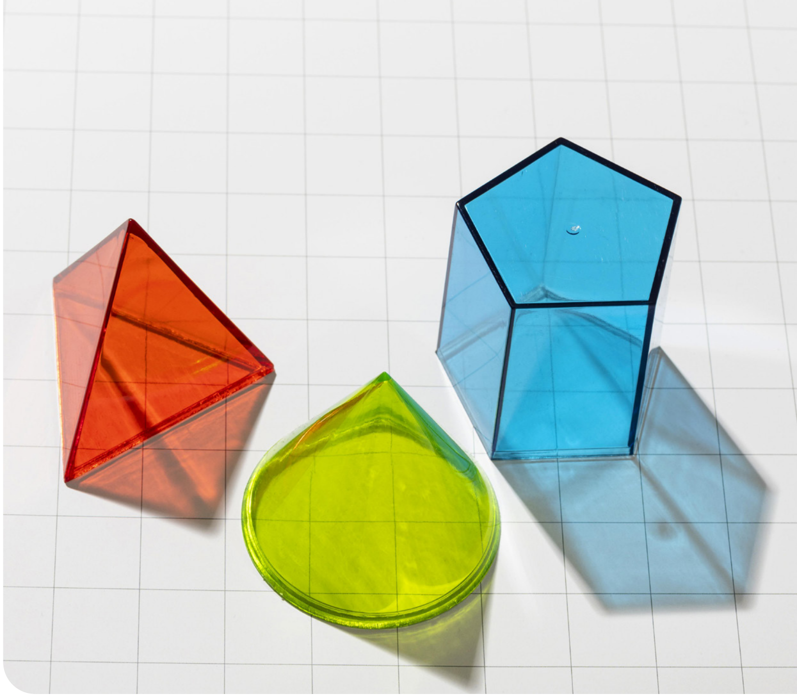
CIENCIA BÁSICA Y CULTURA

Boletín de Ciencias Básicas

Año 2024 Número 8 22 de abril

Regresión lineal simple a partir del teorema de proyección

Juan Gustavo Rueda Escobedo
(Coordinación de matemáticas de la DCB)



En las ciencias experimentales es usual ajustar modelos a partir de datos obtenidos del fenómeno que se estudia. Un caso típico es el tratar de ajustar datos a un modelo lineal de la forma $y = mx + b$, donde x es la variable independiente, y la variable dependiente, y m y b son parámetros que hay que determinar. Si se tienen n parejas de datos $\{x_i, y_i\}$, los valores de m y b que minimizan la distancia de la recta a los puntos están dados por [2, Sec. 2.2.1]:

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad ; \quad b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Estas expresiones generalmente son referidas como fórmulas de regresión lineal simple y es común que se enseñen sin dar mucho contexto sobre cómo se obtienen, debido a que entender su deducción habitual requiere haber cursado un módulo de estadística y uno o dos de cálculo. Sin embargo, también es posible llegar a ellas a partir del álgebra lineal, asignatura que se toma, por lo regular, en el segundo semestre de una carrera.

Al tener n pares $\{x_i, y_i\}$ y la relación $mx + b = y$ podemos imaginar que hay un operador lineal L que va de ~^{2} a ~^{n} que genere los valores y_i. Aquí, $\{m,b\}$ está en el dominio de la operación y la colección de mediciones $\{y_1, y_2,..., y_n\}$ está en el codominio. Esto puede representarse de la siguiente forma:

$$\underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}}_{=A} \underbrace{\begin{bmatrix} m \\ b \end{bmatrix}}_{=v} = \underbrace{\begin{bmatrix} y_1 & 1 \\ y_2 & 1 \\ \vdots & \vdots \\ y_n & 1 \end{bmatrix}}_{=W}$$

Así, se busca al vector $v = \{m,b\}$ que produce a las mediciones representadas por el vector $v = \{y_1, y_2,..., y_n\}$ a través del operador L, que en este caso se representa por medio de la matriz A. Si el fenómeno es perfectamente lineal y los datos no han sido corrompidos por ruido de medición, $\{y_1, y_2,..., y_n\}$ debe estar en el espacio rango de L (denotado aquí por ~{L}. Equivalentemente, w debe estar en el espacio columna de A. Sin embargo, lo normal es que este no sea el caso, resultando en un sistema de ecuaciones lineales incompatible, i.e., no existe solución para v. Ante esta situación, lo que se puede hacer es buscar el elemento de ~{L} más cercano a w. Esto se puede lograr proyectando a w sobre ~{L}. Si w_{proy} es la proyección de w sobre ~{L}, entonces, la ecuación $Av = w_{proy}$ es compatible y tiene al menos una solución. Además, si al menos hay una $i,j \epsilon$ {1,2,...,n} con $i \neq j$, tal que $x_i \neq x_j$, la solución será única, ya que la matriz A define un mapa inyectivo[1]. Dicha solución es la deseada pues proporciona los valores de m y n que resultan en la recta más cercana a las mediciones $\{y_1, y_2,..., y_n\}$[2].

