Regresión lineal múltiple. Ejemplo

Las variables a utilizar son,

Y: gasto semanal en alimentación (miles de $)
X_1: ingreso semanal (miles de $)
X_2: tamaño de la familia

Las observaciones se presentan el las hojas de excel Ejemmult.xls

Se tiene una muestra aleatoria de n = 15 observaciones, y el modelo propuesto es,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

donde el error ε cumple todos los supuestos de un modelo de regresión lineal.

Estimación puntual

$$X^T X = \begin{pmatrix} n & \sum X_{1i} & \sum X_{2i} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{1i} X_{2i} \\ \sum X_{2i} & \sum X_{1i} X_{2i} & \sum X_{2i}^2 \end{pmatrix} = \begin{pmatrix} 15 & 42 & 55 \\ 42 & 188.08 & 140.8 \\ 55 & 140.8 & 219 \end{pmatrix}$$

$$X^T \underline{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_{1i} Y_i \\ \sum X_{2i} Y_i \end{pmatrix} = \begin{pmatrix} 8.07 \\ 32.036 \\ 28.96 \end{pmatrix}$$

$$(X^T X)\,\underline{\hat{\beta}} = \begin{pmatrix} 15 & 42 & 55 \\ 42 & 188.08 & 140.8 \\ 55 & 140.8 & 219 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 8.07 \\ 32.063 \\ 28.96 \end{pmatrix} = X^T \underline{Y} \quad \text{Ecs.normales}$$

La solución del sistema es,

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T \underline{Y} = \begin{pmatrix} 1.35985 & -0.090255 & -0.28202 \\ -0.090255 & 0.016549 & 0.012602 \\ -0.28202 & 0.012602 & 0.06729 \end{pmatrix} \begin{pmatrix} 8.07 \\ 32.063 \\ 28.96 \end{pmatrix} = \begin{pmatrix} -0.16046 \\ 0.148727 \\ 0.076915 \end{pmatrix}$$

El modelo ajustado es entonces,

$$\hat{Y} = -0.16046 + 0.148727\, X_1 + 0.076915\, X_2$$

La suma de cuadrados de residuos esta dada por,

$$SC_e = \Sigma\, e_i^2 = \underline{Y}^T \underline{Y} - \underline{\hat{\beta}}^T X^T \underline{Y} = 5.7733 - 5.70118 = 0.07212$$

asi que el estimador de σ^2 es

$$\hat{\sigma}^2 = \frac{SC_e}{n-p} = \frac{0.07212}{15-3} = 0.00601 \quad , \quad \hat{\sigma} = 0.07752$$

Se hace notar que la varianza estimada de Y, sin tomar en cuenta el modelo de regresión , es

$$S_Y^2 = \frac{\Sigma\, Y_i^2 \;-\; n\,\overline{Y}^2}{n-1} = \frac{5.7733 - \;15\;(0.538)^2}{15-1} = \frac{1.43164}{14} = 0.10226 \quad , \quad S_Y = 0.319781$$

Es conveniente notar que, al comparar $\hat{\sigma}^2$ con S_Y^2 , la información muestral nos dice que $V(Y \mid X_1, X_2) \leq V(Y)$.

La varianza estimada del vector $\hat{\underline{\beta}}^T$ se obtiene así,

$$\hat{V}(\hat{\underline{\beta}}) = \hat{\sigma}^2 (X^T X)^{-1} = (0.07752)^2\, (X^T X)^{-1} = \begin{pmatrix} 0.00817 & -0.000556 & -0.001964 \\ -0.000556 & 9.94(10)^{-5} & 7.57(10)^{-5} \\ -0.001964 & 7.57(10)^{-5} & 0.000404 \end{pmatrix}$$

Coeficientes de correlación y coeficiente de determinación

$$R^2 = 1 - \frac{SC_e}{SC_T} = 1 - \frac{0.07212}{\underline{Y}^T \underline{Y} - n\overline{Y}^2} = 1 - \frac{0.07212}{1.43164} = 1 - 0.0504 = 0.9496$$

El coeficiente de *determinación ajustado* esta dado por

$$R_{ad}^2 = 1 - \frac{SC_e\,/\,(n-p)}{SC_T\,/\,(n-1)} = 1 - \frac{0.07212\,/\,(15-3)}{1.43164\,/\,(15-1)} = 1 - 0.0588 = 0.9412$$

El valor dado por la raíz cuadrada de R^2 se llama *correlación múltiple*. Su valor en este caso es R = 0.9745.

