UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2011, Vol. 4, No. 2

Un método para determinar observaciones influyentes en la SCE al ajustar modelos de diseños factoriales

A method for Determining Influential Observations in the SSE when Fitting Models in Factorial Designs

Blanca Cecilia Ubaque Lopez[a]
blancaubaque@hotmail.com

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

Resumen

En este artículo se presenta la generalización de la estadística Q_i y el criterio para determinar observaciones influyentes para la SCE al ajustar un modelo lineal $Y = X\beta + e$ de rango incompleto en diseños factoriales. En el análisis de residuales se generaliza la estadística mencionada bajo el supuesto clásico $e_i \sim N(0; \sigma^2)$ y $Cov(e_i, e_j) = 0$ si $i \neq j$).

***Palabras clave*:** diseño de experimentos, modelo de diseño factorial, suma de cuadrados residual, observaciones influyentes en la SCE.

Abstract

This paper presents the generalization of the Q_i statistics and the criterion to identify influential observations for the SCE when a linear model $Y = X\beta + e$ of incomplete rank is fitted using factorial design. Throughout the residual analysis, the later statistics is generalised under the classic assumption $e_i \sim N(0; \sigma^2)$ y $Cov(e_i, e_j) = 0$ si $i \neq j$).

***Key words*:** Experiment Design, Factorial model, residual sum squares, influential observation for the SCE.

1. La estadística Q_i

En la revista *Comunicaciones en Estadística* (Vol. 2, No. 2, página 139), se expone la metodología para calcular la estadística Q_i que evaluada para el i-ésimo

[a]Docente, Colegio Nuestra Señora del Rosario.
[b]Docente, Facultad de Estadística. Universidad Santo Tomás.

registro, mide el cambio en la suma de cuadrados residual 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

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

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.

Para el criterio de selección de observaciones influyentes, se asocia a la estadística $T_i = \frac{\sqrt{Q_i(1-h_{ii})}}{s}$ una distribución $T_{(n-p)}$ 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. Generalización de Q_i

En un modelo de diseño factorial $Y = X\beta + e$ la matriz $(X'X)^{-1}$ no existe y el cálculo de la suma de cuadrados residual SCE se realiza mediante las ecuaciones

- $\hat{\beta} = GX'Y$ para G una inversa generalizada de $X'X$.
- $\hat{Y} = X\hat{\beta} = XGXY'$.
- $e = Y - \hat{Y} = (I - XGX')Y$.
- $SCE = Y'(I - H)Y$, con $H = XGX'$.

Como en los modelos de rango completo se define la estadística

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

donde e_i es el residual calculado para la i-ésima observación y h_{ii} es el i-ésimo valor en la diagonal de la matriz H.

Con el supuesto clásico $e_i \sim N(0; \sigma^2)$ y $Cov(e_i, e_j) = 0$ para $i \neq j$, el criterio para detectar observaciones influyentes resulta

$$\frac{\sqrt{Q(i)(1-h_{ii})}}{s} = \frac{e_i}{s} = T_i \sim T_{(n-r)} \tag{3}$$

con r el rango de la matriz de diseño X. Esta distribución 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 $|T_i| \geq t_{\alpha/2}$

3. Ejemplo

Para los siguientes datos simulados y el modelo $y_{ijk} = \mu + A_i + B_j + C_{ij} + e_{ijk}$

OBS	Y	A	B
1	23.5	a1	b1
2	24.6	a1	b1
3	21.5	a1	b1
4	24.1	a1	b1
5	29.4	a1	b2
6	28.7	a1	b2
7	28.6	a1	b2
8	27.5	a1	b2
9	29.4	a2	b1
10	35.5	a2	b1
11	34.2	a2	b1
12	33.8	a2	b1
13	43.5	a2	b2
14	41.6	a2	b2
15	39.8	a2	b2
16	40.7	a2	b2

Para estos datos, se presenta a continuación la ANOVA del modelo ajustado con todas las observaciones.

