UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2010, Vol. 3, No. 2

Un método para determinar un grupo de observaciones influyentes en la SCE al ajustar modelos de rango completo

Implementation of a Method to Determine a Group of Influential Observations in the SCE when fitting Full Rank Models

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

Resumen

En este artículo se presenta un método que permite establecer la existencia de un grupo de observaciones influyentes para la SCE al ajustar un modelo lineal de rango completo. En el análisis de residuales se utiliza la estadística Q_i y 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, observaciones influyentes, parámetros estimados, suma de cuadrados residual.

Abstract

Throughout this paper is presented a method that allows us to identify the presence of an influential group of observations for the RSS after fitting a linear model of full rank. Over the residual analysis a Q_i statistic is used and the criteria is built, using the distribution of the statistic mentioned under the assumption than $e_i \sim N(0; \sigma^2)$ and $Cov(e_i, e_j) = 0$ if $(i \neq j)$.

***Key words*:** Estimated Parameters, Full rank Model, Influential Observations, Residual Sum of Squares.

1. La estadística Q_i

En la revista *Comunicaciones en Estadística* de la Universidad Santo Tomás, volumen 2 número 2 página 139, se expone la metodología para calcular la estadística

[a]Docente. Facultad de Estadística. Universidad Santo Tomás.

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

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

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.

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:

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

Se asocia a la estadística $T_i = \dfrac{\sqrt{Q_i(1-h_{ii})}}{s}$ 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. El método propuesto

Considerando que la estadística

$$Q_i = SCE - SCE(i)$$

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, se define:

- La estadística

$$Q_{(i,j)} = SCE(i) - SCE(i,j) = SCE - SCE(j|i) \tag{2}$$

que mide el cambio en la SCE cuando el modelo se ajusta después de eliminar la pareja de registros (i,j), $SCE(i,j)$ es la suma de cuadrados residual al ajustar el modelo después de eliminar la pareja de registros (i,j) y $SCE(j|i)$ es la suma de cuadrados residual antes de ajustar el modelo, para ello se elimina el j-ésimo registro después de eliminar el registro i.

- La estadística

$$Q_{(j|i)} = SCE(i) - SCE - SCE(j|i) \tag{3}$$

que mide el cambio en la SCE cuando antes de ajustar el modelo se elimina el j-ésimo registro después de eliminar el registro i.

De lo anterior se deduce

$$Q_{(i,j)} = SCE - SCE(i) + Q_{(j|i)}$$

es decir

$$Q_{(i,j)} = Q_i + Q_{(j|i)} \tag{4}$$

lo cual significa que $Q_{(i,j)}$, es el impacto sobre la suma de cuadrados residual SCE al eliminar la pareja de registros (i, j), se puede calcular como la suma de Q_i, el impacto de eliminar el i-ésimo registro y el impacto $Q_{(j|i)}$, calculado al eliminar el registro j después de eliminar el registro i. En consecuencia, $Q_{(i,j)}$ toma su valor máximo cuando Q_i y $Q_{(j|i)}$ son valores máximos. Para determinar la pareja de observaciones (i, j) que mayor impacto genera sobre SCE se sugiere el siguiente procedimiento:

- Para los n registros tomados en el análisis, estime el modelo $Y = X\beta + e$ y calcule la estadística Q_i para $i = 1, 2, \cdots n$.
- Elimine la observación para la cual Q_i toma el valor máximo, ajuste el modelo y calcule nuevamente la estadística $Q_{(j|i)}$ para las $n-1$ observaciones presentes.
- Cuando el procedimiento descrito se generaliza aplicando en cada paso el criterio de influencia establecido con la estadística T_i para un nivel de significancia α y se detiene cuando $|T_i| < t_{\alpha/2}$, como resultado se obtiene el grupo de observaciones que mayor impacto generan en la SCE al ajustar el modelo.

3. Ejemplo

Se ilustra el desarrollo teórico expuesto en las secciones anteriores con el siguiente ejemplo aplicado a datos simulados para el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$. Para los datos simulados, inicialmente se ajusta el modelo con las 20 observaciones, se ilustra la salida típica calculada en SAS, con la prueba de normalidad de los residuales. Para $i = 1, 2, \ldots 20$, se calculan las estadísticas Q_i, T_i y el P-valor de T_i, además se muestra la variación porcentual de la suma de cuadrados residual SCE resultante de eliminar el ı-ésimo registro. En este paso y utilizando como criterio de influencia el P-valor de T_i, se caracteriza la observación más influyente para la SCE. Se elimina esta observación, se ajusta nuevamente el modelo y se describen los cambios ocurridos en la SCE y en la prueba de normalidad de los residuales, luego se repite el proceso hasta que el criterio determine que ya no hay observaciones influyentes.

