Modelo de Regresión Lineal Múltiple. Análisis de Varianza.

Dr. Víctor Aguirre

Propósito

- Se verá cómo probar si es significativo, globalmente, el modelo. ¿Es al menos una variable significativa?
- ¿El modelo explica una cantidad de variación significativa de la variable dependiente?
- También se verá si es significativa la contribución de un subconjunto adicional de variables explicativas.

Significancia global del modelo.

- ¿Es al menos una variable significativa? Es equivalente a probar

$$H_0 : b_1 = b_2 = ... = b_r = 0$$

$$H_1 : \text{al menos una } b_j \neq 0 \text{ para } j = 1,2,...,r$$

- El cálculo del estadístico de prueba se visualiza por medio de la tabla de análisis de varianza.

Identidad de Análisis de Varianza (ANDEVA).

- Proposición 14

Considere el EMC del modelo de Regresión Lineal Múltiple con b_0 presente. Entonces:

$$SCT = SCR(X_1,\ldots,X_r) + SCE(X_1,\ldots,X_r)$$

(la variación total se puede descomponer en dos fuentes de variación)

Tabla de Análisis de Varianza (ANDEVA)

Fuente	Grados de Libertad	Suma de Cuadrados	Promedio de los Cuadrados	F
Modelo (Regresión)	*r*	*SCR*	*CMR=SCR/r*	$F_{calc} = \frac{CMR}{CME}$
Error (Residuos)	*n-r-1*	*SCE*	*CME=SCE/(n-r-1)*	
Total	*n-1*	*SCT*		

La tabla enfatiza la partición de la variación de Y, también muestra cómo se obtiene el estadístico de prueba.

Regla de rechazo.

Rechazar H_0 si $F_{calc} > F^{a}(r, n-r-1)$

o equivalent emente

Rechazar H_0 si $Valor\ P < \boldsymbol{a}$

donde

$$Valor\ P = P(F(r, n-r-1) > F_{calc})$$

Valores críticos F^{α} (α=5%).

$$P(F(2,14) > 3.74) = 0.05$$

Grados de libertad para el denominador	Grados de libertad para el numerador								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	161	199	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59

Valores críticos F^{α} (α=5%).

Grados de libertad para el denominador	Grados de libertad para el numerador								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96

Relación entre F_{calc} y R^2.

$$\text{Dado que } R^2 = \frac{SCR}{SCT} \quad \text{entonces}$$

$$F_{calc} = \frac{R^2 / r}{(1 - R^2)/(n - r - 1)}$$

Ejemplo, Y=Consumo textiles.

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	8460.936433	4230.468217	136.68309	6.514E-10
Residuos	14	433.3129785	30.95092704		
Total	16	8894.249412			

$$R^2(\,precio, ing / cápita) = 0.9512\ ;$$

$$F_{calc} = \frac{0.9512 / 2}{(1-0.9512)/14} = 136.4426$$

$$F^{0.05}(2,14) = 3.74$$

Contribución incremental.

Supongamos que se acepta que las variables

$$X_1,\dots,X_r$$

estén en el modelo.

Se desea ver si un conjunto adicional

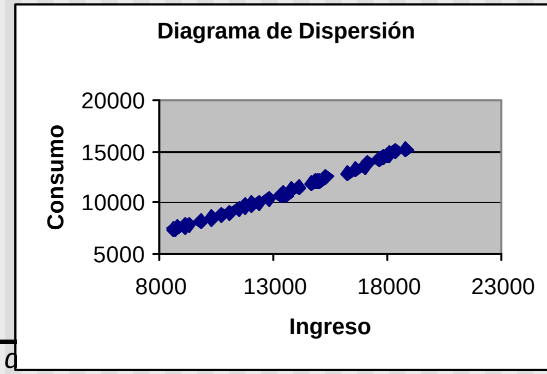
$$X_{r+1},\dots,X_{r+k}$$

explica una cantidad significat iva de variación de Y.

Ejemplo: Y= Consumo real per cápita USA.

Resumen

Estadísticas de la regresión	
Coeficiente de	0.9986271
R2 Original	0.99725609
R^2 ajustado	0.99717769
Error típico	133.09196
Observacione	37



ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	225324426	225324426	12720.5132	1.99E-46
Residuos	35	619971.442	17713.4698		
Total	36	225944398			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	463.176573	98.7911607	4.68844146	4.0964E-05	262.61961	663.733537
Ingreso	0.77941865	0.00691064	112.785253	1.99E-46	0.76538928	0.79344802

¿Vale la pena incluir X_2=Interés a 3 meses y X_3=Inflación como variables explicativas adicionales?

Contribución incremental.

Para medir la contibución incremental de

$X_{r+1},\ldots,X_{r+k}$ dado que $X_1,\ldots,X_r$ están

presentes en el modelo se calcula

$$SCR(X_{r+1},\ldots,X_{r+k} \mid X_1,\ldots,X_r) =$$

$$= SCR(X_1,\ldots,X_r,X_{r+1},\ldots,X_{r+k}) - SCR(X_1,\ldots,X_r)$$

$$= SCE(X_1,\ldots,X_r) - SCE(X_1,\ldots,X_r,X_{r+1},\ldots,X_{r+k})$$

y se usa la siguiente tabla de ANDEVA para ver si hay significancia.

Tabla de ANDEVA, de contribución incremental.

