Comunicaciones en Estadística
Junio 2009, Vol. 2, No. 1

Un criterio que compara las estadísticas Q_i y $DF\beta_j(i)$ para el análisis de residuales en modelos de rango completo

Some criterium for comparing the Q_i and $DF\beta_j(i)$ statistics for the analysis of residuals in complete range models

Luis Francisco Rincón Suárez[a]
franciscorincon@usantotomas.edu.co

Resumen

En este artículo se presenta un criterio para comparar las estadísticas Q_i y $DF\beta_j(i)$ comunmente usadas en el análisis de residuales para identificar observaciones influyentes en la estimación de modelos de rango completo. El criterio se construye usando la distribución de las estadísticas mencionadas bajo el supuesto clásico $e_i \sim N(0;\sigma^2)$ y $Cov(e_i, e_j) = 0$ si $i \neq j$).

***Palabras clave*:** Modelo de rango completo,Suma de cuadrados residual, parámetros estimados, observaciones influyentes en la SCE, observaciones influyentes en la estimación de los parámetros..

Abstract

This paper presents, a criterion to compare the Q_i and $DF\beta_j(i)$ statistics, commonly used in the analysis of residuals to identify influential observations on the estimates of full rank linear models. The criterion uses the distribution of these statistics, under the classical assumption $e_i \sim N(0;\sigma^2)$ and $Cov(e_i, e_j) = 0$, $i \neq j$.

***Key words*:** Full rank models, residual sum squares, estimated parameters, influential observations in the RSS, influential observations in the parameters estimation.

1. La estadística Q_i

Cuando el modelo $Y = \beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_x+e$ con k variables regresoras escrito en forma matricial

$$Y = X\beta + e$$

[a]Docente, Facultad de estadística. Universidad Santo Tomás

se transforma a la forma

$$Y^* = X\beta^* + e^*$$

donde

$$Y^* = Y + R \quad \text{para} \quad R' = (r_1, r_2, \cdots, r_n)$$

un vector de constantes de dimensión $n \times 1$

Resultan ciertos los siguientes resultados en la estimación de los dos modelos:

- $\hat{\beta}^* = \hat{\beta} + (X'X)^{-1}X'R$ para $\hat{\beta}$ el vector de dimensión $p \times 1$ de parámetros estimados por el método de mínimos cuadrados.
- $\hat{Y}^* = \hat{Y} + X(X'X)^{-1}X'R$ con $\hat{Y} = X\hat{\beta}$.
- $e^* = e + (I - H)R$ con $e = Y - \hat{Y}$.
- $SCE^* = SCE + R'(I - H)(2Y + R)$ para $SCE = Y'(I - H)Y$.

Particularmente la ecuación correspondiente al i-ésimo residual satisface

$$e_i^* = e_i + (1 - h_{ii})r_i \tag{1}$$

y en este caso $e_i^* = 0$ cuando $r_i = \dfrac{-e_i}{(1 - h_{ii}}$ siendo e_i el i-ésimo residual obtenido al ajustar el modelo $Y = X\beta + e$

Si el vector de alteraciones $R = r_i\vec{e}_i$, para $\vec{e}_i$ el i-ésimo vector canónico de dimensión n, el modelo $Y = X\beta + e$ sólo se altera en el i-ésimo valor de Y y además

$$SCE^* = SCE - \frac{e_i^2}{(1 - h_{ii})} \tag{2}$$

Este resultado permite definir la estadística $Q(i)$ que evaluada para el i-ésimo registro, mide el cambio en la SCE cuando el modelo $Y = X\beta + e$ se ajusta después de eliminar este registro. Dicha estadística se calcula con la expresión

$$Q(i) = \frac{e_i^2}{1 - h_{ii}} = SCE - SCE(i) \tag{3}$$

donde SCE es la suma de cuadrados residual cuando el modelo se ajusta con los n registros y $SCE(i)$ es la suma de cuadrados residual cuando el modelo se ajusta sin el i-ésimo registro.

1.1. Criterio para detectar observaciones influyentes

Con el supuesto clásico $e_i \sim N(0;\sigma^2)$ y $Cov(e_i, e_j) = 0$ para $i \neq j$ se deduce s^2 el estimador insesgado de σ^2 la varianza residual del modelo original:

$$\frac{e_i}{\sigma} \sim N(0,1) \quad \text{y} \quad \frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

De lo anterior resulta

$$\frac{e_i}{s} \sim T_{n-p} \tag{4}$$

y

$$\frac{\sqrt{Q(i)(1-h_{ii})}}{s} = \frac{e_i}{s} = T_i \sim T_{(n-p)} \tag{5}$$

distribución que permite establecer el siguiente criterio para la clasificación de observaciones influyentes.

