
En el análisis de la posible influencia del peso, X_1 y del nivel de ácido úrico, X_2, sobre el nivel de colesterol, Y, en los individuos de una población, se seleccionó al azar a 10 personas de la población en estudio, anotándose el valor, que en ellos tomaban, las tres variables antes mencionadas. Los resultados obtenidos fueron los siguientes:

X_1	50	80	75	80	68	75	70	80	90	60
X_2	40	70	50	65	55	60	60	62	69	63
Y	200	350	250	300	340	340	300	360	400	220

Se pide:

a) **Determinar el hiperplano de regresión muestral de Y sobre X_1, X_2.**

b) **Contrastar, a nivel $\alpha = 0'05$, si el hiperplano determinado explica suficientemente bien a la variable Y en función de X_1 y X_2.**

---

a) Para calcular el hiperplano (en este caso plano al haber sólo dos variable regresoras) de regresión de Y sobre X_1, X_2,

$$y_t = \hat{\beta}_0 + \hat{\beta}_1\, x_1 + \hat{\beta}_2\, x_2$$

debemos determinar y resolver, previamente, el *sistema de ecuaciones normales*

$$\left\{\begin{array}{rcrcrcl} n\,\hat{\beta}_0 & + & \hat{\beta}_1 \displaystyle\sum_{j=1}^{n} x_{1j} & + & \hat{\beta}_2 \displaystyle\sum_{j=1}^{n} x_{2j} & = & \displaystyle\sum_{j=1}^{n} y_j \\ \hat{\beta}_0 \displaystyle\sum_{j=1}^{n} x_{1j} & + & \hat{\beta}_1 \displaystyle\sum_{j=1}^{n} x_{1j}^2 & + & \hat{\beta}_2 \displaystyle\sum_{j=1}^{n} x_{1j}\, x_{2j} & = & \displaystyle\sum_{j=1}^{n} x_{1j}\, y_j \\ \hat{\beta}_0 \displaystyle\sum_{j=1}^{n} x_{2j} & + & \hat{\beta}_1 \displaystyle\sum_{j=1}^{n} x_{1j}\, x_{2j} & + & \hat{\beta}_2 \displaystyle\sum_{j=1}^{n} x_{2j}^2 & = & \displaystyle\sum_{j=1}^{n} x_{2j}\, y_j \end{array}\right.$$

que para los datos del enunciado queda igual a

$$\left\{\begin{array}{rcrcrcr} 10\cdot\hat{\beta}_0 & + & 728\cdot\hat{\beta}_1 & + & 594\cdot\hat{\beta}_2 & = & 3060 \\ 728\cdot\hat{\beta}_0 & + & 54174\cdot\hat{\beta}_1 & + & 43940\cdot\hat{\beta}_2 & = & 228370 \\ 594\cdot\hat{\beta}_0 & + & 43940\cdot\hat{\beta}_1 & + & 36024\cdot\hat{\beta}_2 & = & 185380 \end{array}\right.$$

sistema de tres ecuaciones con tres incógnitas, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, que tiene como soluciones los valores

$$\begin{array}{rcr} \hat{\beta}_0 & = & -55'58 \\ \hat{\beta}_1 & = & 4'2301 \\ \hat{\beta}_2 & = & 0'9029 \end{array}$$

El hiperplano buscado será, por tanto,

$$y_t = -55'58 + 4'2301\, x_1 + 0'9029\, x_2$$

mediante el cual, si el Análisis de la Regresión Lineal Múltiple, que haremos a continuación, permite aceptarlo como modelo, podríamos considerar como razonable que un individuo de la población en estudio con un peso de $x_1 = 85$ kgr y un nivel de ácido úrico de $x_2 = 66$, tenga un nivel de colesterol de

$$y_t = -55'58 + 4'2301 \cdot 85 + 0'9029 \cdot 66 = 363'57.$$

En la determinación de los coeficientes de regresión, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, hemos utilizado el sistema de ecuaciones normales. Equivalentemente, podríamos haber utilizado la notación matricial empleada en EII-sección 6.6, siendo

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} 1 & 50 & 40 \\ 1 & 80 & 70 \\ \vdots & \vdots & \vdots \\ 1 & 60 & 63 \end{pmatrix}$$

$$Y = \begin{pmatrix} 200 \\ 350 \\ \vdots \\ 220 \end{pmatrix}$$

y siendo los coeficientes de regresión iguales a

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1}\, X'Y$$

es decir,

$$X' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 50 & 80 & \cdots & 60 \\ 40 & 70 & \cdots & 63 \end{pmatrix}$$

la matriz traspuesta de la matriz X, obtenida de esta última intercambiando las filas y las columnas, siendo

$$(X'X)^{-1} = \begin{pmatrix} 10 & 728 & 594 \\ 728 & 54174 & 43940 \\ 594 & 43940 & 36024 \end{pmatrix}^{-1} = \begin{pmatrix} 5'4148 & -0'0325 & -0'0496 \\ -0'0325 & 0'0019 & -0'0018 \\ -0'0496 & -0'0018 & 0'0031 \end{pmatrix}$$

la inversa del producto de las matrices $X'X$, y siendo

$$X'Y = \begin{pmatrix} 3060 \\ 228370 \\ 185380 \end{pmatrix}$$

Por último, será

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1} X'Y = \begin{pmatrix} 5'4148 & -0'0325 & -0'0496 \\ -0'0325 & 0'0019 & -0'0018 \\ -0'0496 & -0'0018 & 0'0031 \end{pmatrix} \cdot \begin{pmatrix} 3060 \\ 228370 \\ 185380 \end{pmatrix}$$

$$= \begin{pmatrix} -55'58 \\ 4'2301 \\ 0'9029 \end{pmatrix}$$

b) Una vez determinado el hiperplano de regresión muestral por uno u otro procedimiento, en este apartado vamos a analizar si éste explica suficientemente bien a la variable Y en función de X_1 y X_2, contrastando la hipótesis nula H_0:*ninguna de las variables independientes X_i tiene valor en la explicación de Y*, o equivalentemente $H_0 : \beta_1 = ... = \beta_k = 0$, frente a la alternativa de que alguna X_i sirve para explicar a la variable Y.