- Ajuste inicial del modelo con las 20 observaciones.

Obs	Y	X1	X2
1	38.91	4.2	1.9
2	25.57	1.2	3.1
3	26.71	1.4	2.8
4	26.67	1.3	2.9
5	23.94	1.8	2.2
6	33.43	1.5	3.0
7	26.07	2.0	1.9
8	25.13	2.2	1.6
9	24.32	2.4	1.3
10	26.24	2.6	1.0
11	25.27	2.5	1.3
12	32.76	3.3	2.3
13	28.81	3.0	1.7
14	25.35	3.2	1.5
15	26.93	3.4	2.1
16	28.80	3.8	2.6
17	29.74	4.0	1.8
18	29.62	4.5	2.3
19	30.28	4.6	1.5
20	34.78	4.8	3.2

La siguiente es la salida del ajuste del modelo

Analysis of Variance

Fuente	DF	Sum of Squares	Mean Square	F-Valor	Pr > F
Modelo	2	145.63275	72.81638	8.29	0.0031
Error	17	149.34390	8.78494		
Total correg	19	294.97665			

Root MSE	2.96394	R-cuadrado	0.4937
Media dependiente	28.46650	Adj R-Sq	0.4341
Coeff Var	10.41203		

Parámetros estimados

Variable	DF	Parameter Estimate	Standard Error	Valor t	Pr > |t|
Intercepto	1	16.85548	3.09710	5.44	<.0001
X1	1	2.16424	0.59210	3.66	0.0020
X2	1	2.55581	1.04811	2.44	0.0260

Tests para normalidad

Test		--Estadístico--		-----P-valor------
Shapiro-Wilk	X	0.818491	Pr < W	0.0017
Kolmogorov-Smirnov	D	0.24573	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.224689	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.285642	Pr > A-Sq	<0.0050

La siguiente es la salida de los valores calculados de las estadísticas Q_i, T_i y el P-valor para las 20 observaciones.

OBS	Y	T	QI	PORCENT	P_VALOR
1	38.91	2.735781	74.474782	49.867976	0.014081
2	25.57	0.609178	4.31418	2.888760	0.550461
3	26.71	0.111903	0.13287	0.088975	0.912210
4	26.67	0.138609	0.21023	0.140770	0.891387
5	23.94	0.821165	6.54848	4.384838	0.422918
6	33.43	1.909870	39.819893	26.663219	0.073171
7	26.07	0.010125	0.00099	0.000663	0.992039
8	25.13	0.194368	0.37245	0.249396	0.848191
9	24.32	0.355000	1.30181	0.871689	0.726955
10	26.24	0.405438	1.83429	1.228236	0.690212
11	25.27	0.107500	0.11859	0.079412	0.915650
12	32.76	0.973090	8.88769	5.951158	0.344152
13	28.81	0.376841	1.34048	0.897583	0.710953
14	25.35	0.764101	5.66150	3.790915	0.455282
15	26.93	0.894445	7.48149	5.009574	0.383572
16	28.80	0.986754	9.79388	6.557942	0.337600
17	29.74	0.125805	0.15485	0.103690	0.901361
18	29.62	0.962537	9.77693	6.546588	0.349272
19	30.28	0.123039	0.16328	0.109333	0.903519
20	34.78	0.216738	0.69088	0.462611	0.830991

Las estadísticas máx$(Q_i) = 74.4747$, $T_i = 2.7357$ y el P-valor=0.014081 calculadas para el primer registro $(y, x_1, x_2) = (38.91, 4.2, 1.9)$, caracterizan esta observación como influyente según el criterio definido y el valor de $Q_i = 74.47$ corresponde al valor en el cual se disminuye la SCE si se ajusta el modelo después de eliminar este registro con una variación porcentual de 49.87 %.

- En el segundo paso, se elimina la observación caracterizada como influyente y se repite el proceso para las 19 observaciones restantes. Resultado de ajustar el modelo después de eliminar el primer registro son: la $SCE = 74.86$ en el nuevo modelo, el coeficiente de determinación aumenta $R^2 = 0.58$ y las pruebas de normalidad de los residuales mejoran. Las estadísticas y prueba de normalidad de los residuales y los valores de las estadísticas Q_i, T_i para el modelo ajustado con las 19 observaciones son las siguientes.