La solución propuesta puede obtenerse por medio del siguiente método[3]:

1. Primero, se debe obtener una base ortonormal para ~{L} .
2. Segundo, usando la base ortonormal, se calcula la proyección de w sobre ~{L}.
3. Finalmente, se resuelve la ecuación $Av = w_{proy}$.

[1] Esto sucede porque las columnas de A serán linealmente independientes resultando en *rank* (A) = *dim* (~^{2}) = 2. El único caso donde esto no sucede es cuando $x_1 = x_2 = ... = x_n$, ya que ambas columnas resultan paralelas.
[2] Recuerde que la proyección de un vector v sobre un subespacio resulta en el elemento del subespacio más cercano a v. Véase [1, Prop. 6.61].
[3] La manera de obtener una base ortonormal puede ser consultada en [1, Prop. 6.32]. El cálculo de la proyección se explica en [1, Prop. 6.55]. Finalmente, cómo cambia la matriz asociada a un operador al modificar la base, puede ser revisado en [1, Prop. 3.84].

Diseño y formación editorial: L.D.G. Luis Enrique Vite Rangel

A continuación, se procederá de dicha manera.

Paso 1. Para obtener una base ortonormal, aplicaremos el procedimiento de Gram-Schmidt a las columnas de la matriz A. Por simplicidad en el cálculo, se toma como primer vector a la segunda columna de A. Al normalizar se obtiene:

$$\mathbf{e}_1 = \frac{1}{\sqrt{n}}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$$

Para calcular al segundo vector de la base, restamos de la primera columna su proyección sobre $\mathbf{e}_1$ y procedemos a normalizar. Primero, se resta de la segunda columna su proyección:

$$\bar{\mathbf{e}}_2 = \begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} - \left\langle \begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}, \frac{1}{\sqrt{n}}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}\right\rangle \frac{1}{\sqrt{n}}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = \begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} - \frac{\sum_{i=1}^{n} x_i}{n}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = \frac{1}{n}\begin{bmatrix}nx_1 - \sum_{i=1}^{n} x_i\\ nx_2 - \sum_{i=1}^{n} x_i\\ \vdots \\ nx_n - \sum_{i=1}^{n} x_i\end{bmatrix};$$

luego se calcula su norma:

$$\|\bar{\mathbf{e}}_2\| = \frac{1}{n}\sqrt{\left(nx_1 - \sum_{i=1}^{n} x_i\right)^2 + \left(nx_2 - \sum_{i=1}^{n} x_i\right)^2 + \cdots + \left(nx_n - \sum_{i=1}^{n} x_i\right)^2} = \frac{1}{\sqrt{n}}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$

por último, se normaliza, resultando así en el segundo vector de la base:

$$\mathbf{e}_2 = \frac{\bar{\mathbf{e}}_2}{\|\bar{\mathbf{e}}_2\|} = \frac{\frac{1}{\sqrt{n}}}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}\begin{bmatrix}nx_1 - \sum_{i=1}^{n} x_i\\ nx_2 - \sum_{i=1}^{n} x_i\\ \vdots \\ nx_n - \sum_{i=1}^{n} x_i\end{bmatrix}$$

Al tener la base ortonormal $B_{\perp} = \{e_1, e_2\}$ se expresarán todas las demás cantidades con respecto a ella.

