Revista Colombiana de Estadística

No.35

UNA GENERALIZACIÓN DE LA ESTADÍSTICA DFBeta EN MODELOS DE REGRESIÓN LINEAL SIMPLE

LUIS FRANCISCO RINCÓN[a] Y LUIS ALBERTO LÓPEZ[b]

[a]Profesor Asociado del Departamento de Matemáticas y Estadística de la Universidad Nacional de Colombia. e-mail:lurincon@ciencias.ciencias.unal.edu.co
[b]Profesor Asociado del Departamento de Matemáticas y Estadística de la Universidad Nacional de Colombia. e-mail: llopez@ciencias.ciencias.unal.edu.co
Proyecto de investigación financiado por COLCIENCIAS y CINDEC.

RESUMEN En este trabajo se presenta una metodología para el cálculo de la estadística $DFBeta$, cuando se quiere medir la influencia que ejerce un grupo de k observaciones $k < n$, en la estimación de los parámetros obtenidos via mínimos cuadrados, para un modelo de Regresión Lineal Simple.

PALABRAS CLAVES. Mínimos Cuadrados , Regresión Lineal Simple , Residuales, outliers, ajuste, puntos influyentes.

1.INTRODUCCIÓN

En BELSLEY et al (1980) se presenta como medida de diagnóstico en Regresión, la estadística $DFBeta(i)$ que mide la influencia que ejerce la i-ésima observación sobre el estimador de mínimos cuadrados del vector $\beta = (\beta_0, ..., \beta_p)$ asociado a un modelo de Regresión lineal. Cuando se remueve la i-ésima observación esta estadística se obtiene a partir de la expresión:

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

con c_i la línea i de la matriz $c = (X'X)^{-1} X'$, $\hat{e}_i = Y_i - \hat{Y}_i$, y h_{ii} el i-ésimo elemento en la diagonal de la matriz $H = X\,(X'X)^{-1} X'$ la cual se conoce como "Matriz Hat",

nombre dado por TUKEY (1977).

La remoción de un conjunto de k observaciones y su influencia en la estimación de los parámetros no ha sido estudiada cuando se usa como método de diagnóstico la estadística *DFBeta*; en BECKMAN and COOK (1983), MARASINGHE (1985), PAUL and FUNG (1991), se presentan algunas alternativas basadas en otros estadísticos, que no conducen a evaluar los cambios en los parámetros, sino los cambios en los cuadrados medios del error. Considerando la importancia que tiene la estadística *DFBeta*; al medir los cambios que ejerce una observación sobre los parámetros del modelo, se presenta en este trabajo una generalización que notaremos $DFBeta(1,...,k)$; y que permite medir la influencia de k observaciones en la estimación de los parámetros asociados a un Modelo de Regresión Lineal Simple.

2. CÁLCULO DE LA ESTADÍSTICA $DFBeta(1,...,K)$

Para el modelo de Regresión $Y_i = \alpha + \beta X_i + e_i$, $i = 1, 2, \ldots, n$, con X_i vector de valores conocidos , y e_i vector de errores independientes e idénticamente distribuidos $N(0, \sigma^2)$, sin pérdida de generalidad interesa medir la influencia que ejercen las primeras k observaciones en la estimación de los parámetros via mínimos cuadrados. Para ello modificamos la componente y_i en cada una de ellas, con constantes arbitrarias γ_i, $i = 1, \ldots, k$ definiendo

$$y_i^* = \begin{cases} y_i + \gamma_i & \text{si} \quad i = 1, \ldots, k \\ y_i & \text{si} \quad k < i \leq n \end{cases}$$

