma} - H\vec{\gamma}) \\ &= \hat{\epsilon} - (I - H)\vec{\gamma} \qquad (7)\end{aligned}$$

Bajo la misma partición la ecuación (7) se expresa como

$$\begin{bmatrix} \hat{\epsilon}_1^* \\ \hat{\epsilon}_2^* \end{bmatrix} = \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \end{bmatrix} + \begin{bmatrix} I - H_{11} & -H_{12} \\ -H_{21} & I - H_{22} \end{bmatrix} \begin{bmatrix} \vec{\gamma}_1 \\ 0 \end{bmatrix}$$

De tal manera que el vector $\vec{\gamma}_1$ se hace $\hat{\epsilon}_1^* = 0$ está dado por

$$\hat{\gamma}_1 = -(I - H_{11})^{-1}\hat{\epsilon}_1 \tag{8}$$

término que al reemplazarse en la ecuación (5) proporciona la expresión para calcular los valores de la estadística $DFBeta(\vec{Y}_1)$ según

$$\hat{\beta} - \hat{\beta}^*_{\vec{Y}_1} = -(X'X)^{-1}X_1'\vec{\gamma}_1$$

$$DFBeta(\vec{Y}_1) = (X'X)^{-1}X_1'(I - H_{11})^{-1}\vec{\epsilon}_1. \tag{9}$$

Nótese que $DFBeta(\vec{Y}_1)$ es un vector de dimensión $r \times 1$ el cual mide el efecto que tienen los $k$ registros del bloque $\vec{Y}_1$, en la estimación vía mínimos cuadrados en cada una de las componentes del vector de parámetros $\vec{\beta}$, siendo $\vec{\beta}$ el vector de parámetros estimados en presencia de todas las observaciones y $\hat{\beta}^*_{\vec{Y}_1}$ el vector de parámetros estimados despues de eliminar las observaciones contenidas en el bloque $Y_1$.

El anterior resultado se puede resumir en el siguiente teorema.

**Teorema 1.** *En un modelo de regresión lineal* $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$*, particionado como:*

$$\begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \vec{\beta} + \begin{bmatrix} \vec{\epsilon_1} \\ \vec{\epsilon_2} \end{bmatrix}$$

*con* $\vec{Y}_1$ *de dimensión* $k$*,*

$$DFBeta(\vec{Y}_1) = (X'X)^{-1}X_1'(I - H_{11})^{-1}\vec{\epsilon}_1$$

*siendo* $H_{11} = X_1(X'X)^{-1}X_1'$.

## 3. Distribución de probabilidad de la estadística $DFBeta(\vec{Y}_1)$

Para el método particionado (7) y bajo el supuesto de normalidad de los residuales se satisface que

$$\hat{\epsilon}_1 \sim N(0, \sigma^2(I - H_{11}))$$

y se muestra, Rincón (1999) que $\hat{\gamma}_1$ definido en (8) satisface

$$\hat{\gamma}_1 \sim N(0, \sigma^2(I - H_{11})^{-1}) \tag{11}$$

es decir que cada una de las componentes $\gamma_i,\ i = 1, ..., k$ de $\hat{\gamma}_1$ se distribuye según

$$\hat{\gamma}_1 \sim N(0, \sigma^2 H_i) \tag{12}$$

donde $H_i$ es el $i$-ésimo elemento de la diagonal de la matriz $(I - H_1)^{-1}$.

Conocida la distribución de $\hat{\gamma}_1$ obtenida en (10) y reescribiendo la estadística $DFBeta(\vec{Y}_1)$ como

$$DFBeta(\vec{Y}_1) = -(X'X)^{-1}X_1'\hat{\gamma}_1$$

se obtiene que

$$E(DFBeta(\vec{Y}_1)) = -(X'X)^{-1}X_1'E(\hat{\gamma}_1) = 0$$

y

$$\begin{aligned} Var(DFBeta(\vec{Y}_1)) &= (X'X)^{-1}X_1'V(\hat{\gamma}_1)[(X'X)^{-1}X_1']' \\ &= \sigma^2 C(I - H_{11})^{-1}C' \end{aligned}$$

