Modelo con k regresores

Contents

1 Modelo y su estimación

Supongamos que se tiene el siguiente modelo de regresión lineal:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i, \ i = 1, 2, \cdots, n$$

- 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:

$$i = 1 \Rightarrow y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + e_1$$

$$i = 2 \Rightarrow y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + e_2$$

$$\cdots$$

$$i = n \Rightarrow y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + e_n$$

Agrupando:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Finalmente:

$$y = X\beta + e$$

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

En este caso, el vector de parámetros estimados es:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix}$$

Los residuos, igual que en los apartados precedentes se calculan

$$e = y - \hat{y} = y - X\beta$$

La suma de residuos al cuadrado será:

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

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

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

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

2 Residuos

La ecuación del modelo se puede expresar como

$$y = X\beta + e = \hat{y} + e$$

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

2.1 Matriz H

Se define la matriz H como:

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

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 la siguiente propiedad para la suma de los residuos al cuadrado:

$$\sum e_i^2 = e^T e = (y - Hy)^T (y - Hy) = y^T y - y^T Hy - y^T H^T y + y^T H^T Hy = y^T y - y^T H^T y$$

$$= y^T y - (Hy)^T y = y^T y - (X\beta)^T y = y^T y - \beta^T (X^T y)$$

Finalmente

$$\sum e_i^2 = y^T y - \beta^T (X^T y)$$

2.2 Ortogonalidad de residuos y regresores

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

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

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:

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

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i,\ i = 1, 2, \cdots, n$$

Si sumamos en ambos miembros desde 1 hasta n

$$\sum y_i = \sum \beta_0 + \beta_1 \sum x_{1i} + \beta_2 \sum x_{2i} + \cdots + \beta_k \sum x_{ki} + \sum e_i$$

Teniendo en cuenta que los residuos suman cero

$$\sum y_i = n\beta_0 + \beta_1 \sum x_{1i} + \beta_2 \sum x_{2i} + \cdots + \beta_k \sum x_{ki}$$

Y dividiendo entre n

$$\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k$$

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

$$y_i - \bar{y} = \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i, \;\; i = 1, 2, \cdots, n$$

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

$$\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

Que en este caso se expresa como

$$y_a = X_a \beta_a + e$$

donde β_a es el vector de coeficientes del modelo excepto β_0.

3.2 Estimación del modelo utilizando matrices de covarianzas

Se puede demostrar que $X_a^T e = 0$:

$$X_a^T e = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{bmatrix}^T \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix} = \begin{bmatrix} \sum (x_{1i} - \bar{x}_1) e_i \\ \sum (x_{2i} - \bar{x}_2) e_i \\ \cdots \\ \sum (x_{ki} - \bar{x}_k) e_i \end{bmatrix} = \begin{bmatrix} \sum x_{1i} e_i - \bar{x}_1 \sum e_i \\ \sum x_{2i} e_i - \bar{x}_2 \sum e_i \\ \cdots \\ \sum x_{ki} e_i - \bar{x}_k \sum e_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

Por tanto, partiendo de la ecuación en diferencias a la media:

$$y_a = X_a \beta_a + e$$

y multiplicando ambos miembros por X_a^T:

$$X_a^T y_a = X_a^T X_a \beta_a + X_a^T e \Rightarrow \beta_a = \left(X_a^T X_a \right)^{-1} \left(X_a^T Y_a \right) = \left(\frac{1}{n-1} X_a^T X_a \right)^{-1} \left(\frac{1}{n-1} X_a^T Y_a \right)$$

$$\beta_a = S_{XX}^{-1} S_{Xy}$$

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

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

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

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

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

3.3 Residuos

Los residuos en este modelo se definen como

$$e = y_a - X_a \beta_a$$

Se ha demostrado que $X_a^T e = 0$. Por último vamos a demostrar otra propiedad análoga a la obtenida con el modelo con matrices de datos:

$$\sum e_i^2 = e^T e = (y_a - X_a\beta_a)^T (y_a - X_a\beta_a) = y_a^T y_a - y_a^T X_a \beta_a - \beta_a^T X_a^T y_a - \beta_a^T X_a^T X_a \beta_a$$

$$= (n-1)s_y^2 - (n-1)S_{Xy}^T \beta_a - (n-1)\beta_a^T S_{Xy} + (n-1)(S_{XX}^{-1} S_{Xy})^T S_{XX} \beta_a$$

Finalmente:

$$\sum e_i^2 = (n-1)s_y^2 - (n-1)\beta_a^T S_{Xy}$$

3.4 Aplicación a los datos

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

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

```
# en primer lugar vamos a calcular las matrices de covarianzas con la función de R cov()
(S = cov(d))
```

```
##            kid_score    mom_hs     mom_iq  mom_work   mom_age
## kid_score 416.596205 1.9864731 137.244279 2.1105672 5.0719235
## mom_hs      1.986473 0.1687562   1.742053 0.1232267 0.2380403
## mom_iq    137.244279 1.7420527 225.000000 2.0344125 3.7116099
## mom_work    2.110567 0.1232267   2.034413 1.3956908 0.4326955
## mom_age     5.071923 0.2380403   3.711610 0.4326955 7.2957770
```

```
(Sxx = S[c(3,5),c(3,5)])
```

```
##            mom_iq  mom_age
## mom_iq  225.00000 3.711610
## mom_age   3.71161 7.295777
```

```
(Sxy = S[c(3,5),1] )
```

```
##     mom_iq    mom_age
## 137.244279   5.071923
```

```
(beta_a = solve(Sxx) %*% Sxy)
```

```
##              [,1]
## mom_iq  0.6035720
## mom_age 0.3881286
```

Vamos a comprobar que las matrices de covarianzas se pueden calcular a partirde X_a e Y_a:

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

```
n = nrow(d)
1/(n-1) * t(Xa) %*% Xa
```

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

```
1/(n-1) * t(Xa) %*% ya
```

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

Falta calcular β_0. Utilizamos la fórmula

$$\beta_0 = \bar{y} - (\beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k)$$

```
( beta0 = mean(d$kid_score) - colMeans(d[,c(3,5)]) %*% beta_a )
```

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

Por último, para la suma de los residuos al cuadrado:

```
# varianza de y
(sy2 = S[1,1] )
```

```
## [1] 416.5962
```

```
(RSS = (n-1)*sy2 - (n-1)*t(beta_a) %*% Sxy)
```

```
##          [,1]
## [1,] 143665.4
```