UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2013, Vol. 6, No. 1

Cómo observar el impacto del i-ésimo registro sobre el coeficiente de determinación al ajustar un modelo de regresión lineal múltiple

How to observe the influence of i-th record upon the coefficient of determination in multiple regression models

Luis Alejandro Fernández[a]
luisfc91@yahoo.com

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

Resumen

En este artículo se expone un procedimiento para calcular el coeficiente de determinación R_i^2, del modelo de regresión lineal múltiple $Y = X\beta + e$ ajustado después de eliminar el i-ésimo registro. El procedimiento permite observar el comportamiento del coeficiente de determinación, cuando el registro eliminado es influyente para la suma de cuadrados residual SCE según la estadística Q_i. Se incluye la sintaxis para realizar los cálculos en R.

***Palabras clave*:** modelo lineal de rango completo, suma de cuadrados residual, observaciones influyentes en la SCE, coeficiente de determinación R^2.

Abstract

This article exposes a procedure to calculate the coefficient of determination R_i^2, of a multiple linear regression model $Y = X\beta + e$ adjusted after the elimination of a ith record of the data. The procedure allows to observe the behavior of the coefficient of determination when the eliminated record, influences the sum of squared error of the residual SSE by the statistic Q_i. It is included in this article the program in R to make the calculations.

***Key words*:** linear model of full range, sum of squared error, influent observations in the SSE, coefficient of determination R^2.

1. Justificación

En Rincón (2009*b*) se expone la metodología para calcular la estadística Q_i que, evaluada para el i-ésimo registro, mide el cambio en la suma de cuadrados residual

[a]Estadístico, Facultad de Estadística. Universidad Santo Tomás.
[b]Docente, Facultad de Estadística. Universidad Santo Tomás.

SCE cuando el modelo de rango completo $Y = X\beta + e$ se ajusta después de eliminar este registro. Dicha estadística se calcula con la expresión (Rincón 2009a)

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

donde $h_{ii} = X_i(X'X)^{-1}X_i'$, 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.

Los resultados logrados en este trabajo son muy útiles para el análisis de residuales en el ajuste del modelo, ya que permiten observar simultáneamente el impacto que genera eliminar una observación, sobre la suma de cuadrados residual SCE, y sobre el coeficiente de determinación R^2.

2. Marco teórico

En esta sección se expone el marco teórico utilizado para calcular el estadístico R_i^2, cuando el modelo se ajusta con o sin intercepto.

2.1. Modelo sin intercepto

En el modelo sin intercepto el estadístico R_i^2 está dado por

$$R_i^2 = \frac{SCR_i}{SCT_i} = \frac{SCT_i - SCE_i}{SCT_i} = 1 - \frac{SCE_i}{SCT_i}$$

donde

- SCT_i: es la suma de cuadrados total del modelo después de eliminar el i-ésimo registro y en el modelo sin intercepto está dada por la expresión

  $$SCT_i = Y'Y - y_i^2$$

  Si denotamos por $SCT_{(i)}$ el vector de valores SCT_i para $i = 1, 2, \cdots n$, usamos la siguiente expresión para calcularlo en R

  $$SCT_{(i)} = 1_n Y'Y - Diag(YY') \tag{2}$$

  con 1_n un vector columna de dimensión n con valores iguales a 1 y $Diag(YY')$ la matriz diagonal con valores y_i^2.

- SCE_i: es la suma de cuadrados residual del modelo calculada después de eliminar el i-ésimo registro y se calcula con la expresión

  $$\begin{aligned} SCE_i &= SCE - \frac{e_i^2}{1 - h_{ii}} \\ &= SCE - Q_i \end{aligned}$$

- SCR_i: es la suma de cuadrados de regresión calculada después de eliminar el i-ésimo registro y está dada por la expresión

$$\begin{aligned} SCR_i &= SCT_i - SCE_i \\ &= Y'Y - y_i^2 - (SCE - Q_i) \\ &= Y'Y - y_i^2 - SCE + Q_i \end{aligned}$$

