Modelo con k regresores

Contents

1 Modelo y su estimación

Supongamos que se tiene el siguiente modelo de regresión lineal:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,\ i = 1, 2, \cdots, n$$

- El término y_i se conoce como *variable respuesta*, y las x se conocen como *regresores*.
- El término e_i representa el error del modelo.

La ecuación del modelo se puede escribir en notación matricial:

$$i = 1 \Rightarrow y_1 = b_0 + b_1 x_{11} + b_2 x_{21} + \cdots + b_k x_{k1} + e_1$$

$$i = 2 \Rightarrow y_2 = b_0 + b_1 x_{12} + b_2 x_{22} + \cdots + b_k x_{k2} + e_2$$

$$\cdots$$

$$i = n \Rightarrow y_n = b_0 + b_1 x_{1n} + b_2 x_{2n} + \cdots + b_k x_{kn} + e_n$$

Agrupando:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Finalmente:

$$y = XB + e$$

Esta ecuación es válida para cualquier número de regresores y cualquier número de observaciones.

En este caso, el vector de parámetros estimados es:

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix}$$

Los residuos, igual que en los apartados precedentes se calculan

$$e = y - \hat{y} = y - XB$$

La suma de residuos al cuadrado será:

$$RSS(B) = \sum e_i^2 = e^T e = y^T y - y^T XB - B^T X^T y + B^T X^T XB$$

El método de mínimos cuadrados consiste en minimizar dicha suma, con lo que se obtiene:

$$B = (X^T X)^{-1} X^T y$$

Todo lo presentado en los apartados precedentes es aplicable en este caso también.

2 Residuos

La ecuación del modelo se puede expresar como

$$y = XB + e = \hat{y} + e$$

Es decir, los datos y se descomponen en parte perteciente a la recta ($\hat{y} = XB$) y parte no perteneciente a la recta o residuos (e). Ambas se pueden calcular ahora ya que se conoce B.

2.1 Matriz H

Se define la matriz H como:

$$\hat{y} = XB = X(X^T X)^{-1} X^T y = Hy$$

La matriz $H = X(X^T X)^{-1} X^T$ se denomina en inglés *hat matrix*. Es sencillo comprobar que la matriz H es simétrica ($H^T = H$) e idempotente ($H \cdot H = H$).

Los residuos se pueden expresar en función de dicha matriz como:

$$e = y - \hat{y} = (I - H)y$$

Se suele utilizar para derivar resultados teóricos. Por ejemplo, utilizando esta matriz se puede demostrar que $\sum e_i^2 = y^T y - B^T (X^T y)$.

2.2 Ortogonalidad de residuos y regresores

Otra propiedad importante de los residuos es que $X^T e = 0$. Efectivamente, sustituyendo el valor de B en la ecuación del modelo

$$y = XB + e = X(X^T X)^{-1} X^T y + e$$

Multiplicando por la izquierda por X^T se obtiene

$$X^T y = (X^T X)(X^T X)^{-1} X^T y + X^T e \Rightarrow X^T y = X^T y + X^T e \Rightarrow X^T e = 0$$

Si excribimos dicha propiedad en función de las componentes de las matrices:

$$X^T e = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}^T \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

Este producto equivale a las siguientes ecuaciones:

$$\sum e_i = 0,\ \sum x_{1i} e_i = 0,\ \sum x_{2i} e_i = 0, \cdots,\ \sum x_{ki} e_i = 0$$

La primera ecuación indica que los residuos siempre suman cero. Las siguientes ecuaciones indican que el vector residuos es ortogonal a las columnas de la matriz X (consideradas estas columnas como vectores). Por tanto es ortogonal al espacio vectorial generado por dichos vectores.

