Modelos de Regresión y Series de Tiempo (MRST)
2025 - 02

Clase 8 – Forma matricial del MRLM y estimación de sus parámetros

Docente: Natalia Jaramillo Quiceno

Escuela de Ingenierías

natalia.jaramilloq@upb.edu.co

Regresión lineal múltiple
Forma matricial del MRLM

Suponga que tiene $n > k$ observaciones disponibles, donde:

$\boldsymbol{y_i}$ es la i-ésima observación de la variable respuesta

$\boldsymbol{x_{ij}}$ denota la i-ésima observación de la variable independiente x_j

Obs. (i)	**Variable respuesta (y)**	**Variables regresoras (x1, …, xk)**			
		x1	**x2**	**…**	**xk**
1	y_1	x_{11}	x_{12}	…	X_{1k}
2	y_2	x_{21}	x_{22}	…	X_{2k}
…	…				
n	y_n	x_{n1}	x_{n2}	…	X_{nk}

Así, el modelo muestral se puede escribir como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

Donde se supone que los residuales son independientes e idénticamente distribuidos, es decir:

$$\varepsilon_i \backsim N(0, \sigma^2)$$

- Se asume que las xj son fijas, lo cual es cierto para observaciones en diseños experimentales, pero no necesariamente en el caso de diseños observacionales.
- Es necesario que las observaciones de cada variable sean INDEPENDIENTES.

Regresión lineal múltiple

Forma matricial del MRLM

Definiendo los siguientes vectores y matrices…

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{\boldsymbol{n} \times \boldsymbol{1}} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(\boldsymbol{k}+\boldsymbol{1}) \times \boldsymbol{1}} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{\boldsymbol{n} \times (\boldsymbol{k}+\boldsymbol{1})} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{\boldsymbol{n} \times \boldsymbol{1}}$$

Entonces, la **notación matricial de MRLS** es:

$$\underset{\boldsymbol{n} \times \boldsymbol{1}}{\boldsymbol{y}} = \underset{\boldsymbol{n} \times \boldsymbol{1}}{\boldsymbol{X}\beta} + \underset{\boldsymbol{n} \times \boldsymbol{1}}{\varepsilon}$$

Regresión lineal múltiple

Estimación de parámetros por Mínimos Cuadrados

Así como para el MRLS, comenzamos definiendo la función de mínimos cuadrados:

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Se debe minimizar la función S respecto a β_0, β_1, …, β_k. Por ende, los estimadores de β_0, β_1, …, β_k por mínimos cuadrados deben satisfacer:

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \qquad \textbf{(1)}$$

y

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0, \quad j = 1, 2, \ldots, k \qquad \textbf{(2)}$$

Regresión lineal múltiple
Estimación de parámetros por Mínimos Cuadrados

Volviendo a la notación matricial….

Podemos escribir la función de mínimos cuadrados de la forma: $S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Al desarrollar esta expresión encontramos que: $S(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \underline{\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}} - \underline{\mathbf{y}'\mathbf{X}\boldsymbol{\beta}} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$

Escalar idéntico

$$S(\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Nuevamente, tenemos que se debe minimizar la función S respecto a $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$. Por ende, los estimadores de $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ por mínimos cuadrados deben satisfacer:

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

que se simplifica a

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

Ecuaciones normales de mínimos cuadrados

Al resolver se obtiene

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Estimador de $\boldsymbol{\beta}$ por mínimos cuadrados

Regresión lineal múltiple

Estimación de parámetros por Mínimos Cuadrados

Revisemos por un momento la forma matricial de las ecuaciones normales

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

$$\underset{(\boldsymbol{k}+\mathbf{1})\times(\boldsymbol{k}+\mathbf{1})}{\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix}} \underset{(\boldsymbol{k}+\mathbf{1})\times\mathbf{1}}{\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}} = \overset{(\boldsymbol{k}+\mathbf{1})\times\mathbf{1}}{\begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}}$$

Notar que:

- Los elementos diagonales de $\boldsymbol{X}'\boldsymbol{X}$ son las sumas de cuadrados de los elementos de las columnas de $\boldsymbol{X}$.
- Los elementos fuera de la diagonal son las sumas de los productos cruzados de las columnas de $\boldsymbol{X}$.
- Los elementos de $\boldsymbol{X}'\boldsymbol{y}$ son las sumas de los productos cruzados de $\boldsymbol{X}$ por las observaciones $\boldsymbol{y_i}$.

Regresión lineal múltiple

Estimación de parámetros por Mínimos Cuadrados

Así, el modelo de regresión que corresponde a los niveles de las variables regresoras $\boldsymbol{x}' = [1, x_1, x_2, \dots, x_k]$ es

$$\hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

Y el vector de valores ajustados $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ que corresponde a los valores observados $\boldsymbol{y}_{\boldsymbol{i}}$ es:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

A la matriz $n \times n$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'$, se le suele llamar **matriz sombrero**.

Finalmente, los ***n*** residuales se pueden obtener también de manera matricial como sigue:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \longrightarrow \mathbf{e} = \underline{\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Regresión lineal múltiple
Ejemplo en RStudio

Se tiene la siguiente información de una localidad turística sobre los turistas extranjeros llegados de 5 países de procedencia:

Observación (i)	**Y** Número de turistas	X_1 Ingresos medios anuales (miles de euros)	X_2 Distancia (cientos de km)
1	18	5	17
2	25	10	15
3	7	2	32
4	12	4	25
5	19	6	20

- Exprese de forma matricial las ecuaciones normales $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$
- Determine el vector solución para $\hat{\boldsymbol{\beta}}$
- Presente el modelo de regresión lineal ajustado

Base de datos

08 – turistas.txt

Regresión lineal múltiple
Seguimiento – Clase 8

Volviendo al caso estudiado en el seguimiento de la clase 7 (base de datos 07 – DB riesgo 7.txt)…

A partir del muestreo de 113 hospitales en Estados Unidos, se evaluó la incidencia de diferentes factores en la probabilidad de que un paciente adquiera una infección mientras está hospitalizado. Las variables analizadas fueron:

Variable Respuesta (y): riesgo de infección en porcentaje **(riesgo)**

Potencial predictor (x_1): tiempo de hospitalización promedio de los pacientes **(tiempo)**

Potencial predictor (x_2): edad promedio de los pacientes **(edad)**

Potencial predictor (x_3): índice de rayos X realizados **(tasarayosx)**

1. Genere los vectores y matrices requeridos y responda los siguientes puntos:

- Exprese de forma matricial las ecuaciones normales $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$
- Determine el vector solución para $\hat{\boldsymbol{\beta}}$
- Presente el modelo de regresión lineal ajustado

Base de datos
07 – riesgo.txt

2. Verifique sus resultados utilizando la función lm() de Rstudio.

MUCHAS GRACIAS

Natalia Jaramillo Quiceno
e-mail: natalia.jaramilloq@upb.edu.co