Paso 2. Ahora, se procede a calcular la proyección de $\mathbf{w}$ sobre ~{L}. La proyección se obtiene de la siguiente manera:

$$\mathbf{w}_{proy} = \langle \mathbf{w}, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \mathbf{w}, \mathbf{e}_2\rangle \mathbf{e}_2$$

Se pondrá el foco en obtener a $\langle \mathbf{w}, \mathbf{e}_1\rangle$ y a $\langle \mathbf{w}, \mathbf{e}_2\rangle$. Estos términos resultan en:

$$\langle \mathbf{w}, \mathbf{e}_1\rangle = \frac{y_1 + y_2 + \cdots + y_n}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} y_i \quad ;$$

$$\langle \mathbf{w}, \mathbf{e}_2\rangle = \frac{\sqrt{n}}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}\left(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i\right)$$

Nótese que los términos $\langle \mathbf{w}, \mathbf{e}_1\rangle$ y $\langle \mathbf{w}, \mathbf{e}_2\rangle$ son las coordenadas de $\mathbf{w}_{proy}$ con respecto a la base $B_{\perp}$, así que estas cantidades son lo único que se requiere para describir a $\mathbf{w}_{proy}$.
Paso 3. Para poder resolver la ecuación $A\mathbf{v} = \mathbf{w}_{proy}$, se expresará en relación con la base $B_{\perp}$. Se comenzará con la matriz A. Denote a la primera columna de A como ¥ y a la segunda como 1. Así se tiene que:

$$\chi = \begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} = \langle \chi, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \chi, \mathbf{e}_2\rangle \mathbf{e}_2 \quad ; \quad 1 = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = \langle 1, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle 1, \mathbf{e}_2\rangle \mathbf{e}_2$$

A continuación, se calculan las coordenadas de x y 1 con respecto a la base $B_{\perp}$:

$$\langle \chi, \mathbf{e}_1\rangle = \frac{\sum_{i=1}^{n} x_i}{\sqrt{n}} \quad ; \quad \langle \chi, \mathbf{e}_2\rangle = \frac{\frac{1}{\sqrt{n}}\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} = \frac{1}{\sqrt{n}}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\langle 1, \mathbf{e}_1\rangle = \frac{n}{\sqrt{n}} = \sqrt{n} \quad ; \quad \langle 1, \mathbf{e}_2\rangle = \frac{\sqrt{n}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}} = 0$$

De esta manera, el lado izquierdo de la ecuación puede expresarse equivalentemente de la siguiente forma:

$$\begin{aligned} A\mathbf{v} = \chi m + 1b &= \left(\langle \chi, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \chi, \mathbf{e}_2\rangle \mathbf{e}_2\right)m + \left(\langle 1, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle 1, \mathbf{e}_2\rangle \mathbf{e}_2\right)b \\ &= \left(\langle \chi, \mathbf{e}_1\rangle m + \langle 1, \mathbf{e}_1\rangle b\right)\mathbf{e}_1 + m\langle \chi, \mathbf{e}_2\rangle \mathbf{e}_2 \end{aligned}$$

Por otra parte, el lado derecho de la ecuación corresponde a:

$$\mathbf{w}_{proy} = \langle \mathbf{w}, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \mathbf{w}, \mathbf{e}_2\rangle \mathbf{e}_2$$

Al combinar ambos lados se llega a la siguiente expresión:

$$\left(\langle \chi, \mathbf{e}_1\rangle m + \langle 1, \mathbf{e}_1\rangle b\right)\mathbf{e}_1 + m\langle \chi, \mathbf{e}_2\rangle \mathbf{e}_2 = \langle \mathbf{w}, \mathbf{e}_1\rangle \mathbf{e}_1 + \langle \mathbf{w}, \mathbf{e}_2\rangle \mathbf{e}_2$$

Se igualan los coeficientes de cada elemento de la base $B_{\perp}$ y se obtiene el siguiente sistema de ecuaciones:

$$\begin{aligned} \langle \chi, \mathbf{e}_1\rangle m + \langle 1, \mathbf{e}_1\rangle \bar{b} &= \langle \mathbf{w}, \mathbf{e}_1\rangle \\ m\langle \chi, \mathbf{e}_2\rangle &= \langle \mathbf{w}, \mathbf{e}_2\rangle \end{aligned}$$

Se resuelve para m y para b y se tiene que:

$$m = \frac{\langle \mathbf{w}, \mathbf{e}_2\rangle}{\langle \chi, \mathbf{e}_2\rangle}$$

$$b = \frac{\langle \mathbf{w}, \mathbf{e}_1\rangle}{\langle 1, \mathbf{e}_1\rangle} - \frac{\langle \chi, \mathbf{e}_1\rangle\langle \mathbf{w}, \mathbf{e}_2\rangle}{\langle 1, \mathbf{e}_1\rangle\langle \chi, \mathbf{e}_2\rangle} = \frac{\langle \mathbf{w}, \mathbf{e}_1\rangle\langle \chi, \mathbf{e}_2\rangle - \langle \chi, \mathbf{e}_1\rangle\langle \mathbf{w}, \mathbf{e}_2\rangle}{\langle 1, \mathbf{e}_1\rangle\langle \chi, \mathbf{e}_2\rangle}$$