2.2. Modelo con intercepto

Para el modelo con intercepto el estadístico R_i^2 está dado por

$$R_i^2 = \frac{SCR_{mi}}{SCT_{mi}} \tag{3}$$

Donde

- SCT_m es la suma de cuadrados total ajustada por la media, puesto que el modelo tiene intercepto y SCT_{mi} es la suma de cuadrados del total ajustado por la media calculada despúes de eliminar el i-ésimo registro. En el análisis de varianza o tabla de ANOVA del modelo con intercepto

$$\begin{aligned} SCT_m &= Y'Y - n\bar{Y}^2 \\ &= Y'Y - n\left(\frac{\sum_{i=1}^n y_i}{n}\right)^2 \\ &= Y'Y - \frac{1}{n}\left(\sum_{i=1}^n y_i\right)^2 \end{aligned}$$

entonces

$$SCT_{mi} = Y'Y - y_i^2 - \frac{1}{(n-1)}\left[\left(\sum_{i=1}^n y_i\right) - y_i\right]^2$$

de donde

$$\begin{aligned} SCT_{mi} &= Y'Y - y_i^2 - \frac{1}{(n-1)}\left[\left(\sum_{i=1}^n y_i\right)^2 - 2y_i\sum_{i=1}^n y_i + y_i^2\right] \\ &= Y'Y - y_i^2 - \frac{1}{(n-1)}(n^2\bar{Y}^2 - 2y_i n\bar{Y} + y_i^2) \\ &= Y'Y + \frac{1}{(n-1)}(2y_i n\bar{Y} - n^2\bar{Y}^2 - ny_i^2) \end{aligned}$$

Denotamos por $SCT_{m(i)}$ el vector de valores SCT_{mi} para $i = 1, 2, \cdots, n$ y una expresión para calcularlo en R está dada por

$$SCT_{m(i)} = 1_n Y'Y + (Y 2n\bar{Y} - 1_n n^2 \bar{Y}^2 - Diag(YY')n)\frac{1}{(n-1)}$$

- SCR_{mi} es la suma de cuadrados de regresión ajustada por la media y SCR_{mi} es la suma de cuadrados de regresión ajustada por la media y calculada despúes de eliminar el i-ésimo registro. Del análisis de la ANOVA

$$SCR_m = SCT_m - SCE$$

y

$$SCR_{mi} = SCT_{mi} - SCE_i$$

es decir

$$\begin{aligned} SCR_{mi} &= Y'Y + \frac{1}{(n-1)}(2y_i n\bar{Y} - n^2\bar{Y}^2 - ny_i^2) - (SCE - Q_i) \\ &= Y'Y - SCE + Q_i + \frac{1}{(n-1)}(2y_i n\bar{Y} - n^2\bar{Y}^2 - ny_i^2) \end{aligned}$$

Una expresión que facilita el cálculo del vector $SCR_{m(i)}$, que contiene los valores SCR_{mi} está dada por

$$SCR_{m(i)} = 1_n(Y'Y - SCE) + Q_{(i)} + (Y 2n\bar{Y} - 1_n n^2\bar{Y}^2 - Diag(YY')n)\frac{1}{(n-1)}$$

3. Sintaxis en R

3.1. Ejemplo

Para ilustrar el cálculo de la estadística $R^2_{(i)}$ en R se toman los siguientes datos de un curso de pregrado en la Universidad Santo Tomás. Los datos contienen para 32 estudiantes las variables:

- Pe: el peso del estudiante medido en kilogramos.
- Ta: la talla del estudiante medida en centímetros.
- Co: el contorno del estudiante medido en centímetros.

	Peso	Talla	Contorno
1	75,0	171	93
2	78,0	169	93
3	56,0	161	82
4	55,0	155	72
5	63,0	159	86
6	45,0	153	65
7	55,0	162	85
8	60,0	172	82
9	80,3	168	101
10	75,0	176	92
11	67,0	170	83
12	56,0	170	75
13	80,0	178	93
14	54,0	153	74
15	72,0	170	86
16	59,0	165	76
17	86,0	174	94
18	61,0	180	71
19	65,0	160	98
20	82,0	188	96
21	70,0	180	78
22	70,0	155	91
23	58,0	162	76
24	63,0	175	83
25	72,0	167	98
26	92,0	183	99
27	72,0	180	93
28	75,0	175	95
29	55,0	168	68
30	75,0	173	92
31	74,0	173	97
32	55,0	155	72

3.2. Modelo sin intercepto

Para los anteriores datos contenidos en el archivo CURSO2.cvs, en el modelo sin intercepto $Pe = \beta_1 Ta + \beta_2 Co + e$ la siguiente es la sintaxis en R para calcular las estadísticas listadas a continuación.