Fuente	Grados de Libertad	Suma de Cuadrados	Promedio de los Cuadrados	F
Modelo Original	***r***	$SCR(X_1,\ldots,X_r)$		
Contribución Incremental	***k***	$A = SCR(X_{r+1},\ldots X_{r+k} \mid X_1,\ldots,X_r)$	A/k	$F_{calc} = \dfrac{A/k}{CME}$
Modelo Nuevo	***r+k***	$SCR(X_1,\ldots,X_r,X_{r+1},\ldots X_{r+k})$		
Error (Modelo Nuevo)	***n-r-k-1***	***SCE***	***CME***	
Total	***n-1***	***SCT***		

Se acepta que la contibución incremental, de $X_{r+1},\ldots,X_{r+k}$ dado que $X_1,\ldots,X_r$, es significativa si $F_{calc} > F^a(k, n-r-k-1)$

Ejemplo: Y= Consumo real per cápita USA.

Resumen

Estadísticas de la regresión	
Coeficiente de correl	0.99898249
R 2 Nuevo	0.99796601
R^2 ajustado	0.9977811
Error típico	118.009955
Observaciones	37

ANÁLISIS DE VARIANZA

	Grados de libert	*ma de cuadra*	*edio de los cuadr*	*F*	*alor crítico de F*
Regresión	3	225484828	75161609.41	5397.07907	1.8123E-44
Residuos	33	459569.534	13926.3495		
Total	36	225944398			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	498.813059	88.2317586	5.653441201	2.6795E-06	319.304054	678.322064
Int 3 meses	-23.4036927	11.9670917	-1.955670872	0.05901444	-47.7509432	0.94355778
Inflación	-4.0249502	9.68259812	-0.41568907	0.68032833	-23.7243598	15.6744594
Ingreso	0.7883778	0.00674182	116.9384938	8.6198E-45	0.77466147	0.80209414

$$SCR(Int3meses, Inflación \mid Ingreso) =$$

$$225484828 - 225324426 = 160402$$

Ejemplo: Y= Consumo real per cápita USA.

Fuente	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Modelo Original	1	225324426			
Contribución Incremental	2	160402	80201	5.76	0.00715587
Regresión	3	225484828			
Residuos	33	459569.53	13926		
Total	36	225944398			

Como *ValorP* < 0.05 entonces *Int3meses* e *Inflación* explican una cantidad adicional significativa de variación de *Consumo* aún dado que *Ingreso* está presente en el modelo.

Relación de F_{calc} con R^2 Original y R^2 Nueva

$$X_1, \ldots, X_r \rightarrow R_O^2$$

$$X_1, \ldots, X_r, X_{r+1}, \ldots, X_{r+k} \rightarrow R_N^2$$

$$F_{calc} = \frac{(R_N^2 - R_O^2)/k}{(1 - R_N^2)/(n - r - k - 1)}$$

Ejemplo: Y= Consumo real per cápita USA.

R 2 Original	0.997256088					
R 2 Nuevo	0.997966006					
n	37			**gl**		
k (adicionales)	2	**num**	0.000354959	2	**F calc**	**Valor P**
r (original)	1	**denom**	6.16362E-05	33	5.75894	0.007155871

Contribución incremental de una sola variable.

Resulta de interés medir la contibución incremental de una sola variable explicativa, por ejemplo :

¿Es significativa la variación explicada por X_1 dado que $X_2,\dots,X_r$ están presentes en el modelo?

Se puede demostrar que en este caso

$$F_{calc} = \frac{(R^2_{X_1,X_2,\dots,X_r} - R^2_{X_2,\dots,X_r})}{(1 - R^2_{X_1,X_2,\dots,X_r})/(n-r-1)} = \left(\frac{\hat{b}_1}{E\hat{E}(\hat{b}_1)} \right)^2$$

$$= (\text{Estadístico t})^2$$

Contribución incremental de una sola variable.

De lo anterior se sigue que, cuando se prueba si es significativa X_1 o $H_1 : b_1 \neq 0,$ se interpreta también como ver si X_1 explica una cantidad significativa de variación dado que $X_2,\ldots,X_r$ están presentes en el modelo.

Lo mismo aplica para el resto de las variables en el modelo.

Ejemplo: Y= Consumo real per cápita USA.

Resumen

Estadísticas de la regresión	
Coeficiente de corr	0.99886449
R 2 Original	0.997730269
R^2 ajustado	0.997596756
Error típico	122.8141741
Observaciones	37

	Coeficientes	*Error típico*	*Estadístico t*
Intercepción	485.0209621	91.53	5.299043529
Inflación	-18.0477823	6.7717	-2.66516635
Ingreso	0.783855965	0.0066	118.9328828

Resumen

Estadísticas de la regresión	
Coeficiente de correl	0.998982485
R 2 Nuevo	0.997966006
R^2 ajustado	0.997781098
Error típico	118.0099551
Observaciones	37

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	498.8130593	88.23175859	5.653441201	2.67947E-06
Int 3 meses	-23.4036927	11.96709172	-1.955670872	0.059014444
Inflación	-4.0249502	9.682598116	-0.41568907	0.680328333
Ingreso	0.788377804	0.006741816	116.9384938	8.61981E-45

R 2 Original	0.997730269
R 2 Nuevo	0.997966006

				gl		
n	37					
k (adicionales)	1	**num**	0.000235737	1	**F calc**	**Valor P**
r (original)	2	**denom**	6.16362E-05	33	3.82465	0.059014444

$$(-1.955670872)^2 = 3.82465$$