Los nuevos estimadores α^* y β^*, obtenidos por mínimos cuadrados del modelo modi-

ficado satisfacen:

$$\hat{\alpha}^* = \frac{\sum_{i=1}^{k} \gamma_i}{n} - \frac{\sum_{i=1}^{k} \gamma_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \bar{x} + \hat{\alpha}$$
$$\hat{\beta}^* = \frac{\sum_{i=1}^{k} \gamma_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \hat{\beta} \qquad (1)$$

siendo $\hat{\alpha}$ y $\hat{\beta}$ las estimaciones obtenidas via mínimos cuadrados del modelo original. Para las nuevas estimaciones, los nuevos residuales $\hat{e}_i^* = y_i^* - \hat{y}_i^*$ safisfacen:

$$\hat{e}_i^* - \hat{e}_i = \gamma_i (1 - h_{ii}) - \sum_{i \neq j}^{k} \gamma_j h_{ij} \qquad \text{si} \quad 1 \leq i \leq n$$
$$\hat{e}_j^* - \hat{e}_j = -\sum_{i=1}^{k} \gamma_i h_{ij} \qquad \text{si} \quad k < j \leq n \qquad (2)$$

siendo h_{ij} el elemento de la i-ésima fila y de la j-ésima columna de la matriz H. Los desarrollos algebraicos necesarios para obtener 1 y 2 se presentan en los apéndices 1 y 2 respectivamente.

La expresión 2 permite determinar constantes de ajuste γ_i que hacen $\hat{e}_i^* = 0$ para $i = 1, ..., k$, como solución del sistema :

$$-\hat{e}_i = \gamma_i (1 - h_{ii}) - \sum_{i \neq j}^{k} \gamma_j h_{ij} \qquad \text{con} \quad i = 1, ..., k \qquad (3)$$

Las cuales coinciden con las constantes de ajuste del modelo de DRAPER and JOHN (1981) diseñado para detectar un grupo de k outliners.

Dado que estimar el modelo después de ajustar las respuestas Y_i con las constantes γ_i para $1 \leq i \leq k$. equivale a estimar el modelo después de eliminar las k observaciones

seleccionadas, y como la estimación de los parámetros después del ajuste es función de la estimación inicial y de las constantes de ajuste, esto permite la construcción de la estadística

$$DFBeta(1,...,k) = \left(\hat{\alpha}^* - \hat{\alpha}; \hat{\beta}^* - \hat{\beta}\right)$$

con

$$\begin{aligned} \hat{\alpha}^* - \hat{\alpha} &= \frac{\sum_{i=1}^{k} \hat{\gamma}_i}{n} - \frac{\sum_{i=1}^{k} \hat{\gamma}_i (x_i - \bar{x})}{\sum_{i=1} (x_i - \bar{x})^2} \bar{x} \\ \hat{\beta}^* - \hat{\beta} &= \frac{\sum_{i=1}^{k} \hat{\gamma}_i (x_i - \bar{x})}{\sum_{i=1} (x_i - \bar{x})^2} \end{aligned} \tag{4}$$

Estadística que mide el efecto en la estimación de los parámetros obtenidos por mínimos cuadrados, cuando se eliminan o corrigen las primeras k observaciones.

3. APLICACIÓN

Usando la estadística $DFBeta(1,...,k)$ definida en (4) explícitamente se presentan los cálculos para medir el efecto de eliminar las dos primeras observaciones en un modelo de Regresión Lineal Simple. La aplicación de (3) proporciona el sistema:

$$\begin{aligned} \hat{e}_1 &= \gamma_2 h_{12} - \gamma_1 (1 - h_{11}) \\ \hat{e}_2 &= \gamma_1 h_{21} - \gamma_2 (1 - h_{22}) \end{aligned}$$

cuya solución

$$\gamma_1 = \frac{h_{12} e_2 + (1 - h_{22}) e_1}{(1 - h_{11})(1 - h_{22}) - h_{12}^2}; \qquad \gamma_2 = -\frac{(1 - h_{11}) e_2 + h_{21} e_1}{(1 - h_{11})(1 - h_{22}) - h_{12}^2}$$