- $Q_i = SCE - SCE_{(i)}$
- Pvalor.
- $SCR_{(i)}$.
- $SCE_{(i)}$.

- $SCT_{(i)}$.
- $R^2_{(i)}$.
- Var $\%R^2$.

```
rm(list=ls(all=TRUE))

   Base=read.csv("CURSO2.csv")
    attach(Base)
    x=cbind(Pe,Ta,Co)

      A=lm(Pe~Ta+Co-1)
     n=nrow(x)
     p=ncol(x)-1
     Y=matrix(x[,1],ncol=1)
     J=matrix(1,nrow=n,ncol=1)
     X=matrix(J)
   for(i in 2:(p+1)){
   X=cbind(X,x[,i])
        }
    X=X[,-1]
   H=X%*%solve(t(X)%*%X)%*%t(X)
    E=A$res
    SCE=t(E)%*%E
    SCT=t(Y)%*%Y
    SCR=SCT-SCE
    S=sqrt(anova(A)[3,3])

     # Qi
      hii=J-diag(H)
     Q=(E^2)/hii

     NSCE=(J%*%SCE)-Q

       ########################################

         SCTi= J%*%t(Y)%*%Y - Y*Y
         SCEi= J%*%SCE - Q
         SCRi= SCTi - SCEi
         R2i= SCRi/SCTi
          R2=SCR/SCT

        Rcuadrado=cbind(Y,Q,PvalT,SCRi,SCEi,SCTi,R2i,
       ((R2i - J%*%R2)/J%*%R2)*100)

       colnames(Rcuadrado)=c("Obs","Q","SCR(i)",
```

```
                "SCE(i)","SCT(i)",
                 "R2(i)","Var.% R2")
```

3.2.1. Salidas Modelo sin intercepto

```
 N    Q(i)       SCR(i)     SCE(i)    SCT(i)    R2(i)  Var.% R2
 1   4.188    142250.997  1099.093  143350.09  0.992  -0.026
 2  27.028    141814.837  1076.253  142891.09  0.992  -0.013
 3  78.000    144813.809  1025.281  145839.09  0.993   0.038
 4   5.360    144852.169  1097.921  145950.09  0.992  -0.012
 5  21.399    143924.208  1081.882  145006.09  0.993  -0.006
 6  55.566    145902.375  1047.715  146950.09  0.993   0.028
 7 146.524    144993.333   956.757  145950.09  0.993   0.086
 8  26.839    144298.648  1076.442  145375.09  0.993   0.000
 9   2.937    141426.656  1100.344  142527.00  0.992  -0.032
10   6.770    142253.596  1096.494  143350.09  0.992  -0.025
11   1.680    143384.489  1101.601  144486.09  0.992  -0.022
12  16.797    144752.606  1086.484  145839.09  0.993  -0.004
13  47.568    141519.377  1055.713  142575.09  0.993   0.000
14  22.277    144978.086  1081.004  146059.09  0.993   0.000
15  17.362    142705.171  1085.919  143791.09  0.992  -0.015
16   2.349    144393.159  1100.931  145494.09  0.992  -0.016
17 155.294    140631.104   947.986  141579.09  0.993   0.072
18  16.294    144167.103  1086.987  145254.09  0.993  -0.008
19 146.506    143793.315   956.775  144750.09  0.993   0.080
20  40.816    141188.625  1062.465  142251.09  0.993  -0.006
21  63.688    143035.497  1039.593  144075.09  0.993   0.019
22   1.147    142972.956  1102.134  144075.09  0.992  -0.025
23   5.985    144513.794  1097.296  145611.09  0.992  -0.013
24   8.768    143911.577  1094.513  145006.09  0.992  -0.014
25  22.334    142710.143  1080.947  143791.09  0.992  -0.011
26 214.164    139621.973   889.117  140511.09  0.994   0.109
27   1.753    142689.562  1101.528  143791.09  0.992  -0.026
28   0.183    142246.992  1103.098  143350.09  0.992  -0.029
29   0.048    144846.857  1103.233  145950.09  0.992  -0.015
30   7.345    142254.154  1095.936  143350.09  0.992  -0.024
31   4.129    142399.939  1099.151  143499.09  0.992  -0.026
32   5.360     144852.169  1097.921  145950.09  0.992  -0.012
```