Analysis of Variance

Fuente	DF	Sum of Squares	Mean Square	F-Valor	Pr > F
Modelo	2	105.30049	52.65024	11.25	0.0009
Error	16	74.86912	4.67932		
Total correg	18	180.16961			

Root MSE	2.16317	R-cuadrado	0.5845
Media dependiente	27.91684	Adj R-Sq	0.5325
Coeff Var	7.74863		

Parámetros estimados

Variable	DF	Parameter Estimate	Standard Error	Valor t	Pr > |t|
Intercepto	1	17.56645	2.26738	7.75	<.0001
X1	1	1.70625	0.44712	3.82	0.0015
X2	1	2.62776	0.76516	3.43	0.0034

Tests para normalidad

Test	--Estadístico--		-----P-valor------	
Shapiro-Wilk	X	0.907531	Pr < W	0.0667
Kolmogorov-Smirnov	D	0.117669	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.070528	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.524632	Pr > A-Sq	0.1636

OBS	Y	T	QI	PORCENT	P_VALOR
1	25.57	1.01240	6.36358	8.49961	0.326411
2	26.71	0.27872	0.43961	0.58718	0.784028
3	26.67	0.33981	0.67421	0.90052	0.738417
4	23.94	1.14589	6.79244	9.07241	0.268689
5	33.43	2.50599	36.55217	48.82143	0.023390
6	26.07	0.04544	0.01063	0.01420	0.964313
7	25.13	0.18242	0.17484	0.23353	0.857542
8	24.32	0.35019	0.67569	0.90250	0.730759
9	26.24	0.74406	3.30005	4.40776	0.467624
10	25.27	0.01009	0.00055	0.00074	0.992068

11	32.76	1.62681	13.30828	17.77539	0.123306
12	28.81	0.76628	2.96488	3.96009	0.454663
13	25.35	0.74801	2.90799	3.88410	0.465305
14	26.93	0.90422	4.10095	5.47749	0.379291
15	28.80	0.96264	5.01969	6.70462	0.350049
16	29.74	0.28596	0.43253	0.57772	0.778575
17	29.62	0.77128	3.42119	4.56956	0.451781
18	30.28	0.42676	1.07576	1.43685	0.675237
19	34.78	0.28417	0.6559444	0.87612	0.779918

En el nuevo análisis de residuales en el ajuste del modelo después de eliminar la primera observación caracterizada como influyente, las estadísticas máx$(Q_i) = 36.5521$, $T_i = 2.5059$ y el P-valor=0.0233 calculadas para el registro $(y, x_1, x_2) = (33.43, 1.5, 3)$, caracterizan esta observación como influyente según el criterio definido y el valor de $Q_i = 36.5521$ corresponde al valor en el cual se disminuye la SCE si se ajusta el modelo después de eliminar este registro con una variación porcentual de 48.82 %. El valor de la estadística $Q_{(i,j)} = 111.0268 = 74.4747 + 36.5521$ es el valor máximo de variación de la SCE al eliminar una pareja de observaciones, en este caso las observaciones $(38.91, 4.2, 1.9)$ y $(33.43, 1.5, 3)$.

- En el tercer paso, se elimina la segunda observación caracterizada como influyente y se repite el proceso para las 18 observaciones restantes. Resultado de ajustar el modelo después de eliminar este registro son: la $SCE = 38.3169$ en el nuevo modelo, el coeficiente de determinación aumenta $R^2 = 0.7413$ y las pruebas de normalidad de los residuales mejoran. Las estadísticas y prueba de normalidad de los residuales y los valores de las estadísticas Q_i, T_i para el modelo ajustado con las 18 observaciones son las siguientes.

