OPINION

Cuidado al calcular e interpretar el Coeficiente de Determinación con el software EXCEL en el caso de modelos lineales sin término constante

Por: Arturo Rubio Donet

El objetivo de estos comentarios es compartir experiencias logradas durante el procesamiento y análisis de datos e intercambiar opiniones sobre sistemas de procesamiento de uso comercial difundidos en nuestro medio. Se agradece la atención a estos comentarios así como las opiniones y sugerencias que hicieran llegar.

Los modelos de regresión que no incorporan a un término constante comprenden a modelos que por definición deben pasar por el origen, esto es si todas las variables independientes son cero, la variable dependiente también será cero, estos es:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_p X_{pi} + \varepsilon_i$$

Para estos casos las ecuaciones normales no comprenderán desviaciones respecto a los promedios de las variables sino directamente las sumas de cuadrados de las variables. Esto significa que no deberá considerarse el término de corrección en las sumas de cuadrados.

En los casos donde hay necesidad de estimar modelos lineales que pasan por el origen, es decir que no contienen un término constante, utilizando el procesador EXCEL se deberá tener especial cuidado para evaluar el valor del coeficiente de determinación (R^2) resultante y en su interpretación ya que el método que utiliza este software para este caso específico no es adecuado ya que utiliza una relación donde se resta el término de corrección que en este caso no corresponde por ser un modelo sin término constante:

$$R^2 = \frac{\hat{\beta}' X' Y - n\overline{Y}^2}{Y' Y - n\overline{Y}^2}$$

En este caso específico, al usar esta relación para el cálculo el coeficiente de determinación como indicador del grado de ajuste logrado al modelo puede resultar un desconcertante e ilógico valor negativo.

Contrariamente, al utilizar la relación $R^2 = \frac{\hat{\beta}' X' Y}{Y' Y}$ se obtendrán elevados coeficientes de determinación si el promedio de la variable dependiente es alto, arribándose a la conclusión que se llega a un mejor ajuste de un modelo sin término constante que con él. Por tanto, hay que tener sumo cuidado al calcular e interpretar el valor de este coeficiente cuando el modelo no contiene a un término constante.

EJEMPLOS NUMERICOS:

Sean los datos:

Y	1	4	6	9	15
X	1	3	5	7	12

Asumiendo un modelo con término constante $Y=B_o+B_1X$ las ecuaciones normales correspondientes son:

$$X'X\beta = X'Y$$

$$\begin{bmatrix} 5 & 28 \\ 28 & 228 \end{bmatrix}\begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \begin{bmatrix} 35 \\ 286 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} b_O \\ b_1 \end{bmatrix} = \begin{bmatrix} -0.078652 \\ 1.264045 \end{bmatrix}$$

Análisis de variancia sin ajustar por la media es

Fuente	G.L.	S.C.	CM
Regresión	2	358.764	179.382
Residual	3	0.236	0.079
Total	5	359.000	

Análisis de variancia ajustada por la media es:

Fuente	G.L.	S.C.	CM
Regresión	1	113.764	113.764
Residual	3	0.236	0.079
Total	4	114.000	

$$R^2 = \frac{\hat{\beta}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} = \frac{358.764 - 5(7)^2}{359 - 5(7)^2} = 0.997930$$

A través de la Estimación Lineal del software EXCEL el resultado considerando término constante es:

Pendiente	1.264045	-0.078652	Término Constante
Estadística t	0.033236	0.2224438	Estadística t
R^2	0.997930	0.280449	Error estándar
Fc	1446.43	3	GL(Residual)
SC(Regresión)	113.764	0.236	SC(Residual)

Asumiendo un modelo sin término constante $Y=B_1X$ la ecuación normal es:

$$X'X\beta = X'Y$$

$$[228][b_1] = [286]$$

$$\hat{\beta} = [b_1] = [1.254386]$$

El análisis de variancia será

Fuente	G.L.	S.C.	CM
Regresión	1	358.754	358.754
Residual	4	0.246	0.0615
Total	5	359.000	

$$R^2 = \frac{\hat{\beta}'X'Y}{Y'Y} = \frac{358.764}{359} = 0.9993$$ Es un valor muy alto

A través de la Estimación Lineal del software EXCEL el resultado considerando término constante es:

Pendiente	1.254386	0	Término Constante
Estadística t	0.016411		Estadística t
R^2	0.997845	0.24779	Error estándar
Fc	1852.571	4	GL(Residual)
SC(Regresión)	113.754	0.246	SC(Residual)

Sean los datos:

Y	1	4	6	9	15
X	1	3	5	7	2

Asumiendo un modelo con término constante $Y=B_o+B_1X$ las ecuaciones normales correspondientes son:

$$X^{'}X\beta = X^{'}Y$$

$$\begin{bmatrix} 5 & 18 \\ 18 & 228 \end{bmatrix}\begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \begin{bmatrix} 35 \\ 286 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} b_O \\ b_1 \end{bmatrix} = \begin{bmatrix} 5.448276 \\ 0.431034 \end{bmatrix}$$

Análisis de variancia sin ajustar por la media es

Fuente	G.L.	S.C.	CM
Regresión	2	249.31	124.65
Residual	3	109.69	36.56
Total	5	359.00	

Análisis de variancia ajustada por la media es:

Fuente	G.L.	S.C.	CM
Regresión	1	4.31	4.31
Residual	3	109.69	36.56
Total	4	114.00	

$$R^2 = \frac{\hat{\beta}^{'}X^{'}Y - n\overline{Y}^2}{Y^{'}Y - n\overline{Y}^2} = \frac{249.31 - 5(7)^2}{359 - 5(7)^2} = 0.0378$$

A través de la Estimación Lineal del software EXCEL el resultado considerando término constante es:

Pendiente	0.431034	5.448276	Término Constante
Estadística t	1.255	5.266	Estadística t
R^2	0.0378	6.047	Error estándar
Fc	0.12	3	GL(Residual)
SC(Regresión)	4.31	109.69	SC(Residual)

Asumiendo un modelo sin término constante $Y=B_1X$ la ecuación normal es:

$$X^{'}X\beta = X^{'}Y$$

$$[88][b_1] = [136]$$

$$\hat{\beta} = [b_1] = [1.545455]$$

El análisis de variancia

Fuente	G.L.	S.C.	CM
Regresión	1	210.18	210.18
Residual	4	148.82	37.20
Total	5	359.000	

$$R^2 = \frac{\hat{\beta}^{'}X^{'}Y}{Y^{'}Y} = \frac{210.18}{359} = 0.5854$$

Se aprecia que es un valor muy alto

A través de la Estimación Lineal del software EXCEL el resultado considerando término constante será:

Pendiente	1.545455	0	Término Constante
Estadística t	0.650		Estadística t
R^2	-0.305	6.0996	Error estándar
Fc	-0.936	4	GL(Residual)
SC(Regresión)	-34.8182	148.82	SC(Residual)

Se aprecia que resulta un coeficiente negativo, esto ocurre por que EXCEL usa la relación:

$$R^2 = \frac{\hat{\beta}^{'}X^{'}Y - n\overline{Y}^2}{Y^{'}Y - n\overline{Y}^2} = \frac{210.18 - 5(7)^2}{359 - 5(7)^2} = \frac{-34.82}{114} = -0.3054$$