3.3. Comentarios

De los valores compilados en la tabla anterior se deduce:

- Para el modelo $Pe = \beta_1 Ta + \beta_2 Co + e$ ajustado con todos los registros, la suma de cuadrados residual $SCE = 1103.2808$ y es el registro 26 el de mayor impacto sobre la SCE, la suma de cuadrados residual se reduce en $Q_i = 214.164$ con una variación porcentual de 19.41 %, si este registro es eliminado.
- El valor del coeficiente R^2, considerando los datos completos es $R^2 = 0.992519$ y el mayor valor del coeficiente $R^2(i)$ se presenta al eliminar el registro de mayor impacto sobre la SCE, es decir, de los datos se deduce que en el ejemplo, al eliminar el registro 26 se presenta el mayor aumento del valor del R^2.
- El coeficiente R^2, puede disminuir cuando se elimina otro registro, como ilustración eliminar el registro 9.
- En general en este modelo a simple vista no se presentan cambios significativos en el R^2, al ajustar el modelo eliminando algún registro.

3.4. Modelo con intercepto

Para los anteriores datos contenidos en el archivo CURSO2.cvs, en el modelo sin intercepto $Pe = \beta_0 + \beta_1 Ta + \beta_2 Co + e$ la siguiente es la sintaxis en R para calcular las estadísticas listadas a continuación.

- $Q_i = SCE - SCE_{(i)}$
- Pvalor.
- $SCR_{(i)}$.
- $SCE_{(i)}$.
- $SCT_{(i)}$.
- $R^2_{(i)}$.
- Var $\%R^2$.

```
rm(list=ls(all=TRUE))

   Base=read.csv("CURSO2.csv")
   attach(Base)
   x=cbind(Pe,Ta,Co)
   A=lm(Pe~Ta+Co)
   n=nrow(x)
   p=ncol(x)
   Y=matrix(x[,1],ncol=1)
   J=matrix(1,nrow=n,ncol=1)
```

```
X=matrix(J)
for(i in 2:p){
X=cbind(X,x[,i])
}
H=X%*%solve(t(X)%*%X)%*%t(X)
E=A$res
SCEm=anova(A)[3,2]
SCTm=t(Y)%*%Y - n*(mean(Y)^2)
SCRm=SCTm-SCEm
S=sqrt(anova(A)[3,3])
# Qi
hii=J-diag(H)
  Q=(E^2)/hii
   ########################################
SCEmi=J%*%SCEm - Q
SCTmi= J%*%t(Y)%*%Y + (2*n*mean(Y)*Y -
J%*%n^2*mean(Y)^2 - n*Y*Y)%*%(1/(n-1))
SCRmi=SCTmi-SCEmi
R2i= SCRmi/SCTmi
R2=SCRm/SCTm
Rcuadrado=cbind(Y,Q,PvalT,SCRmi,SCEmi,
SCTmi,R2i,((R2i - J%*%R2)/J%*%R2)*100)
colnames(Rcuadrado)=c("Obs","Q(i)","SCRm(i)",
"SCEm(i)","SCTm(i)","R2(i)","Var.% R2")
```