permite caracterizar la estadística

$$DFBeta\,(1,2) = \left(\frac{\hat{\gamma}_1 + \hat{\gamma}_2}{n} - \frac{\hat{\gamma}_1\,(x_1 - \bar{x}) + \hat{\gamma}_2\,(x_2 - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\bar{x};\ \frac{\hat{\gamma}_1\,(x_1 - \bar{x}) + \hat{\gamma}_2\,(x_2 - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Los cálculos anteriores se pueden generalizar para obtener la estadística $DFBeta\,(i,j)$ de cualquier pareja de observaciones $(\mathcal{O}_i, \mathcal{O}_j)$ $i \neq j$ definiendo:

$$\hat{\alpha}^* - \hat{\alpha} = \frac{\hat{\gamma}_i + \hat{\gamma}_j}{n} - \frac{\hat{\gamma}_i(x_i - \bar{x}) + \hat{\gamma}_j(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\bar{x}$$

$$\hat{\beta}^* - \hat{\beta} = \frac{\hat{\gamma}_i(x_i - \bar{x}) + \hat{\gamma}_j(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\bar{x}$$

cuando

$$\hat{\gamma}_i = \frac{h_{ij}e_j + (1 - h_{jj})\,e_i}{(1 - h_{ii})(1 - h_{jj}) - h_{ij}^2};\quad \hat{\gamma}_j = -\frac{(1 - h_{ii})\,e_j + h_{ji}e_i}{(1 - h_{ii})(1 - h_{jj}) - h_{ij}^2}$$

y así

$$DFBeta\,(i,j) = \left(\frac{\hat{\gamma}_i + \hat{\gamma}_j}{n} - \frac{\hat{\gamma}_i\,(x_i - \bar{x}) + \hat{\gamma}_j\,(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\bar{x};\ \frac{\hat{\gamma}_i\,(x_i - \bar{x}) + \hat{\gamma}_j\,(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Al considerar las opciones

-Signo y tamaño de las constantes de ajuste $\hat{\gamma}_i, \hat{\gamma}_j$.

-Posición de la observación con respecto a $\bar{x}$.

Se puede deducir si el efecto individual de una observación **anula** el efecto de otra.

o si los efectos individuales se suman presentándose **efecto aditivo de la pareja de observaciones** sobre la estimación.

4.EJEMPLO NUMÉRICO

Como ilustración numérica para los datos de Mickey, Dunn y Clark (1967) citados en DRAPER and JOHN (1981), se presenta en la tabla 1 calculada en una hoja electrónica:

-La constante de ajuste $\hat{\gamma}_1$ y la estadística

$DFBeta(1) = \left(\hat{\alpha}^* - \hat{\alpha}; \hat{\beta}^* - \hat{\beta}\right) = (A_1 - A,\ B_1 - B\ )$, para cada observación.

-Las constantes de ajuste $\hat{\gamma}_1$ y $\hat{\gamma}_i$ y la estadística

$DFBeta(1,i) = \left(\hat{\alpha}^* - \hat{\alpha}; \hat{\beta}^* - \hat{\beta}\right) = (A_2 - A,\ B_2 - B\ )$, para las posibles parejas $(\mathcal{O}_1, \mathcal{O}_i)$, $i \geq 2$.

-Las constantes de ajuste $\hat{\gamma}_1, \hat{\gamma}_2$ y $\hat{\gamma}_i$ y la estadística $DFBeta(1,2,i) = \left(\hat{\alpha}^* - \hat{\alpha}; \hat{\beta}^* - \hat{\beta}\right) = (A_3 - A,\ B_3 - B\ )$, para todas las posibles ternas $(\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_i)$, $i \geq 3$.

-El residual $\hat{e}_i$ para cada observación.

-Los resultados del análisis de varianza del Modelo inicial.

