"MODELO LINEAL DE RANGO COMPLETOò

Modelo matemático

$$y = g(x_1, x_2,... x_k)$$

y	variable a explicar
$x_1, x_2,...x_k$	variables explicativas

De esta manera, conocidos los valores de las variables explicativas y la forma específica de g, es posible estimar el valor correspondiente de la variable y.

Modelo estadístico

$$y = g(x_1, x_2,... x_k) + \varepsilon$$

Incluye un término de error ε (aleatorio y no observable) que contempla todos los posibles factores generadores de discrepancias entre los valores observados y los valores estimados por el modelo: selección de variables, selección de individuos, selección de la ecuación, errores de diseño y errores de medición.

Modelo lineal

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Por lo general, esta ecuación lineal constituye una buena aproximación a la verdadera relación

Las constantes $\beta_1, \beta_2, ... \beta_k$, se denominan parámetros, son desconocidas y se estiman a partir de información reportada por observaciones de las variables

Al efectuar n observaciones de las k+1 variables, las relaciones entre ellas quedan descritas mediante el conjunto de n ecuaciones:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \qquad (i = 1,2,...,n)$$

que en forma matricial pueden escribirse como:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

o bien:

$$\mathbf{y_{nx1}} = \mathbf{X_{nxk}}\, \beta_{\mathbf{kx1}} + \varepsilon_{\mathbf{nx1}}$$

expresión que se conoce como modelo lineal general.

Ajuste del modelo

Al hallarse un estimador $\tilde{\beta}$ del vector de parámetros y sustituir en la ecuación anterior, se obtiene un ajuste:

$$\mathbf{y} = \mathbf{X}\tilde{\beta} + e$$

surgiendo un término de error (aleatorio y observable) que se denomina residuo.

Ajuste mínimo cuadrático

Es un procedimiento de estimación mediante el cual se obtiene el estimador **b**, que minimiza la suma de cuadrados de los residuos:

$$SCE = \sum_{i=1}^{n} e_i^2 = e^t e$$

donde:

$$e = \mathbf{y} - \mathbf{X}\tilde{\beta}$$

Al derivar SCE respecto de $\tilde{\beta}$ e igualando al vector nulo, se obtienen las ecuaciones normales:

$$\mathbf{X}^t\mathbf{X}\tilde{\beta} = \mathbf{X}^t\mathbf{y}$$

de donde se deduce el estimador mínimo cuadrático de β**:**

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

asumiendo que $\mathbf{X_{nxk}}$ es una matriz de rango k (k<n). (De ahí el nombre de modelo lineal de rango completo)

Vector de estimaciones

$$\tilde{\mathbf{y}} = \mathbf{Xb}$$

resultando que:

$$\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Xb} = \mathbf{(I - H)y} = \mathbf{Qy} = \mathbf{Q}\varepsilon$$

donde:

H = $\mathbf{X(X^tX)^{-1}X^t}$ (matriz "hat") es una matriz simétrica, idempotente y de rango (o traza) igual a k, y

Q = **I - H** es una matriz simétrica, idempotente y de rango (o traza) igual a n-k.

Nota: A **H** se denomina matriz hat ya que **Hy** = $\tilde{\mathbf{y}}$.

Varianza residual

$$s^2 = \frac{SCE}{n-k} = \frac{\mathbf{e^te}}{n-k} = \frac{\mathring{\mathbf{U}}^t\mathbf{Q}\mathring{\mathbf{U}}}{n-k}$$

Modelo con término constante

En el caso en que se incluye un término constante β_o, el modelo estadístico queda

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_k + \varepsilon$$

siendo este modelo un caso particular del modelo lineal general en el cual:

$$\mathbf{y}_{nx1} = \mathbf{j}_{nx1}\,\beta_o + \mathbf{X}_*\beta_* + \varepsilon$$

$$= (\mathbf{j}_{nx1,}\,\mathbf{X}_*)\begin{pmatrix} \beta_o \\ \boldsymbol{\beta}_* \end{pmatrix} + \varepsilon$$

$$= \mathbf{X}_{nx(k+1)}\,\beta_{(k+1)x1} + \varepsilon$$

es decir, en el que:

$$\mathbf{X}_{nx(k+1)} = (\mathbf{j}_{nx1,}\,\mathbf{X}_*) \qquad y \qquad \beta_{(k+1)x1} = \begin{pmatrix} \beta_o \\ \boldsymbol{\beta}_* \end{pmatrix}$$

donde $\mathbf{X}_*$ es la matriz que contiene la información relativa a las p variables explicativas, y β el vector de coeficientes de regresión correspondientes: $\beta_1, \beta_2, ... \beta_k$.
Resolviendo las ecuaciones normales para este caso, se obtiene:

$$b_o = \bar{y} - \bar{\mathbf{x}}_*^t\,\mathbf{b}_*$$

$$\mathbf{b}_* = (\mathbf{X}_*^t\mathbf{X}_*)^{-1}\mathbf{X}_*^t\mathbf{y} = \mathbf{S}_{xx}^{-1}\,\mathbf{s}_{xy}$$

El vector de estimaciones en este caso:

$$\tilde{\mathbf{y}} = \mathbf{j}_{nx1}\,b_o + \mathbf{X}_*\mathbf{b}_*$$

resultando que:

$$\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}_*\,b_* = (\mathbf{I} - \mathbf{H}_*)\,\mathbf{y} = \mathbf{Q}_*\,\mathbf{y} = \mathbf{Q}_*\,\boldsymbol{\varepsilon}$$

donde:

$\mathbf{H}_* = \mathbf{X}_*\,(\,\mathbf{X}_*^t\,\mathbf{X}_*\,)^{-1}\,\mathbf{X}_*^t$ es una matriz simétrica, idempotente y de rango (o traza) igual a k, y

$\mathbf{Q}_* = \mathbf{I} - \mathbf{H}_*$ es una matriz simétrica, idempotente y de rango (o traza) igual a n-k.

y la varianza residual:

$$s^2 = \frac{SCE}{n - k - 1} = \frac{\mathbf{e}^t\mathbf{e}}{n - k - 1} = \frac{\boldsymbol{\varepsilon}^t\mathbf{Q}_*\boldsymbol{\varepsilon}}{n - k - 1}$$

Modelo de regresión simple(caso bidimensional, k=1)

$$b_o = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s(x, y)}{s^2(x)} = r(x,y)\frac{s(y)}{s(x)}$$

$$s^2 = \frac{SCE}{n-2} = \frac{(n-1)s^2(y)(1-r^2)}{n-2}$$