Analysis of Variance

Fuente	DF	Sum of Squares	Mean Square	F-Valor	Pr > F
Modelo	2	109.76915	54.8845	21.49	<.0001
Error	15	38.31694	2.5544		
Total correg	17	148.0860			

Root MSE	1.59827	R-cuadrado	0.7413
Media dependiente	27.61056	Adj R-Sq	0.7068
Coeff Var	5.78861		

Parámetros estimados

Variable	DF	Parameter Estimate	Standard Error	Valor t	Pr > |t|
Intercepto	1	17.70481	1.67566	10.57	<.0001
X1	1	2.00961	0.33995	5.91	<.0001
X2	1	1.98931	0.58999	3.37	0.0042

Tests para normalidad

Test	-Estadístico--		-----P-valor------	
Shapiro-Wilk	X	0.90989	Pr < W	0.0857
Kolmogorov-Smirnov	D	0.14952	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.05900	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.49665	Pr > A-Sq	0.1942

OBS	Y	T	QI	PORCENT	P_VALOR
1	25.57	0.44624	0.73296	1.912889	0.661793
2	26.71	0.38895	0.49174	1.283374	0.702777
3	26.67	0.36519	0.45192	1.179435	0.720064
4	23.94	1.10031	3.47341	9.064970	0.288538
5	26.07	0.35429	0.35507	0.926684	0.728045
6	25.13	0.11190	0.03597	0.093879	0.912379
7	24.32	0.49678	0.74232	1.937338	0.626545
8	26.24	0.82644	2.22898	5.817224	0.421503
9	25.27	0.02814	0.00236	0.006172	0.977933
10	32.76	2.40763	15.96346	41.661643	0.029383
11	28.81	1.06022	3.09849	8.086500	0.305818
12	25.35	1.10716	3.48028	9.082895	0.285663
13	26.93	1.11686	3.41842	8.921444	0.281622
14	28.80	1.07213	3.41375	8.909262	0.300608
15	29.74	0.26026	0.19584	0.511109	0.798197
16	29.62	1.06583	3.56669	9.308396	0.303354
17	30.28	0.21711	0.15375	0.401270	0.831046
18	34.78	0.66524	1.98124	5.170681	0.515991

En el nuevo análisis de residuales en el ajuste del modelo después de eliminar la segunda observación caracterizada como influyente, las estadísticas máx$(Q_i) = 15.9634$ $T_i = 2.4076$ y el P-valor=0.0293 calculadas para el registro $(y, x_1, x_2) = (32.763.3, 2.3)$, caracterizan esta observación como influyente según el criterio definido y el valor de $Q_i = 15.9634$ corresponde al valor en el cual se disminuye la SCE si se ajusta el modelo después de eliminar este registro con una variación porcentual de 41.66 %. El valor de la estadística $Q_{(i,j,k)} = 126.9902 = 111.0268 + 15.9634$ es el valor máximo de variación de la SCE al eliminar una terna de observaciones, en este caso las observaciones $(38.91, 4.2, 1.9)$, $(33.43, 1.5, 3)$ y $(32.763.3, 2.3)$.

- En el ultimo paso, se elimina la tercera observación $(y, x_1, x_2) = (32.763.3, 2.3)$ caracterizada como influyente y se repite el proceso para las 17 observaciones restantes. Las estadísticas y prueba de normalidad de los residuales y los valores de las estadísticas Q_i, T_i para el modelo ajustado con las 17 observaciones son las siguientes.

Analysis of Variance

Fuente	DF	Sum of Squares	Mean Square	F-Valor	Pr > F
Modelo	2	97.65603	48.82802	30.58	<.0001
Error	14	22.35348	1.59668		
Total correg	16	120.00951			

Root MSE	1.26360	R-cuadrado	0.8137
Media dependiente	27.30765	Adj R-Sq	0.7871
Coeff Var	4.62726		

Parámetros estimados

Variable	DF	Parameter Estimate	Standard Error	Valor t	Pr > |t|
Intercepto	1	18.02227	1.32858	13.57	<.0001
X1	1	1.92520	0.27009	7.13	<.0001
X2	1	1.84179	0.46878	3.93	0.0015

Tests para normalidad

Test	--Estadístico--		-----P-valor------	
Shapiro-Wilk	X	0.91859	Pr < W	0.1398
Kolmogorov-Smirnov	D	0.17624	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.07565	Pr > W-Sq	0.2272
Anderson-Darling	A-Sq	0.49892	Pr > A-Sq	0.1898

OBS	Y	T	QI	PORCENT	P_VALOR
1	25.57	0.37357	0.32276	1.44391	0.71431
2	26.71	0.66116	0.89140	3.98778	0.51923
3	26.67	0.63611	0.86050	3.84953	0.53496
4	23.94	1.26587	2.87869	12.87804	0.22621
5	26.07	0.55234	0.54005	2.41595	0.58942
6	25.13	0.05901	0.00625	0.02799	0.95377
7	24.32	0.56748	0.60573	2.70980	0.57937
8	26.24	1.08454	2.39978	10.73562	0.29644
9	25.27	0.03197	0.00191	0.00855	0.97494
10	28.81	1.48868	3.82738	17.12212	0.15874
11	25.35	1.26273	2.83564	12.68549	0.22731

```
12  26.93  1.19159  2.44503  10.93806  0.25322
13  28.80  1.04991  2.06883   9.25508  0.31154
14  29.74  0.55533  0.56054   2.50764  0.58743
15  29.62  1.03021  2.10907   9.43510  0.32037
16  30.28  0.50580  0.52518   2.34947  0.62086
17  34.78  1.28447  4.78126  21.38937  0.21982
```