Algunas conclusiones que se derivan de los resultados ofrecidos y que se presentan en las dos tablas siguientes son:

-Se observa que es la octava observación con $DFBeta(i) = (-4.244, 03478)$, la de mayor influencia sobre la estimación de los parámetros , si fuera

eliminada, tabla1. Esta influencia se debe a su posición con respecto a $\bar{x}$, ya que su residual $\hat{e}_8 = -5,54$ no es de los más altos.

-En la tabla 1 se observa también que la primera observación aunque presenta el residual más alto, $\hat{e}_1 = 30.285$ no ofrece la mayor influencia en la estimación si fuera eliminada.

-Con la hoja electrónica es muy sencillo calcular la estadística $DFBeta\,(i,j)$ de la pareja conformada por las observaciones i,y j para un i dado. Basta subir la observación i al primer lugar de la tabla para obtener la estadística $DFBeta$ de todas las parejas con la i-ésima observación como primera componente . Realizado el procedimiento, es decir, calculada la estadística $DFBeta$ para todas las posibles parejas, se presenta la tabla 2, en donde la pareja conformada por las observaciones 7 y 8 del listado original, es la de mayor influencia en la estimación de los parámetros, con $DFBeta\,(i,j) = (-12.01, 1.0399)$.

Tabla-1

ESTADÍSTICA DFBeta

N=	21	B=	-1,127	Total=	3912,67
Ymed=	93,6667	A=	109,874	Model=	1604,08
Xmed=	14,381			Error=	2308,59
				S=	11,0229
				R^2=	0,40997
				V(B)=	0,09621
				V(A)=	25,6826

N	X	Y	R1	A1-A	B1-B	R1	R(i)	A2-A	B2-B	R1	R2	R(i)	A3-A	B3-B	E(i)
1	17	121	31,9816	-0,5692	-0,0663										30,285
2	10	83	-16,65	1,62342	-0,0578	31,3565	-15,361	0,93968	-0,1183						-15,604
3	10	83	-16,65	1,62342	-0,0578	31,3565	-15,361	0,93968	-0,1183	30,6391	-16,496	-16,466	2,6686	-0,1779	-15,604
4	11	84	-14,287	1,23032	-0,0382	31,427	-12,934	0,55453	-0,0998	30,7196	-16,273	-13,957	2,24184	-0,1575	-13,477
5	11	86	-12,166	1,04774	-0,0326	31,5181	-10,81	0,36999	-0,0943	30,8167	-16,136	-11,824	2,04315	-0,1515	-11,477
6	7	113	12,1145	-1,5951	0,0708	32,4343	13,2672	-2,324	0,01028	31,8081	-14,4	12,1113	-0,7566	-0,0451	11,0151
7	26	71	-11,321	-0,9587	0,10416	31,3254	-8,6644	-1,2912	0,01475	30,6822	-15,45	-8,7911	0,2159	-0,0363	-9,5721
8	42	57	-15,903	-4,244	0,34777	31,2627	-6,49	-2,2884	0,0771	30,8651	-15,214	-4,2881	-0,2102	-0,023	-5,5403
9	15	102	9,48556	-0,3848	-0,0046	32,5578	11,1579	-1,0321	-0,073	31,916	-14,994	10,4538	0,46987	-0,1233	9,03099
10	9	91	-9,3936	1,02288	-0,04	31,6678	-8,1514	0,32403	-0,1004	30,9701	-16,052	-9,2938	2,02601	-0,1595	-8,7309
11	12	105	9,12554	-0,682	0,0172	32,4588	10,5871	-1,3688	-0,0474	31,8175	-14,761	9,72058	0,14663	-0,0989	8,65003
12	20	94	7,18792	0,11762	-0,032	32,5618	9,26903	-0,4278	-0,1088	31,9172	-15,08	8,81702	1,04663	-0,1577	6,66594
13	11	102	4,79476	-0,4129	0,01284	32,2467	6,18289	-1,1063	-0,0503	31,5938	-15,056	5,24106	0,45437	-0,1037	4,52304
14	9	96	-4,0141	0,4371	-0,0171	31,8752	-2,7637	-0,2663	-0,0779	31,1946	-15,646	-3,8775	1,39265	-0,1355	-3,7309
15	18	93	3,622	-0,0232	-0,0104	32,3025	5,51226	-0,6102	-0,0828	31,6503	-15,162	4,95207	0,88336	-0,1324	3,41196
16	8	104	3,41477	-0,4107	0,01725	32,1492	4,61621	-1,1274	-0,0433	31,4924	-15,125	3,47428	0,49642	-0,1002	3,14207
17	11	100	2,67461	-0,2303	0,00716	32,1557	4,05882	-0,9218	-0,0558	31,4966	-15,155	3,10798	0,64947	-0,1096	2,52304
18	15	95	2,13322	-0,0865	-0,001	32,1771	3,78598	-0,7262	-0,0686	31,5191	-15,21	3,06317	0,79782	-0,1196	2,03099
19	10	100	1,48962	-0,1452	0,00517	32,0959	2,8093	-0,8451	-0,0568	31,4331	-15,255	1,7905	0,75345	-0,1119	1,39605
20	10	100	1,48962	-0,1452	0,00517	32,0959	2,8093	-0,8451	-0,0568	31,4331	-15,239	1,7905	0,75192	-0,1118	1,39605
21	20	87	-0,3602	-0,0059	0,0016	32,0874	1,69058	-0,5434	-0,0741	31,4322	-15,322	1,23038	0,95469	-0,1238	-0,3341