3 El modelo en diferencias a la media

3.1 Modelo

Dada la ecuación del modelo

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,\ i = 1, 2, \cdots, n$$

Si sumamos en ambos miembros desde 1 hasta n

$$\sum y_i = \sum b_0 + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots + b_k \sum x_{ki} + \sum e_i$$

Teniendo en cuenta que los residuos suman cero

$$\sum y_i = n b_0 + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots + b_k \sum x_{ki}$$

Y dividiendo entre n

$$\bar{y} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \cdots + b_k \bar{x}_k$$

Si a la ecuación del modelo le restamos esta última ecuación se obtiene:

$$y_i - \bar{y} = b_1 (x_{1i} - \bar{x}_1) + b_2 (x_{2i} - \bar{x}_2) + \cdots + b_k (x_{ki} - \bar{x}_k) + e_i,\ i = 1, 2, \cdots, n$$

Estas n ecuaciones se pueden expresar en forma matricial de la misma forma que hicimos antes, obteniendo:

$$\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Que en este caso se expresa como

$$\tilde{y} = \tilde{X}\tilde{B} + e$$

donde $\tilde{B}$ es el vector de coeficientes del modelo excepto b_0.

3.2 Estimación del modelo utilizando matrices de covarianzas

Se puede demostrar que $\tilde{X}^T e = 0$, por lo que

$$\tilde{X}^T \tilde{y} = \tilde{X}^T \tilde{X} \tilde{B} + \tilde{X}^T e \Rightarrow S_{Xy} = S_{XX} B^*$$

$$\tilde{B} = S_{XX}^{-1} S_{Xy}$$

donde S_{Xy} es la matriz de covarianzas de X e y, y S_{XX} es la matriz de covarianzas de X:

$$S_{Xy} = \frac{1}{n-1} \tilde{X}^T \tilde{y} = \begin{bmatrix} S_{1y} \\ S_{2y} \\ \cdots \\ S_{ky} \end{bmatrix}$$

$$S_{XX} = \frac{1}{n-1} \tilde{X}^T \tilde{X} = \begin{bmatrix} S_{11} & S_{21} & \cdots & S_{k1} \\ S_{12} & S_{22} & \cdots & S_{k2} \\ \cdots & \cdots & \cdots & \cdots \\ S_{1k} & S_{2k} & \cdots & S_{kk} \end{bmatrix}$$

donde S_{ij} representa la covarianza entre x_i e x_j, y S_{iy} representa la covarianza entre x_i e y.

Las ecuaciones derivadas en este apartado constituyen una alternativa para la estimación de los coeficientes del modelo de regresión lineal.

A modo de resumen:

- Las matrices X e y son matrices de **datos**. Con ellas se pueden estimar los parámetros del modelo haciendo $B = (X^T X)^{-1} X^T y$.
- Las matrices S_{Xy} y S_{XX} son matrices de **covarianzas**. Con ellas se pueden estimar los parámetros del modelo haciendo $\tilde{B} = S_{XX}^{-1} S_{Xy}$.

3.3 Aplicación a los datos

Para comprobar su funcionamiento, vamos a aplicarlo al caso de dos regresores:

```
d = read.csv("datos/kidiq.csv")
```

```
y = matrix(d$kid_score, ncol = 1)
Xa = cbind(d$mom_iq, d$mom_age) # sin columna de unos!!!!
```

```r
(Sxy = cov(Xa,y))
```

```
##             [,1]
## [1,] 137.244279
## [2,]   5.071923
```

```r
(Sxx = cov(Xa))
```

```
##           [,1]     [,2]
## [1,] 225.00000 3.711610
## [2,]   3.71161 7.295777
```

```r
(B_a = solve(Sxx) %*% Sxy)
```

```
##           [,1]
## [1,] 0.6035720
## [2,] 0.3881286
```

Falta calcular b_0. Utilizamos la fórmula

$$b_0 = \bar{y} - (b_1\bar{x}_1 + b_2\bar{x}_2 + \cdots + b_k\bar{x}_k)$$

```r
( beta0_e = mean(d$kid_score) - colMeans(Xa) %*% B_a )
```

```
##          [,1]
## [1,] 17.59625
```

4 Ejercicios propuestos

1. Demostrar que la matriz H es simétrica e idempotente.
2. Utilizando la matriz H demostrar que $\sum e_i^2 = y^T y - B^T(X^T y)$.
3. Demostrar que $\tilde{X}^T e = 0$.
4. Demostrar que $\sum e_i^2 = (n-1)s_y^2 - (n-1)\tilde{B}^T S_{Xy}$.