3.4.1. Salidas Modelo con intercepto

N	Q(i)	SCRm(i)	SCEm(i)	SCTm(i)	R2(i)	Var.% R2
1	1.501	3136.770	611.769	3748.539	0.837	-0.261
2	27.532	3106.150	585.737	3691.887	0.841	0.281
3	28.291	3090.870	584.978	3675.848	0.841	0.223
4	17.230	3055.338	596.039	3651.377	0.837	-0.265
5	0.037	3176.107	613.232	3789.339	0.838	-0.098
6	0.005	2679.855	613.264	3293.119	0.814	-3.006
7	82.128	3120.236	531.142	3651.377	0.855	1.853
8	40.404	3180.221	572.866	3753.087	0.847	0.998
9	4.630	3027.232	608.639	3635.871	0.833	-0.761
10	0.139	3135.408	613.131	3748.539	0.836	-0.305
11	0.982	3196.483	612.288	3808.771	0.839	0.030
12	18.888	3081.467	594.381	3675.848	0.838	-0.082
13	9.577	3040.104	603.693	3643.797	0.834	-0.556
14	6.325	3017.898	606.944	3624.842	0.833	-0.766
15	14.736	3188.076	598.534	3786.610	0.842	0.351

16	0.159	3123.763	613.111	3736.874	0.836	-0.365
17	107.658	2944.365	505.612	3449.977	0.853	1.723
18	0.878	3154.844	612.392	3767.235	0.837	-0.184
19	66.046	3255.960	547.224	3803.184	0.856	2.041
20	3.844	2978.023	609.425	3587.448	0.830	-1.057
21	10.887	3199.285	602.382	3801.668	0.842	0.305
22	29.509	3217.907	583.761	3801.668	0.846	0.889
23	0.617	3105.944	612.652	3718.597	0.835	-0.446
24	30.002	3206.071	583.267	3789.339	0.846	0.845
25	14.642	3187.982	598.628	3786.610	0.842	0.348
26	78.730	2647.296	534.540	3181.835	0.832	-0.833
27	38.180	3211.520	575.090	3786.610	0.848	1.089
28	4.699	3139.968	608.571	3748.539	0.838	-0.160
29	0.881	3038.989	612.389	3651.377	0.832	-0.799
30	1.071	3136.340	612.198	3748.539	0.837	-0.275
31	14.325	3164.349	598.944	3763.294	0.841	0.221
32	17.230	3055.338	596.039	3651.377	0.837	-0.265

De los valores compilados en la tabla anterior para el modelo con intercepto se deduce:

- Para el modelo $Pe = \beta_0 + \beta_1 Ta + \beta_2 Co + e$ ajustado con todos los registros, la suma de cuadrados residual $SCE = 613.269623$ y es ahora el registro 17 el de mayor impacto sobre la SCE, la suma de cuadrados residual se reduce en $Q_i = 107.6575$ con una variación porcentual de 19.41 %, si este registro es eliminado.
- El valor del coeficiente R^2, considerando los datos completos, es $R^2 = 0.8389903$ y también en este modelo al eliminar el registro de mayor impacto aumenta el valor del R^2, sin embargo a diferencia del modelo anterior, en el modelo con intercepto, la mayor variación positiva del R^2 es 2.041 y se presenta al eliminar el registro 19 que no es el registro de mayor impacto para la SCE.
- El coeficiente R^2 puede disminuir cuando se elimina otro registro, como ilustración eliminar el registro 6 la variación porcentual del R^2 es del -3.006 %.
- En general hay evidencia para pensar que en este modelo se presenta una mayor variación del R^2, que en el modelo sin intercepto.

3.5. Conclusiones

- En general al ajustar un modelo $Y = X\beta + e$ la mayor variación del coeficiente de determinación $R^2_{(i)}$ no corresponde con la eliminación del registro de mayor impacto para la suma de cuadrados residual.

- Se recomienda utilizar los resultados logrados en este trabajo para intentar caracterizar escenarios para el estadístico $R^2_{(i)}$.
- Un trabajo interesante con continuación de estos logros es construir la distribución del estadístico $R^2 - R^2_{(i)}$

Recibido: 27 de febrero de 2013
Aceptado: 22 de marzo de 2013

Referencias

Rincón, L. F. (2009*a*), *Curso Básico de Modelos Lineales*, Universidad Santo Tomás.

Rincón, L. F. (2009*b*), 'Un criterio que compara las estadísticas q_i y $df\beta_j(i)$ para el análisis de residuales en modelos de rango completo', *Comunicaciones en Estadística* (2), 139–146.