El método de mínimos cuadrados

Pablo Andrés Úsuga Gómez, Juan Pablo Robledo Meza, Camilo Montoya Arango
Mateo Sebastián Mora Montero, Daniel Andrés Hernandez Pedraza

Enero 18 de 2025

Contexto del problema: Suponga que tiene un conjunto de datos numéricos, $D = \{y_1, y_2, ..., y_n\}$ donde cada y_i está asociado a un vector de p variables independientes $x_i = (x_{i1,}, x_{i2,}, ..., x_{ip})$, donde la notación indica que dicho vector de p valores puede variar con cada i. Suponga que tiene un modelo matemático $Y(x, \theta)$ que es una función de un vector $x \in \mathbb{R}^p$, y de un vector de parámetros θ en cierto espacio parametral $E[\theta]$.De todos esos posibles parámetros posibles que puede asignar al modelo, usted desea encontrar los que mejor se ajusten a sus datos iniciales. La pregunta es: ¿ qué queremos decir por mejor ajuste ?, matemáticamente necesitamos establecer una medida con la cual aclaremos dicho término, entonces se define una métrica que permita medir la diferencia entre cada dato inicial y_i con la información del modelo evaluado en el vector de dependencias x_i. Esta métrica en principio puede ser arbitraria, pero usualmente son métricas inducidas por normas, especialmente las normas euclidianas. Ahora que tenemos una manera de medir, es necesario establecer un criterio de optimalidad extra para determinar los parámetros del modelo que optimicen esta función que planteamos. Digamos que d representa la función distancia entre cada dato y el resultado del modelo, entonces un criterio de optimalidad arbitrario sería cualquier función F de las distancias de todos los datos con el modelo $F(D, Y(\theta)) = F(d(y_1, Y(x_1, \theta)), d(y_2, Y(x_2, \theta), .., d(y_n, Y(x_n, \theta)))$, y con esta, la pregunta formalmente es encontrar un conjunto de parámetros $\theta^* \in E[\theta]$ que minimice dicha función. El criterio o método de mínimos cuadrados consiste en utilizar la métrica estándar del valor absoluto, y el objetivo es minimizar la función

$$F(D, Y(\theta)) = F(d(y_1, Y(x_1, \theta)), d(y_2, Y(x_2, \theta)), .., d(y_n, Y(x_n, \theta))) = \sum_{i=1}^{n} d(y_i, Y(x_i, \theta))^2 = \sum_{i=1}^{n} |y_i - Y(x_i, \theta)|^2$$

y por tanto el problema se traduce en encontrar θ^* de modo que se minimice la suma de las distancias cuadradas de los datos iniciales y el modelo evaluado en cada vector de dependencia, o sea, hallar

$$\theta^* \in E[\theta] : min_{\theta \in E[\theta]}(F(D, Y(\theta))) = F(D, Y(\theta^*))$$

Ahora presentamos algunos ejemplos concretos para ilustrar el razonamiento anterior. Un caso fácil de analizar es tener un conjunto de n pares ordenados $P = \{(x_i, y_i) : 1 \leq i \leq n\}$, que corresponde a los datos iniciales, y digamos que queremos ajustar dichos datos a un modelo lineal de la forma $Y = mx + b$, los parámetros de este modelo son $\theta = (m, b)$. Tomemos la métrica usual entre dos valores reales, $d(y_i, Y(x_i, \theta)) = |y_i - (mx_i + b)|$. Entonces el problema de mínimos cuadrados para este modelo consiste en hallar los parámetros $\theta^* = (m^*, b^*)$ tales que se minimice la suma de los cuadrados de las distancias

$$F(P, Y(\theta)) = \sum_{i=1}^{n} d(x_i, Y(x_i, \theta))^2 = \sum_{i=1}^{n} |x_i - (mx_i + b)|^2$$

De modo más general consideremos el modelo de regresión lineal múltiple que se emplea en ciencia de datos. Los datos iniciales son unos valores reales $\{y_1, y_2, ..., y_n\}$, el modelo es $y(x, B) = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$, donde x es un vector de dimensión k y B el vector de parámetros de dimensión $k+1$. Queremos minimizar la suma de los cuadrados de las distancias, esto es :

$$D = \sum_{i=1}^{n} |y_i - Y(x_i, B)|^2 = \sum_{i=1}^{n} |y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + .. + b_kx_{ik})|^2$$