La i-ésima observación es influyente para la SCE al ajustar el modelo $Y = X\beta + e$ si $|T_i| \geq t_{\alpha/2}$

2. Estadística $DFBeta(i)$

Del resultado $\hat{Y}^* = \hat{Y} + X(X'X)^{-1}X'R$ con $\hat{Y} = X\hat{\beta}$ y $r_i = \dfrac{-e_i}{(1-h_{ii}}$ esta estadística mide el cambio en la estimación del vector de parámetros β al ajustar el modelo $Y = X\beta + e$ después de eliminar el i-ésimo registro. Se calcula con la expresión

$$DFBeta(i) = \hat{\beta} - \hat{\beta}(i) = \frac{(X'X)^{-1}X_i'e_i}{1-h_{ii}} \tag{6}$$

donde $\hat{\beta}$ es el vector de parámetros estimados con la totalidad de los registros, $\hat{\beta}(i)$ es el vector de parámetros estimados después de eliminar el i-ésimo registro y X_i la i-ésima fila de la matriz X.

2.1. El criterio de comparación

Con el supuesto $e_i \sim N(0;\sigma^2)$ y $Cov(e_i, e_j) = 0$ para $i \neq j$ se deduce:

$$\frac{(X'X)^{-1}X_i'e_i}{1-h_{ii}} \sim N(0;\sigma^2\frac{((X'X)^{-1}X_i')(X_i(X'X)^{-1})}{(1-h_{ii})^2}) \tag{7}$$

Para cada j, $j = 1, 2, \cdots, p$ se define m_j el j-ésimo elemento en la diagonal de la matriz

$$M = \frac{(X'X)^{-1}X_i')(X_i(X'X)^{-1})}{(1-h_{ii})^2}$$

entonces para cada j

$$DF\beta_j(i) = \hat{\beta}_j^*(i) - \hat{\beta}_j(i) \sim N(0; m_j\sigma^2) \tag{8}$$

satisface

$$\frac{DF\beta_j(i)}{s\sqrt{m_j}} \sim T_{n-p} \tag{9}$$

y denotando por

$$\frac{DFBeta_j(i)}{s\sqrt{m_j}} = T_j(i) \sim T_{(n-p)}$$

finalmente se obtiene

$$|T_j(i)| = |T_i| = \frac{|e_i|}{s} \text{ para todo } j \text{ y todo } i \tag{10}$$

El último resultado relaciona las distribuciones de las estadísticas $Q(i)$ y $DFBeta_j(i)$ y permite concluir que **un registro es influyente en la estimación de los parámetros sí y sólo sí es influyente en la suma de cuadrados residual**.