La matriz de correlaciones simples entre Y, X_1y X_2 es

	Y	X_1	X_2
Y	1	0.9425	-0.1265
X_1	0.9425	1	-0.3776
X_2	-0.1265	-0.3776	1

Las correlaciones parciales $r_{YX_1|X_2}$ y $r_{YX_2|X_1}$, estan dadas por

$$r_{YX_1!X_2}^2 = \frac{R_{Y|X_1X_2}^2 - r_{YX_2}^2}{1 - r_{YX_2}^2} = \frac{0.9496 - (-0.1265)^2}{1 - (-0.1265)^2} = \frac{0.9336}{0.98399} = 0.9488 \quad , \quad r_{YX_2!X_1}^2 = \frac{R_{Y|X_1X_2}^2 - r_{YX_1}^2}{1 - r_{YX_1}^2} = 0.5488$$

Intervalos de confianza

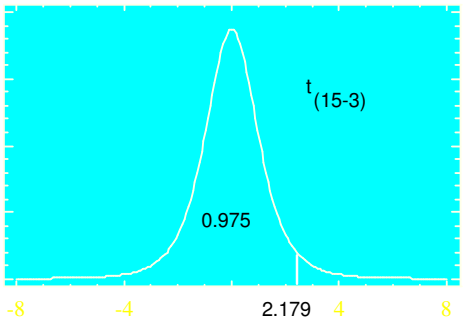
$$\hat{Y} = -0.16046 + 0.148727\, X_1 + 0.076915\, X_2$$
$$(0.090389) \quad (0.009971) \qquad (0.020107)$$

Para β_2, el intervalo esta dado por:

[0.076915 -2.179(0.02011),0.076915+2.179(0.02011)]

es decir,

$\beta_2 \in$ [0.0331,0.1207] con 95 % de confianza.



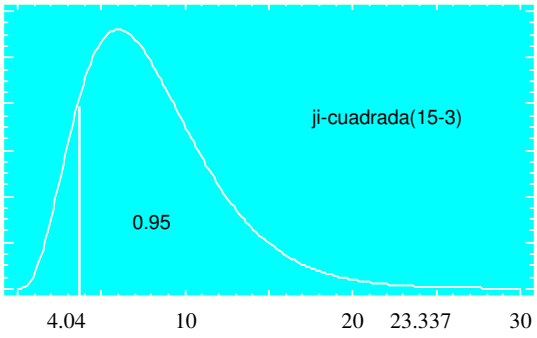
Para σ^2, el intervalo esta dado por:

[(15-3) 0.00601/ 23.337, (15-3) 0.00601/ 4.404]

es decir,

$\sigma^2 \in$ [0.0031, 0.0164] con 95 % de confianza.

$\sigma \in$ [0.0557, 0.12797] con aprox. 95 % de confianza



□

Prueba de hipótesis

H_0: $\beta_2 = 1$ vs H_1: $\beta_2 < 1$

Estadística de prueba: $T = \dfrac{\hat{\beta}_2 - \beta_2^*}{S_{\hat{\beta}_2}} \sim t_{(n-p)} = t_{12}$

$$T = \frac{0.076915 - 1}{0.020107} = -45.9$$

valor p << 0.05, se rechaza H_0

H_0: $\beta_1 = 2\beta_2$ $(\beta_1 - 2\beta_2 = 0)$

H_1: $\beta_1 \neq 2\beta_2$ $(\beta_1 - 2\beta_2 \neq 0)$

$$\hat{\beta}_1 - 2\hat{\beta}_2 = 0.14727 - 2(0.076915) = -0.005103$$

$$\hat{V}(\hat{\beta}_1 - 2\hat{\beta}_2) = \hat{V}(\hat{\beta}_1) + 4\hat{V}(\hat{\beta}2) - 2(2)\hat{C}ov(\hat{\beta}_1, \hat{\beta}_2) =$$
$$= 9.94(10)^{-5} + 4(0.000404) - 4(7.57)(10)^{-5} =$$
$$= 0.0014126$$