La sustitución de los coeficientes en la expresión para m resulta en:

$$m=\frac{\langle w,e_2\rangle}{\langle \chi,e_2\rangle}=\frac{\frac{1}{\sqrt{n}}\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)}{\frac{1}{\sqrt{n}}\left(n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right)}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

lo cual corresponde a la expresión inicialmente dada para m . Para poder obtener una expresión para b , primero se requieren las siguientes relaciones:

$$\langle 1,e_1\rangle\langle \chi,e_2\rangle=\sqrt{n}\cdot\frac{1}{\sqrt{n}}\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$\langle w,e_1\rangle\langle \chi,e_2\rangle=\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}y_i\right)\left(\frac{1}{\sqrt{n}}\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\right)=\frac{1}{n}\sum_{i=1}^{n}y_i\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

$$\begin{aligned}\langle \chi,e_1\rangle\langle w,e_2\rangle&=\left(\frac{\sum_{i=1}^{n}x_i}{\sqrt{n}}\right)\left(\frac{\frac{1}{\sqrt{n}}}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}}\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)\right)\\&=\frac{1}{n}\cdot\frac{\sum_{i=1}^{n}x_i\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}}\end{aligned}$$

$$\begin{aligned}\langle w,e_1\rangle\langle \chi,e_2\rangle-\langle \chi,e_1\rangle\langle w,e_2\rangle&=\frac{1}{n}\sum_{i=1}^{n}y_i\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}-\frac{1}{n}\frac{\sum_{i=1}^{n}x_i\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}}\\&=\frac{1}{n}\cdot\frac{\sum_{i=1}^{n}y_i\left(n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right)-\sum_{i=1}^{n}x_i\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}}\\&=\frac{\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}}\end{aligned}$$

Así, al sustituir en la expresión para b , se llega a:

$$b=\frac{\langle w,e_1\rangle\langle \chi,e_2\rangle-\langle \chi,e_1\rangle\langle w,e_2\rangle}{\langle 1,e_1\rangle\langle \chi,e_2\rangle}=\frac{\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}x_iy_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}$$

De esta manera, se recuperan las fórmulas con las que se dio comienzo a este texto.

A pesar de que la solución presentada parezca compleja, sólo requiere de álgebra para seguirse. Además, es muy fácil de generalizar a problemas de mayor dimensión, ya que la estrategia de tres pasos planteada es aplicable a cualquier problema de la forma Av=w, con A y w conocidas y v el vector que se desea encontrar. De hecho, esta estrategia está en el centro del cálculo de la pseudo inversa de Moore-Penrose, uno de los resultados más destacados del álgebra lineal, el cual se abordará en un futuro número de este boletín.

Referencias

[1] Axler, S. (2024). Linear Algebra Done Right (4ta ed.). Springer Nature. https://doi.org/10.1007/978-3-031-41026-0

[2] Montgomery, D.C., Peck, E.A., & Vining, G.G. (2006). Introducción al Análisis de Regresión Lineal (3ra ed.). Compañía Editorial Continental.

“El gran libro de la naturaleza está escrito con símbolos matemáticos”

Galileo Galilei

“Para nosotros los físicos, creer en la separación entre el pasado, presente y futuro solo es una ilusión, aunque una muy convincente”

Albert Einstein

“La química comienza en las estrellas. Las estrellas son la fuente de los elementos químicos, que son los componentes básicos de la materia”

Peter Atkins

“Podemos afirmar que han sido la ingeniería y la tecnología las que han permitido el avance de la sociedad humana”.

Carlos Slim

“El ingeniero ideal es un compuesto... No es un científico, no es un matemático, no es un sociólogo ni un escritor; pero puede usar el conocimiento y las técnicas de cualquiera o todas estas disciplinas para resolver problemas de ingeniería”

Nathan W. Dougherty

“Cuando una puerta se cierra, otra se abre, pero a menudo vemos tanto tiempo y con tanta tristeza la puerta que se cierra que no notamos otra que se ha abierto para nosotros”

Alexander Graham Bel

Diseño y formación editorial: L.D.G. Luis Enrique Vite Rangel