Datos de Mickey, Dunn, y Clark (1967)

Tabla-2

CÁLCULO DE DFBeta PARA TODAS LAS POSIBLES PAREJAS CON EL PRIMER REGISTRO FIJO

			Total=	3912,67			
N=	21	B=	-1,12699	Model=	1604,08		
Ymed=	93,6667	A=	109,874	Error=	2308,59		
Xmed=	14,381			S=	11,0229	V(B)=	0,09621
				R^2=	0,40997	V(A)=	25,6826

N	X	Y	R1	A1-A	B1-B	R1	R(i)	A2-A	B2-B	R1	R2	R(i)	A3-A	B3-B	E(i)
1	26	71	-11,3214	-0,95875	0,10416										-9,57213
2	10	83	-16,6498	1,62342	-0,05776	-11,4663	-16,7393	0,66114	0,04742						-15,604
3	10	83	-16,6498	1,62342	-0,05776	-11,4663	-16,7393	0,66114	0,04742	-11,6319	-17,9433	-17,9421	2,51392	-0,01747	-15,604
4	11	84	-14,2866	1,23032	-0,03825	-11,6045	-14,4897	0,2651	0,06797	-11,7797	-17,7302	-15,6056	2,07511	0,00509	-13,477
5	11	86	-12,1664	1,04774	-0,03257	-11,563	-12,3689	0,08596	0,07327	-11,7369	-17,5999	-13,4762	1,88265	0,01085	-11,477
6	7	113	12,1145	-1,59505	0,0708	-11,6183	12,3737	-2,61307	0,1792	-11,725	-15,8787	11,0943	-0,90542	0,11763	11,0151
7	17	121	31,9816	-0,56916	-0,06632	-8,66442	31,3254	-1,29123	0,01475	-8,85271	-15,4644	30,6939	0,21191	-0,03585	30,285
8	42	57	-15,9026	-4,24397	0,34777	-24,598	-37,2049	-12,012	1,03992	-23,7103	-15,2114	-34,158	-9,64058	0,91236	-5,54031
9	15	102	9,48556	-0,38483	-0,00465	-10,7613	8,88296	-1,2717	0,09465	-10,9526	-18,4768	8,10034	0,54542	0,0327	9,03099
10	9	91	-9,39361	1,02288	-0,04002	-11,3005	-9,37069	0,0634	0,06404	-11,4491	-17,5522	-10,618	1,89806	-0,00079	-8,73094
11	12	105	9,12554	-0,68196	0,0172	-11,053	8,82569	-1,59557	0,11833	-11,2231	-16,2676	7,86622	0,04789	0,06165	8,65003
12	20	94	7,18792	0,11762	-0,03198	-10,6106	6,05162	-0,79952	0,07069	-10,8155	-16,5812	5,54379	0,79155	0,01732	6,66594
13	11	102	4,79476	-0,41291	0,01284	-11,2316	4,59813	-1,34712	0,11564	-11,3948	-16,7215	3,55906	0,35895	0,05636	4,52304
14	9	96	-4,01412	0,4371	-0,0171	-11,3125	-3,99117	-0,52339	0,08707	-11,4578	-17,1126	-5,21124	1,2657	0,02385	-3,73094
15	18	93	3,622	-0,02322	-0,01038	-11,0658	2,6715	-0,95422	0,09415	-11,271	-16,6606	2,04696	0,65688	0,04003	3,41196
16	8	104	3,41477	-0,41072	0,01725	-11,368	3,55172	-1,38989	0,12253	-11,4949	-16,7063	2,29538	0,37941	0,0594	3,14207