Estadística de prueba: $T = \dfrac{\hat{\beta}_1 - 2\hat{\beta}_2}{S_{\hat{\beta}_1 - 2\hat{\beta}_2}} = \dfrac{-0.005103}{\sqrt{0.0014126}} = -0.13577$. Al comparar este valor con el valor en tablas de una $t_{(n-p)} = t_{(12)}$, resulta que valor p = 0.89, asi que no se rechaza H_0. Por otra parte, como se sabe que $t^2_{(k)} = F_{(1,k)}$, tambien se podria haber comparado $T^2 = 0.018434$ con el valor de tablas de una $F_{(1,12)}$, llegando a la misma conclusión.

Tabla de análisis de varianza

ANOVA

f.v.	SC	gl	CM	F
Regresión	1.359542	3-1=2	0.679771	113.3
Error	0.072098	15-3=12	0.00601	
Total	1.43164	15-1=14		

Al comparar el valor de F = 113.3 con el valor de tablas de una distribución $F_{(2,12)}$, es evidente que se rechaza la hipótesis H_0: $\beta_1 = \beta_2 = 0$ a favor de la hipótesis H_1: por lo menos una β_i es diferente de cero.

Predicción puntual

Sean $X_1 = 2.5$ y $X_2 = 4$. La sustitución de estos valores en el modelo ajustado da como resultado

$$\underline{X}_0^T \hat{\underline{\beta}} = = 0.16046 + 0.148727\ (2.5) + 0.076915\ (4) = 0.51902\ ,$$

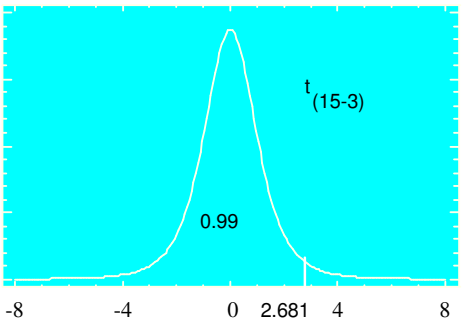
donde $\underline{X}_0^T = (1,2.5,4)$.

La predicción puntual del valor de Y cuando $X_1 = 2.5$ y $X_2 = 4$, es $\hat{Y}_0 = 0.51902$, y la predicción puntual del valor esperado de Y cuando $X_1 = 2.5$ y $X_2 = 4$, es también $\hat{y} = 0.51902$.

Predicción por intervalo

El intervalo para estimar el valor esperado de Y cuando $X_1 = 2.5$ y $X_2 = 4$, esta dado por

$$\hat{y}_0 \pm t^{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2\ \underline{X}_0^T (X^T X)^{-1} \underline{X}_0}$$



Se recuerda que $\hat{\sigma}^2 = 0.00601$. Por otro lado, $\underline{X}_0^T (X^T X)^{-1} \underline{X}_0 = 0.073112$. Entonces, el intervalo esta dado por

$$0.51902 \pm 2.681 \sqrt{0.00601\ (0.073112)}\ ,$$

es decir,

$$y_0 \in [0.462821, 0.575219]$$

con 98% de confianza.

El intervalo para estimar el valor de Y_0 cuando $X_1 = 2.5$ y $X_2 = 4$, esta dado por

$$\hat{Y}_0 \pm t^{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2\ [1 + \underline{X}_0^T (X^T X)^{-1} \underline{X}_0]}$$

Al sustituir valores el intervalo queda así,

$$0.51902 \pm 2.681 \sqrt{0.00601\ [1 + 0.073112]}\ ,$$

es decir,

$$Y_0 \in [0.3037, 0.7343]$$

con 98% de probabilidad.