con $C = (X'X)^{-1}X_1'$ para establecer finalmente que

$$DFBeta(\vec{Y}_1) \sim N(0, \sigma^2 C(I - H_{11})^{-1}C') \tag{13}$$

En particular denotaremos por $M_j$ el $j$-ésimo elemento de la diagonal de $C(I - H_{11})^{-1}C'$ para cada $j = 1, ..., r$ la dimensión de la $DFBeta(\vec{Y}_1)$ resulta que

$$DFBeta_j(\vec{Y}_1) \sim N(0, \sigma^2 M_j) \tag{14}$$

Y finalmente se obtiene de la aplicación del estimador insesgado de $\sigma^2$, $\hat{\sigma}^2 = \frac{SCE}{n-r}$ que para cada $j = 1, 2, ..., r$

$$\frac{DFBeta(\vec{Y}_1)}{S\sqrt{M_j}} = t_j \sim t_{(n-r)} \tag{15}$$

De donde resulta que los valores de las estadísticas $t_j$ definidos en (15) son iguales para cada $j$, cuando $\vec{Y}_1$ consta de un única observación y difieren $\vec{Y}_1$ tiene más de una observación, es decir que el bloque $\vec{Y}_1$ cuando tiene más de una observación puede ser influyente sobre alguno o algunos de los $r$ parámetros y no influyente para los demás.

El anterior resultado se puede resumir en el siguiente teorema.

**Teorema 2.** *Para el modelo de regresión lineal múltiple* $\vec{Y} = X\vec{\beta} + \vec{\epsilon}$*, con* $\epsilon \sim N(0, \sigma^2 I)$ *se tiene que*

$$DFBeta(\vec{Y_1}) \sim N(0, \sigma^2 C(I - H_{11})^{-1} C')$$

*donde* $H_{11} = X_1(X'X)^{-1}X_1'$ *y* $C = (X'X)^{-1}X_1'$.

## 4. Ejemplo

Para los datos citados en Cook y Weisberg (1982) tabla 1, se presentan los siguientes resultados, procesados mediante el paquete SAS

1. La estimación del modelo de regresión lineal, con las 21 observaciones.
2. La estadística $DFBeta(\vec{Y_1})$, para el bloque $Y_1$ compuesto por las primeras 4 observaciones.
3. La estimación del modelo de regresión lineal, después de eliminar el bloque $Y_1$.

Tabla 1. Datos de Weisberg (1982)

| **OBS** | $X_1$ | $X_2$ | $X_3$ | $Y$ | **OBS** | $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 58 | 17 | 88 | 13 | 12 | 58 | 23 | 87 | 15 |
| 2 | 62 | 24 | 87 | 28 | 13 | 56 | 20 | 82 | 15 |
| 3 | 80 | 27 | 89 | 42 | 14 | 58 | 18 | 82 | 11 |
| 4 | 62 | 22 | 87 | 18 | 15 | 50 | 19 | 72 | 8 |
| 5 | 75 | 25 | 90 | 37 | 16 | 50 | 18 | 89 | 8 |
| 6 | 62 | 23 | 87 | 18 | 17 | 58 | 18 | 80 | 14 |
| 7 | 80 | 27 | 88 | 37 | 18 | 50 | 18 | 86 | 7 |
| 8 | 58 | 18 | 89 | 14 | 19 | 50 | 20 | 80 | 9 |
| 9 | 62 | 24 | 93 | 19 | 20 | 50 | 19 | 79 | 8 |
| 10 | 62 | 24 | 92 | 20 | 21 | 58 | 19 | 93 | 12 |
| 11 | 70 | 20 | 91 | 15 | | | | | |

1)

| | Grados de libertad | Suma de cuadrados | Promedio de los cuadrado | F | valor crítico de F |
|---|---|---|---|---|---|
| Regresión | 3 | 1890.408134 | 630.1360445 | 59.9022259 | 3.01633E-09 |
| Residuos | 17 | 178.8299616 | 10.51940951 | | |
| Total | 20 | 2069.238095 | | | |

| | Coeficientes | Error típico | Estadístico t |
|---|---|---|---|
| Intercepto | -39.91967442 | 11.89599685 | -3.3557223351 |
| Variable X1 | 0.7156402 | 0.134858185 | 5.306613007 |
| Variable X2 | 1.295286124 | 0.368024265 | 3.519567177 |
| Variable X3 | -0.152122519 | 0.156294043 | -0.973309769 |

| | Probabilidad | Inferior 95% | Superior 95% |
|---|---|---|---|
| Intercepto | 0.003750307 | -65.01806894 | -14.8212799 |
| Variable X1 | 5.79902E-05 | 0.431113903 | 1.000166498 |
| Variable X2 | 0.002630054 | 0.518821712 | 2.071750537 |
| Variable X3 | 0.344046097 | -0.481874587 | 0.177629549 |