3. ejemplo

Se ilustra el desarrollo teórico expuesto en las secciones anteriores con el siguiente script que se ejecuta en R

```
              # Datos para el modelo

  Y<-matrix(c(10.98,11.13,12.51,8.40,9.27,8.73,6.36,8.5,7.82,9.14,
        8.24,8.19,11.88,9.57,10.94,9.58,10.09,8.11,6.83,8.88,7.68,8.47,
        8.86,10.36,11.08),ncol=1);

  X1<-matrix(c(35.3,29.7,30.8,58.8,61.4,71.3,74.4,76.7,70.7,57.5,46.4,28.9,
       28.1,39.1,46.8,48.5,59.3,70,70,74.5,72.1,58.1,44.6,33.4,28.6),ncol=1);

  X2<-matrix(c(5.8,3.4,4.6,6.7,3.2,4.3,5.6,7.8,3.4,4.5,6.3,4.8,3.1,2.3,4.7,5.7,
           4.8,7.2,6.3,4.3,5.6,4.7,4.9,6.3,2.7),ncol=1);

   N<-nrow(X1);
  J<-matrix(1,nrow=N,ncol=1); X<-matrix(c(J,X1,X2),ncol=3);
   P=ncol(X);


 # Estimación del modelo con intercepto
   B<-(solve(t(X)%*%X))%*%t(X)%*%Y;
   H<- (X%*%(solve(t(X)%*%X)))%*%t(X);
   I<-diag(N);
```

```
  e<- (I-H)%*%Y;
  SCE<- t(e)%*%e;
  s2<-SCE[1,1]/(N-P);
  s<-sqrt(s2);

# Análisis de residuales

# Cálculo de las Estadísticas Qi y su criterio Ti
  hii<-J-diag(H);
  ri<- e/hii;
  Qi<- e^2/hii;
  MaxQi<- max(Qi) ;Ti<-abs(e/s);
  MaxTi<- max(Ti);
  t95<-qt(0.95,N-P);

# Cálculo de la Estadística DFbeta
  M<- t((solve(t(X)%*%X))%*%t(X));
  DFbeta0<- (M[,1])*ri ;
  DFbeta1<- (M[,2])*ri ;
  DFbeta2<- (M[,3])*ri ;


# Verificacion de la igualdad de las estadísticas Ti y Tj(i)

# Calculo de la estadística Tj(i) para cada parametro
   hii2<-hii^2 ;
  MG<-((solve(t(X)%*%X))%*%t(X))^2;

# T para Beta0  T0
  Kbeta0<-MG[1,];
  m0<-Kbeta0/hii2;
  K0<-s*sqrt(m0);
  T0<-DFbeta0/K0;

# T para Beta1  T1
  Kbeta1<-MG[2,];
  m1<-Kbeta1/hii2;
  K1<-s*sqrt(m1);
  T1<-DFbeta1/K1;

# T para Beta2  T2
  Kbeta2<-MG[3,];
  m2<-Kbeta2/hii2;
  K2<-s*sqrt(m2);
  T2<-DFbeta2/K2;
```

```
# Resúmen de las estadísticas Qi t DFbeta
  EST<-matrix(c(Y,Qi,DFbeta0,DFbeta1,DFbeta2,Ti),ncol=6) ;
  colnames(EST)<-c("Y","Qi","DFbeta0","DFbeta1","DFbeta2","Ti");
# Comparación de la estadística T
  Tbeta<-matrix(c(Ti,T0,T1,T2),ncol=4);
 colnames(Tbeta)<-c("Ti","T0","T1","T2");
  TG<-abs(Tbeta);

# Decisión sobre la existencia de observaciones influyentes
  Dec<- matrix(c(MaxQi,MaxTi,t95),ncol=3);

# Análisis final.Cálculo de los nuevos parámetros al ajustar el modelo
# sin le registro i-ésimo
  NSCE<-J%*%SCE-Qi;
  VporSCE<-(Qi/SCE[1,1])*100;
  NB0<-J%*%B[1,1]-DFbeta0;
  NB1<-J%*%B[2,1]-DFbeta1;
  NB2<-J%*%B[3,1]-DFbeta2;

 NP<-matrix(c(Qi,Y,NSCE,VporSCE,NB0,NB1,NB2),ncol=7);
colnames(NP)<-c("Qi","Y","NSCE","VporSCE","NB0","NB1","NB2");

 CRIT<-matrix(c(MaxQi,MaxTi,t95),ncol=3);
colnames(CRIT)<-c("MaxQi","MaxTi","t95");
 round(Tbeta,4) ;round(CRIT,4); round(NP,4) ;
```

y del cual se obtienen los siguientes resultados:

La siguiente tabla que contiene para cada observación el valor de las estadísticas T_i y $T_j(i)$ que permite verificar $|T_j(i)| = |T_i|$ para todo i y todo j

	Ti	T0	T1	T2
[1,]	0.7569	0.7569	-0.7569	0.7569
[2,]	0.2026	0.2026	-0.2026	-0.2026
[3,]	1.7660	1.7660	-1.7660	1.7660
[4,]	0.1860	0.1860	0.1860	-0.1860
[5,]	0.2562	0.2562	0.2562	-0.2562
[6,]	0.4908	-0.4908	0.4908	-0.4908
[7,]	1.3848	1.3848	-1.3848	1.3848
[8,]	1.1265	-1.1265	1.1265	1.1265
[9,]	0.5495	-0.5495	-0.5495	0.5495
[10,]	0.1041	0.1041	0.1041	-0.1041
[11,]	1.1259	1.1259	1.1259	-1.1259
[12,]	2.4288	-2.4288	2.4288	-2.4288
[13,]	0.7771	0.7771	-0.7771	-0.7771
[14,]	0.9021	-0.9021	0.9021	0.9021

[15,]	1.2228	1.2228	-1.2228	-1.2228
[16,]	0.1809	-0.1809	-0.1809	0.1809
[17,]	1.1605	1.1605	1.1605	-1.1605
[18,]	0.2704	-0.2704	0.2704	0.2704
[19,]	1.0871	1.0871	-1.0871	-1.0871
[20,]	0.8210	-0.8210	0.8210	-0.8210
[21,]	0.2623	0.2623	-0.2623	-0.2623
[22,]	0.4681	-0.4681	-0.4681	0.4681
[23,]	0.8562	-0.8562	0.8562	-0.8562
[24,]	0.1321	0.1321	-0.1321	0.1321
[25,]	0.0171	-0.0171	0.0171	0.0171

El valor máximo de la estadística Q_i, el valor T_i y el valor $t_{95\,\%}$ que permite decidir que el registro número 12 es influyente para la SCE.