17	11	100	2,67461	-0,23033	0,00716	-11,2731	2,47725	-1,16799	0,11034	-11,4376	-16,6502	1,42966	0,53175	0,0513	2,52304
18	15	95	2,13322	-0,08655	-0,00105	-11,2266	1,50456	-1,01176	0,10255	-11,4221	-16,7099	0,7044	0,63343	0,04677	2,03099
19	10	100	1,48962	-0,14524	0,00517	-11,3093	1,40135	-1,09436	0,10891	-11,4637	-16,7799	0,28063	0,63795	0,04823	1,39605
20	10	100	1,48962	-0,14524	0,00517	-11,3093	1,40135	-1,09436	0,10891	-11,4637	-16,7206	0,28063	0,63216	0,04844	1,39605
21	20	87	-0,36022	-0,00589	0,0016	-11,5085	-1,59268	-1,00065	0,11296	-11,7164	-16,805	-2,10799	0,61187	0,05887	-0,33406

Datos de Mickey, Dunn, y Clark (1967)

APÉNDICE 1

Dadas las observaciones modificadas (x_i, y_i^*) $i = 1, ..., n$ con

$$y_i^* = \begin{cases} y_i + \gamma_i & \text{para } 1 \leq i \leq k \\ y_i & \text{para } k < i \leq n \end{cases}$$

los nuevos estimadores mínimos cuadrados $\hat{\alpha}^*$ y $\hat{\beta}$ del modelo $Y = \alpha + \beta X + e$ satisfacen:

$$\hat{\beta}^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\alpha}^* = \bar{y}^* + \hat{\beta}^* \bar{x}$$

como

$$\bar{y}^* = \bar{y} + \frac{1}{n} \sum_{i=1}^{k} \gamma_i \quad \text{y} \quad \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

se deduce que:

$$\hat{\beta}^* = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(y_i + \gamma_i - \bar{y} - \frac{1}{n} \sum_{i=1}^{k} \gamma_i \right) + \sum_{i=k+1}^{n} (x_i - \bar{x}) \left(y_i - \bar{y} - \frac{1}{n} \sum_{i=1}^{k} \gamma_i \right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \frac{1}{n} \sum_{i=1}^{k} \gamma_i \sum_{i=1}^{n} (x_i - \bar{x}) + \sum_{i=1}^{k} \gamma_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{\sum_{i=1}^{k} \gamma_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \beta + \frac{\sum_{i=1}^{k} \gamma_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

y reemplazando $\hat{\beta}^*$ se obtiene:

$$\hat{\alpha}^* = \left(\bar{y} + \frac{1}{n}\sum_{i=1}^{k}\gamma_i\right) - \left(\hat{\beta} + \frac{\sum_{i=1}^{k}\gamma_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\bar{x}$$

$$= \hat{\alpha} + \frac{1}{n}\sum_{i=1}^{k}\gamma_i - \frac{\sum_{i=1}^{k}\gamma_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\bar{x}$$