2)

$$(I - H_{11}) = \begin{bmatrix} 0.782824035 & 0.063126555 & 0.025450124 & -0.025467757 \\ 0.063126555 & 0.871494757 & -0.084232713 & -0.064611938 \\ 0.025450124 & -0.084232713 & 0.698444531 & -0.068031365 \\ -0.02546776 & -0.064611938 & -0.068031365 & 0.94777967 \end{bmatrix}$$

$$(I - H_{11})^{-1} = \begin{bmatrix} 1.28785352 & -0.096949517 & -0.056285877 & 0.023956453 \\ -0.09694952 & 1.175928839 & 0.153981465 & 0.08861292 \\ -0.05628588 & 0.153981465 & 1.463481404 & 0.114033046 \\ 0.023956453 & 0.08861292 & 0.114033046 & 1.069967446 \end{bmatrix}$$

$$(X'X)^{-1}X_1'(I - H_{11})^{-1} =$$

$$\begin{bmatrix} -0.019263008 & -0.275521575 & -0.35389412 & -0.128377588 \\ 0.011490704 & -0.008399076 & 0.015018195 & -0.000144364 \\ -0.06045298 & 0.04365796 & 0.019974209 & 0.009322737 \\ 0.00759496 & -0.000868941 & -0.010375098 & 2.52344E-05 \end{bmatrix}$$

$$DFBeta(Y_1) = \begin{bmatrix} -2.54838181 \\ 0.03290743 \\ 0.129378546 \\ -0.01744401 \end{bmatrix}$$

3)

| | Grados de libertad | Suma de cuadrados | Promedio de los cuadrado | F | Valor crítico de F |
|---|---|---|---|---|---|
| Regresión | 3 | 1168.645655 | 389.881885 | 44.50559633 | 4.2436E-07 |
| Residuos | 13 | 113.8837567 | 8.760288978 | | |
| Total | 16 | 1283.529412 | | | |

| | Coeficientes | Error típico | Estadístico t |
|---|---|---|---|
| Intercepto | -37.37129261 | 10.9135109 | -3.424314406 |
| Variable X 1 | 0.68273277 | 0.134640227 | 5.070793353 |
| Variable X 2 | 1.165907578 | 0.387805671 | 3.006422197 |
| Variable X 3 | -0.134678508 | 0.146107275 | -0.921778248 |

| | Probabilidad | Inferior 95% | Superior 95% |
|---|---|---|---|
| Intercepto | 0.004525916 | -60.94849496 | -13.79409027 |
| Variable X 1 | 0.000214461 | 0.391860299 | 0.973605241 |
| Variable X 2 | 0.010113218 | 0.328104523 | 2.003710633 |
| Variable X 3 | 0.373434932 | -0.450324024 | 0.180967008 |

De los resultados anteriores se verifica que los valores del vector

$$DFBeta(Y_1) = (-2.548381, 0.032907, 0.129378, -0.017444)$$

corresponden a la expresión $\hat{\beta} - \hat{\beta}^*_{(\vec{Y}_1)}$.

## Referencias


[1] Belsley, D. et al (1980) Regression diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Jhon Wiley.

[2] Cook, R. D. and Weisberg, S. (1982), *R*esiduals and Influence in Regression, New York: Chapman & Hall.

[3] Rincón, Tatiana (1999), *U*na propuesta para caracterizar observaciones influyentes en modelos de regresión lineal múltiple Trabajo de grado (Estadística); Universidad Nacional de Colombia. Facultad de Ciencias, Departamento de Matemáticas y Estadística. Sede: Bogotá.

[4] Rincón Luis F., López Luis A. (1997) *U*na Generalización de la Estadística DFBeta en mdelos de regresión lineal simple, En: Revista Colombiana de Estadística, No. 35.

[5] Searle, S. R., (1971), *L*inear Models, New York: John Wiley & Sons.

[6] Tukey, J., (1971), *E*xploratory Data Analysis, Reading, M.A: Addison Wesley.