Procedimiento GLM

Variable dependiente: Y

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Modelo	4	16726.77500	4181.69375	1394.48	<.0001
Error	12	35.98500	2.99875		
Total no corr	16	16762.76000			

R-cuadrado	Coef Var	Raiz MSE	Y Media
0.951054	5.471374	1.731690	31.65000

Fuente	DF	Tipo I SS	Cuadrado de la media	F-Valor	Pr > F
Término indepe	1	16027.56000	16027.56000	5344.75	<.0001
A	1	513.02250	513.02250	171.08	<.0001
B	1	176.89000	176.89000	58.99	<.0001
A*B	1	9.30250	9.30250	3.10	0.1036

Para los datos, la matriz de diseño X está dada por

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Utilizando R y la sintaxis descrita al finál de este artículo, se presenta la siguiente tabla que contiene para cada observación:

- El valor de la variable respuesta Y.
- El valor de la estadística Q_i.
- El valor de la estadística T
- El p-valor de la estadística T.
- La suma de cuadrados residual $SCE(i)$ resultante de ajustar el modelo sin la i-ésima observación.
- La variación porcentual de la suma de cuadrados residual generada al ajustar el modelo sin la i-ésima observación.

OBS	Y	Qi	Ti	Pval	SCE(i)	Variación
1	23.5	0.00750	0.04331	0.48308	35.9775	-0.02
2	24.6	1.84083	0.67853	0.25516	34.1441	-5.11
3	21.5	4.94083	1.11163	0.14404	31.0441	-13.73
4	24.1	0.6075	0.38979	0.35176	35.3775	-1.68
5	29.4	0.96333	0.49085	0.31619	35.0210	-2.67
6	28.7	0.03000	0.08662	0.46620	35.9550	-0.08
7	28.6	0.00333	0.02887	0.48872	35.9816	-0.00
8	27.5	1.47000	0.60634	0.27779	34.5150	-4.08
9	29.4	19.50750	2.20882	0.02369	16.4775	-54.21
10	35.5	6.90083	1.31375	0.10675	29.0841	-19.17
11	34.2	1.26750	0.56303	0.29189	34.7175	-3.52

12	33.8	0.44083	0.33205	0.37279	35.5441	-1.22
13	43.5	5.88000	1.21269	0.12429	30.1050	-16.34
14	41.6	0.05333	0.11549	0.45498	35.9316	-0.14
15	39.8	3.41333	0.92395	0.18686	32.5716	-9.48
16	40.7	0.65333	0.40423	0.34658	35.3316	-1.81

De la información se deduce que la novena observación con un p-valor igual a 0.0236 se clasifica como influyente para la suma de cudrados residual SCE. La nueva suma de cuadrados residual si el modelo se ajusta sin esta observación será $SCE(9) = 16.4775$ generando una variación porcentual del 54.21 %, lo cual se verifica en la siguiente salida:

Procedimiento GLM

Variable dependiente: Y

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Modelo	4	15881.92250	3970.48063	2650.60	<.0001
Error	11	16.47750	1.49795		
Total no correg	15	15898.40000			

R-cuadrado	Coef Var	Raiz MSE	Y Media
0.977422	3.848772	1.223910	31.80000

Fuente	DF	Tipo I SS	Cuadrado de la media	F-Valor	Pr > F
Término indep	1	15168.60000	15168.60000	10126.2	<.0001
A	1	579.17411	579.17411	386.64	<.0001
B	1	131.24012	131.24012	87.61	<.0001
A*B	1	2.90827	2.90827	1.94	0.1910

El procedimiento descrito de análisis de residuales aplicado al modelo sin interacción $y_{ijk} = \mu + A_i + B_j + e_{ijk}$ proporciona los siguientes resultados:

OBS	Y	Qi	Ti	Pval	NSCE	Variación
1	23.5	0.86327	0.44871	0.33051	44.42423	-1.90
2	24.6	4.62019	1.03806	0.15908	40.66731	-10.20
3	21.5	1.66327	0.62284	0.27208	43.62423	-3.67
4	24.1	2.54327	0.77018	0.22748	42.74423	-5.61
5	29.4	0.00942	0.04688	0.48166	45.27808	-0.02
6	28.7	0.46173	0.32816	0.37401	44.82577	-1.01

7	28.6	0.62481	0.38174	0.35441	44.66269	-1.37
8	27.5	4.04327	0.97109	0.17461	41.24423	-8.92
9	29.4	25.90173	2.45787	0.01439	19.38577	-57.19
10	35.5	2.81558	0.81036	0.21616	42.47192	-6.21
11	34.2	0.05558	0.11385	0.45555	45.23192	-0.12
12	33.8	0.04327	0.10046	0.46076	45.24423	-0.09
13	43.5	10.08481	1.53366	0.07454	35.20269	-22.26
14	41.6	1.14019	0.51568	0.30737	44.14731	-2.51
15	39.8	0.86327	0.44871	0.33051	44.42423	-1.90
16	40.7	0.00481	0.03349	0.48690	45.28269	-0.01

Según la información obtenida, la novena observación también es influyente para la suma de cuadrados residual del modelo sin interacción con un p-valor de 0.0143 y genera una variación porcentual del 57.2 % si el modelo se ajusta sin esta observación.