En el ajuste del modelo sin las 3 observaciones caracterizadas como influyentes en los pasos previos, según el criterio de influencia ya no existen observaciones influyentes, la $SCE = 22.35$, el coeficiente de determinación aumenta $R^2 = 0.81.37$ y las pruebas de normalidad de los residuales ahora no proporcionan evidencia para rechazar la hipótesis de que los residuales tienen distribución normal.

4. Conclusiones

Es necesario al ajustar un modelo de regresión vía mínimos cuadrados, realizar metódicamente el análisis de residuales para detectar las observaciones influyentes en la estimación de la varianza residual. La experiencia acumulada al aplicar este procedimiento estadístico, muestra cómo en ocasiones un solo dato atípico o influyente aumenta significativamente el error en el modelo, lo cual disminuye la calidad de la estimación, y es necesario en el análisis de los datos establecer los diferentes escenarios para que el experto en el tema de interés visualice las diversas opciones que muestra el procedimiento estadístico en el modelamiento de los datos.

Claramente no es suficiente con caracterizar las observaciones influyentes para la estimación de la varianza residual. Se sabe que una observación influyente en este sentido, al ser eliminada genera, cambios en:

- La estimación de los parámetros, lo cual se puede observar con la estadística DFbeta.
- El coeficiente de determinación.
- La desviación estándar del estimador de cada parámetro y por tanto, en su significancia en el modelo.
- Las pruebas para validar los supuestos adoptados sobre los residuales.

Por todo lo anterior, se debe incluir como parte del modelamiento el estudio de estos escenarios en donde se explique y justifique la influencia global que un registro o grupo de registros tiene sobre el comportamiento de las estadísticas usualmente utilizadas para evaluar, valorar y validar el modelo ajustado o recta de regresión recomendada para su uso en producción. Agregar estos análisis al proceso de modelamiento, es importante para que el experto decida o seleccione el escenario o modelo que mejor interpreta el proceso en el cual se genera la información.

Recibido: 20 de agosto de 2010
Aceptado: 24 de septiembre de 2010

Referencias

Draper, N. R. & John, J. A. (1981), 'Influential observations and outliers in regression', *Technometrics* **23**(1), 21 – 26.

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

Morales, M. A. (2000), *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.

Rincón, A. T. (1999), *Propuesta para caracterizar observaciones influyentes en Modelos de Regresión Lineal Múltiple*, Trabajo de Grado, Pregrado en Estadística, Universidad Nacional de Colombia.

Rincón, L. F. (2009), '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), 139 – 146.