APÉNDICE 2

Los nuevos residuales $\hat{e}^* = y_i^* - \hat{y}_i^*$ con $\hat{y}_i^* = \hat{\alpha}^* + \hat{\beta}^* x_i$, satisfacen:

$$\hat{e}_i^* = \begin{cases} (y_i + \gamma_i) - \left(\hat{\alpha}^* + \hat{\beta}^* x_i\right) & \text{si} \quad 1 \leq i \leq k \\ y_i - \left(\hat{\alpha}^* + \hat{\beta}^* x_i\right) & \text{si} \quad k < i \leq n \end{cases}$$

Se deduce que para $1 \leq i \leq k$

$$\hat{e}_i^* = (y_i + \gamma_i) - \left(\hat{\alpha} + \frac{1}{n}\sum_{i=1}^{k}\gamma_i - \frac{\sum_{i=1}^{k}\gamma_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\bar{x}\right) - \left(\hat{\beta} + \frac{\sum_{i=1}^{k}\gamma_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)x_i$$

en consecuencia

$$\dot{e}_i^* - \dot{e}_i = \gamma_i - \frac{1}{n}\sum_{i=1}^{k}\gamma_i - \frac{\sum_{i=1}^{k}\gamma_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i-\bar{x})$$

$$= \gamma_i - \frac{1}{n}\left(\gamma_i + \sum_{j\neq i}^{k}\gamma_j\right) - \frac{\gamma_i(x_i-\bar{x}) - \sum_{j\neq i}^{k}\gamma_j(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i-\bar{x})$$

$$= \gamma_i\left(1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) - \frac{1}{n}\sum_{j\neq i}^{k}\gamma_j - \frac{\sum_{j\neq i}^{k}\gamma_j(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i-\bar{x})$$

como para cada $i \neq j$

$$\frac{\gamma_j}{n} + \frac{\gamma_j(x_j-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \gamma_j\left(\frac{1}{n} - \frac{(x_j-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) = \gamma_j h_{ij}$$

se obtiene finalmente

$$\epsilon_i^* - \epsilon_i = \gamma_i(1-h_{ii}) - \sum_{j\neq i}^{k}\gamma_j h_{ij} \qquad \text{con } 1 \leq i \leq k$$

Los demás residuales $\hat{e}_j^*$ para cada j con $k < j \leq n$ satisfacen:

$$\begin{aligned} \hat{e}_j^* &= y_j - \left(\hat{\alpha}^* + \hat{\beta}^* x_j\right) \\ &= \hat{e}_j - \frac{1}{n}\sum_{i=1}^{k}\gamma_i - \frac{\sum_{i=1}^{k}\gamma_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}(x_j - \bar{x}) \\ &= \hat{e}_j - \sum_{i=1}^{k}\gamma_i h_{ij} \end{aligned}$$

REFERENCIAS

Beckman, R.J. and Cook R.D. (1983).Outliers.Thechnometric, Vol 25, pag 119.

Belsley, D. et all (1980). Regression Diagnostics. Identifying influencial data and sourses of colinearity. John Wiley.

Draper, N.R. and John, J.A. (1981). Influential observation and outliers in regression. Technometrics, Vol 23, pag 21

Marasinghe, G (1985). A multistage procedure for detecting several outliers in linear regression. Technometrics, Vol 27, pag 395.

Paul, S.R. and Fung, K.Y. (1991). A generalized extreme studentized residual. Multiple outliers detection procedure in linear regression. Technometrics, Vol 33 pag 339.

Tukey, J. (1977) Exploratory Data analysis. Addison Wesley.