A continuación se describe el programa o sintaxis en R para realizar los cálculos anteriores. Para ejecutarlo se debe construir un archivo en formato csv que contenga los datos y la matriz de diseño del modelo $Y = X\beta + e$.

4. La sintaxis en R

A continuación presentamos el código computacional utilizado en el desarrollo de este trabajo.

```
  rm(list=ls(all=TRUE))
# Importar los datos
  DATA1<-read.csv("ARTC3.csv",header=TRUE);DATA1;names(DATA1); attach(DATA1);
  library(MASS)

# Definición de matrices con las variables en el data.
Y<-matrix(c(Y),ncol=1);
M<-matrix(c(M),ncol=1);
W1<-matrix(c(A1),ncol=1);
W2<-matrix(c(A2),ncol=1);
W3<-matrix(c(B1),ncol=1);
W4<-matrix(c(B2),ncol=1);
W5<-matrix(c(C11),ncol=1);
W6<-matrix(c(C12),ncol=1);
W7<-matrix(c(C21),ncol=1);
W8<-matrix(c(C22),ncol=1);

N<-nrow(M);
J<-matrix(1,nrow=N,ncol=1);
X<-matrix(c(M,W1,W2,W3,W4,W5,W6,W7,W8),ncol=9);
```

```
P=ncol(X);
r1<-qr(ginv(X))$rank ;
Ymed<-mean(Y);

# Estimación del modelo con intercepto
B<-(ginv(t(X)%*%X))%*%t(X)%*%Y;
H<- (X%*%(ginv(t(X)%*%X)))%*%t(X);
I<-diag(N);
e<- (I-H)%*%Y;
SCE<- t(e)%*%e;
s2<-SCE[1,1]/(N-r1);
s<-sqrt(s2);
VB<-ginv(t(X)%*%X)*s2;

# Analisis de varianza
SCT<-(t(Y)%*%Y);
SCR<-t(B)%*%(t(X)%*%Y);
F<-(SCR/(r1))/(s2);
PvalF<-1-pf(F,r1,N-r1);

# Análisis de residuales modelo con intercepto
H=(X%*%(ginv(t(X)%*%X)))%*%t(X);
I=diag(N);
hii<-J-diag(H);
ri<- e/hii;
Qi<- e^2/hii;
NSCE<-J%*%SCE-Qi;
VporSCE<-(Qi/SCE[1,1])*100;
MaxQi<- max(Qi) ;Ti<-abs(e/s);
MaxTi<- max(Ti);
t95<-qt(0.95,N-r1);
PVT<-1-pt(Ti,N-r1);

# Salida anova modelo con intercepto
SAL1<-matrix(c(SCR, SCE, SCT, F, PvalF),nrow=5);
rownames(SAL1)<-c("SCR","SCE","SCT","F","P-valor F" );
round(SAL1,5)

# Salida análisis de residuales
SAL2<-matrix(c(Y,Qi,Ti,PVT,NSCE,VporSCE),ncol=6) ;
colnames(SAL2)<-c("Y","Qi","Ti","Pval","NSCE","Variación");
round(SAL2,5)
```

Recibido: 10 de junio de 2011
Aceptado: 12 de octubre de 2011

Referencias

Draper, N. R. & John, J. A. (1980), 'Influential observations and outliers in regression', *Technometrics* **22**.

Jiménez, J. A. y Rincón, L. F. (2000), 'Una generalización de la estadística dfbeta', *Revista Colombiana de Estadística, Universidad Nacional de Colombia* **23**(1).

Morales, M. (2000), 'Estudio de algunas consecuencias derivadas de eliminar una observación influyente en modelos de regresión lineal múltiple', *Tesis Especialización en Estadística. Universidad Nacional de Colombia* .

Rincón, A. (1999), 'Propuesta para caracterizar observaciones influyentes en modelos de regresión lineal múltiple', *Tesis de grado Estadística. Universidad Nacional de Colombia* .

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**(2).

Searle, S. R. (1971), *Linear Models*, John Wiley & Sons.