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

Apéndice

Finalmente se anexan los programas que se ejecutan en SAS o en R y que proporcionan el valor de las estadísticas objeto de estudio.

```
PROGRAMA EN SAS
    data a1;
    input Y X1 X2  @@ ;
    cards ;
38.91 4.2 1.9 25.57 1.2 3.1 26.71 1.4 2.8
26.67 1.3 2.9 23.94 1.8 2.2 33.43 1.5 3.0
26.07 2.0 1.9 25.13 2.2 1.6 24.32 2.4 1.3
26.24 2.6 1.0 25.27 2.5 1.3 32.76 3.3 2.3
28.81 3.0 1.7 25.35 3.2 1.5 26.93 3.4 2.1
28.80 3.8 2.6 29.74 4.0 1.8 29.62 4.5 2.3
30.28 4.6 1.5 34.78 4.8 3.2 ;

   proc print ; run ;
   proc reg; model Y=x1 x2 ; run ;
  output out=a2 r=res p=pred;run;
```

```
  proc chart ; hbar res ; run;
  proc univariate normal data=a2; var res ; run;

    proc iml ; reset noprint ;
    use a1 ;
    read all var {X1 X2} into X;
    read all var {y} into y;
    n=nrow(x);
    X=j(n , 1, 1)|| x;
    p=ncol(x);

    obs={1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,
    16,17,18,19,20};
    H=x*inv(t(x)*x)*t(x);
    C=inv(t(x)*x)*t(x) ;
    D=inv(t(X)*x);
    hii=vecdiag(h);
    res=(I(n)-H)*Y ;
    SCE= t(res)*res ;
    gamma=-res/(1-hii);
    s=sqrt(sce/(n-p)) ;
    Qi=(-res)#gamma;
    maxqi=max(qi);
    T=abs(res/s );
    Porcent=(qi/sce)*100;
    p_valor=2*(1-probt(abs(t),(n-p)));
    print obs Y  t qi  porcent  p_valor ;
    quit;


PROGRAMA EN R
 rm(list=ls(all=TRUE))

# PASO 1 ANALISIS CON TODOS LOS DATOS
# Introducir como matrices los datos para el modelo

 Y<-matrix(c(38.91, 25.57,26.71,26.67,23.94,
 26.07,25.13,24.32,26.24,33.43,25.27,28.81,
 25.35,26.93,28.80,29.74,29.62,32.76,30.28,
 34.78),ncol=1);

 X1<-matrix(c(4.2,1.2,1.4,1.3,1.8,2.0,2.2,2.4,
 2.6,1.5,2.5,3.0,3.2,3.4,3.8,
4.0,4.5,3.3, 4.6,4.8),ncol=1);

 X2<-matrix(c(1.9,3.1,2.8,2.9,2.2,1.9,1.6,1.3,
```

```
 1.0,3.0,1.3,1.7,1.5,2.1,2.6,1.8,
 2.3,2.3, 1.5,3.2),ncol=1);

# Estimación del modelo con intercepto
  N<-nrow(X1);
  J<-matrix(1,nrow=N,ncol=1);
  X<-matrix(c(J,X1,X2),ncol=3);
  P=ncol(X);
  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);
   VB<-solve(t(X)%*%X)*s2;

 #Error estandar de los estimadores de los parámetros
  ErB<-sqrt(diag(VB));
  TB<-B/ErB;
  PvalT0<-2*(1-pt(abs(TB[1,1]),N-P));
  PvalT1<-2*(1-pt(abs(TB[2,1]),N-P));
   PvalT2<-2*(1-pt(abs(TB[3,1]),N-P));
  QT<-qt(0.975,N-P);
  LIB<-B-(QT*ErB);
  LSB<- B+(QT*ErB);

#Análisis de varianza
  Ymed<-mean(Y);
  SCTm<-(t(Y)%*%Y)-N*(Ymed^2);
  SCRm<-t(B)%*%(t(X)%*%Y)-N*(Ymed^2);
  F<-(SCRm/(P-1))/(s2);
  PvalF<-1-pf(F,P-1,N-P-1);

#Coeficiente de determinación
 R2<-SCRm/SCTm;
 Rajus<-1-(N-1)*(1-R2)/(N-P);

# 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(0.975,N-P);
```

```
 PVT<-2*(1-pt(Ti,N-P));
 NSCE<-J%*%SCE-Qi;
 VPORSCE<-(Qi/SCE[1,1])*100;

#Salidas
 DATOS<-matrix(c(Y,X1,X2),ncol=3);
 colnames(DATOS)<-c("Y","X1","X2");
 DATOS;

SAL1<-matrix(c(SCRm, SCE, SCTm, F, PvalF),nrow=5);
rownames(SAL1)<-c("SCRm","SCE","SCTm","F","P-valor F" );
 round(SAL1,5);

SAL2<-matrix(c(s2,s,R2,Rajus),nrow=4);
rownames(SAL2)<-c("S^2","s", "R^2", "Rajus");
round(SAL2,5);

SAL3<- matrix(c(B,ErB,TB,LIB,LSB),ncol=5)
rownames(SAL3)<- c("B0", "B1", "B2");
colnames(SAL3)<- c("Estimación","Error estandar",
"Estadistica T","Lim Inf", "Lim Sup")
round(SAL3,5);

SAL4<- matrix(c(PvalT0,PvalT1,PvalT2),ncol=1)
rownames(SAL4)<- c("Pvalor TB0", "Pvalor TB1",
 "Pvalor TB2");
round(SAL4,5);

SAL5<-matrix(c(Y,Qi,Ti,PVT,VPORSCE),ncol=5) ;
 colnames(SAL5)<-c("Y","Qi","Ti","PvalT","VPORSCE");
 round(SAL5,5)
 MaxQi
```