MaxQi	MaxTi	t95
7.4591	2.4288	1.7171

Finalmente la tabla que contiene para cada registro:

- El valor de $Q_i = SCE - SCE(i)$.
- El valor de SCE =NSCE la nueva suma de cuadrados residual cuando el modelo se ajusta sin el i-ésimo registro.
- La variación porcentual VporSCE de la suma de cuadrados residual cuando el modelo se ajusta sin el i-ésimo registro.
- El valor de los nuevos parámetros estimados $\hat{\beta}_0(i)$ =NB0, $\hat{\beta}_1(i)$ =NB1 y $\hat{\beta}_2(i)$ =NB2 cuando el modelo se ajusta sin el i-ésimo registro.

	Qi	Y	NSCE	VporSCE	NB0	NB1	NB2
[1,]	0.7268	10.98	23.5275	2.9966	13.2421	-0.0583	-0.1931
[2,]	0.0519	11.13	24.2024	0.2141	13.2261	-0.0608	-0.1569
[3,]	3.8619	12.51	20.3925	15.9224	12.9356	-0.0545	-0.1801
[4,]	0.0427	8.40	24.2117	0.1759	13.2638	-0.0615	-0.1532
[5,]	0.0849	9.27	24.1694	0.3500	13.2505	-0.0623	-0.1463
[6,]	0.3033	8.73	23.9511	1.2504	13.2936	-0.0636	-0.1443
[7,]	2.3661	6.36	21.8882	9.7555	13.0977	-0.0565	-0.1619
[8,]	1.8158	8.50	22.4385	7.4864	13.7588	-0.0642	-0.2401
[9,]	0.4117	7.82	23.8426	1.6976	13.3433	-0.0585	-0.1976
[10,]	0.0126	9.14	24.2417	0.0520	13.2866	-0.0616	-0.1600
[11,]	1.5621	8.24	22.6922	6.4406	13.2415	-0.0642	-0.1113

```
[12,] 7.4591   8.19  16.7952  30.7537  13.7741  -0.0726  -0.1171
[13,] 0.7835  11.88  23.4709   3.2302  13.0019  -0.0589  -0.1373
[14,] 1.0963   9.57  23.1580   4.5200  13.6532  -0.0617  -0.2229
[15,] 1.7257  10.94  22.5287   7.1148  13.1773  -0.0604  -0.1604
[16,] 0.0385   9.58  24.2158   0.1587  13.2926  -0.0612  -0.1660
[17,] 1.5611  10.09  22.6933   6.4362  13.2707  -0.0630  -0.1513
[18,] 0.0954   8.11  24.1589   0.3933  13.3675  -0.0618  -0.1756
[19,] 1.4460   6.83  22.8083   5.9618  13.0981  -0.0592  -0.1362
[20,] 0.8729   8.88  23.3815   3.5988  13.3102  -0.0656  -0.1295
[21,] 0.0837   7.68  24.1707   0.3450  13.2594  -0.0607  -0.1611
[22,] 0.2537   8.47  24.0006   1.0460  13.3081  -0.0609  -0.1665
[23,] 0.8510   8.86  23.4034   3.5085  13.3685  -0.0627  -0.1562
[24,] 0.0237  10.36  24.2307   0.0976  13.2880  -0.0608  -0.1698
[25,] 0.0004  11.08  24.2539   0.0016  13.2995  -0.0615  -0.1623
```

Referencias

[1] Draper, N. R, and John, J. A, *Influential observations and outliers in regression*, Technometrics, Vol 22, 1980.

[2] Jiménez, José Alfredo y Rincón, Luis Francisco, *Una generalización de la estadística $DFBeta$*, Revista Colombiana de Estadística, Universidad Nacional de Colombia, Vol 23, N1, 2000.

[3] Leon C, Carmen Elena, *Caracterización de observaciones influyentes en la estimación de los parámetros, en Modelos de Regresión Lineal Múltiple*, Trabajo de Grado, Estadística, Universidad Nacional de Colombia, 2002.

[4] Morales, R. Mario Alfonso, *Estudio de algunas consecuencias derivadas de eliminar una observación influyente en modelos de regresión lineal múltiple*, Trabajo de Grado, Especialización en Estadística, Universidad Nacional de Colombia. 2000.

[5] Rincón, Ángela Tatiana, *Propuesta para caracterizar observaciones influyentes en Modelos de Regresión Lineal Múltiple*, Trabajo de Grado, Estadística, Universidad Nacional de Colombia, 1999.

[6] Romero R. Juan de Jesús, *Identificación de observaciones influyentes en un modelo de diseño a través de variables Dummy*, Trabajo de Grado, Especialización en Estadística, Universidad Nacional de Colombia. 2002.

[7] Searle, S. R, *Linear Models*, John Wiley & Sons, 1971.