Para ello utilizaremos la tabla de Análisis de la Varianza para la Regresión Lineal Múltiple

F. variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión múltiple	$SSEX = \sum_{i=1}^{n} (y_{t_i} - \overline{y})^2$	k	$\frac{SSEX}{k}$	$\frac{\frac{SSEX}{k}}{\frac{SSNEX}{n-k-1}}$
Residual	$SSNEX = SST - SSEX$	$n-k-1$	$\frac{SSNEX}{n-k-1}$	
Total	$SST = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$	$n-1$		

Primero calcularemos la *suma de cuadrados debida a la regresión lineal múltiple*

$$SSEX = \sum_{i=1}^{n} (y_{t_i} - \overline{y})^2$$

en donde y_{t_i} representa el valor teórico obtenido por el hiperplano de regresión muestral y_t en el punto (x_{1i}, x_{2i}), $i = 1, ..., 10$; es decir, por ejemplo

$$y_{t_1} = -55'58 + 4'2301 \cdot 50 + 0'9029 \cdot 40 = 192'041.$$

Por otro lado, la media de las y_i es $\overline{y} = \sum_{i=1}^{10} y_i/10 = 306$, con lo que obtenemos la siguiente tabla de cálculos:

y_{t_i}	$y_{t_i} - \overline{y}$	$(y_{t_i} - \overline{y})^2$
$192'041$	$-113'959$	$12986'654$
$346'031$	$40'031$	$1602'481$
$306'822$	$0'822$	$0'677$
$341'517$	$35'517$	$1261'422$
$281'726$	$-24'274$	$589'213$
$315'852$	$9'851$	$97'052$
$294'701$	$-11'299$	$127'667$
$338'808$	$32'808$	$1076'352$
$387'429$	$81'429$	$6630'698$
$255'109$	$-50'891$	$2589'924$
		26962

Por tanto, será

$$SSEX = \sum_{i=1}^{n} (y_{t_i} - \overline{y})^2 = 26962$$

suma de cuadrados que tiene $k = 2$ grados de libertad al haber sólo dos variables regresoras.

Por otro lado, la *suma total de cuadrados* es

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 974200 - \frac{3060^2}{10} = 37840$$

la cual tiene $n - 1 = 9$ grados de libertad.

Por último, la *suma residual de cuadrados* se calcula por diferencia de las otras dos,

$$SSNEX = SST - SSEX = 37840 - 26962 = 10878$$

con grados de libertad igual a la diferencia de grados de libertad de las otras dos sumas de cuadrados, $9 - 2 = 7$.

Los cuadrados medios de la tabla de Análisis de la Varianza se calculan ahora como cociente entre las sumas de cuadrados y sus grados de libertad:

Cuadrado medio correspondiente a la Regresión Lineal Múltiple:

$$\frac{SSEX}{2} = \frac{26962}{2} = 13481$$

Cuadrado medio Residual:

$$\frac{SSNEX}{7} = \frac{10878}{7} = 1554$$

siendo el estadístico del contraste el cociente de estos dos cuadrados medios:

$$F = \frac{SSEX/2}{SSNEX/7} = \frac{13481}{1554} = 8'675.$$

Todos estos cálculos se resumen en la tabla de Análisis de la Varianza para la Regresión Lineal Múltiple

F. variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal múltiple	$SSEX = 26962$	2	13481	$F = 8'675$
Residual	$SSNEX = 10878$	7	1554	
Total	$SST = 37840$	9		

Como este estadístico, antes de obtener la muestra y, por tanto, tomar un valor concreto, se distribuye como una F de Snedecor con grados de libertad el par formado por los grados de libertad de las dos sumas de cuadrados que forman el cociente de F, es decir, en este caso $(2, 7)$, el punto crítico para un nivel de significación $\alpha = 0'05$, será $F_{(2,7);0'05} = 4'7374 < 8'675 = F$, por lo que rechazaremos la hipótesis nula H_0, concluyendo con la alternativa de que el hiperplano de regresión calculado en el apartado anterior es válido para explicar Y en función de X_1 y X_2.

El p-valor del test, no obstante, no es lo suficientemente contundente al estar acotado por los valores

$$0'01 < \text{p-valor} < 0'025.$$

El *coeficiente de correlación múltiple muestral*

$$R_{y.12} = \sqrt{\frac{SSEX}{SST}} = \sqrt{\frac{26962}{37840}} = \sqrt{0'7125} = 0'8441$$

está en línea con las conclusiones anteriores: no está demasiado cerca de 1, pero es significativamente cercano a dicho valor; o mejor dicho, es significativamente grande como para que el estadístico

$$F' = \frac{R^2_{y.12}/2}{(1 - R^2_{y.12})/(10 - 2 - 1)}$$

(igual, como vimos en CB-sección 10.4.1, al estadístico F antes determinado), rechace la hipótesis nula $H_0 : \rho_{y.12} = 0$, de ser cero el coeficiente de correlación múltiple entre Y y el resto de las —en este caso dos— variables regresoras X_i.