Para resolver este problema de optimización vamos a reinterpretarlo en el contexto del álgebra lineal. Los datos D podemos verlos como un vector $y \in \mathbb{R}^n$, además, considere la matriz de los coeficientes de las variables independientes

$$A = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \tag{1}$$

Con el vector columna de parámetros $B = [b_0, b_1, .., b_k]$, el producto $Y = XB$ resulta ser el modelo al que deseamos ajustar los datos. Note que $Y \in \mathbb{R}^n$, y es el producto de dicha matriz por un vector B, luego Y pertenece al espacio columna de la matriz X, denotado $Col(X)$. De este modo podemos reformular el problema de minimización de modo equivalente a encontrar un vector $y^* \in Col(X)$ tal que minimice $||y - y^*||^2$. Dicha norma es la euclidiana. Note que $Col(X)$ es un subespacio vectorial de $\mathbb{R}^n$, pues corresponde al generado de los vectores columna de X. Vamos a introducir los conceptos de complemento ortogonal y proyección ortogonal para dar solución a este problema de álgebra lineal.

Complemento ortogonal: Dado $U \subseteq \mathbb{R}^n$, el complemento ortogonal de U en $\mathbb{R}^n$ es el conjunto de todos los vectores $v \in \mathbb{R}^n$ que son ortogonales a cada vector de U, esto es :

$$U^{\perp} := \{v \in \mathbb{R}^n : \forall u \in U, v \cdot u = 0\}$$

Donde el producto punto es el usual en $\mathbb{R}^n$.

Teorema de descomposición ortogonal: Si $U \subseteq \mathbb{R}^n$ es un subespacio vectorial, entonces para todo $v \in \mathbb{R}^n$, existen únicos vectores $u \in U$, $w \in U^{\perp}$ tales que $v = u + w$.

Con este teorema podemos definir la proyección ortogonal del siguiente modo:

Proyección ortogonal : Sea $U \subseteq \mathbb{R}^n$ un subespacio vectorial. Dado $v \in \mathbb{R}^n$ definimos la proyección ortogonal de v sobre U como el vector $u \in U$ que se enuncia en el teorema de descomposición ortogonal. Y lo denotamos por $P_U(v) := u$.

Ahora probamos que la proyección ortogonal es el vector que minimiza la distancia a un subespacio.

Distancia mínima a un subespacio vectorial: Suponga que $U \subseteq \mathbb{R}^n$ es un subespacio vectorial, $v \in \mathbb{R}^n$ y $u \in U$. Entonces

$$||v - P_U(v)|| \leq ||v - u||$$

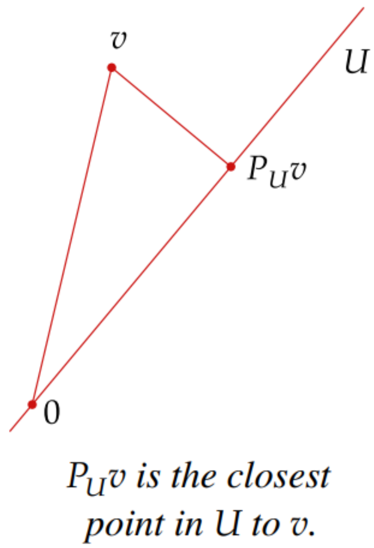

Figura 1: Proyección ortogonal $P_U(v)$. [4]

En particular, la proyección ortogonal de v sobre U minimiza la distancia a dicho subespacio.

Prueba: Como $0 \leq ||P_U(v) - u||$, $v - P_U(v) \in U^{\perp}$ y $P_U(v) - u \in U$ entonces, por teorema de Pitágoras se sigue:

$$\begin{aligned} ||v - P_U(v)||^2 &\leq ||v - P_U(v)||^2 + ||P_U(v) - u||^2 \\ &= ||(v - P_U(v)) + (P_U(v) - u)||^2 \\ &= ||v - u||^2 \end{aligned}$$

Tomando raíces se tiene lo pedido: $||v - P_U(v)|| \leq ||v - u||$

Con esto sabemos que el vector proyección es el que minimiza la distancia a nuestro problema, ahora vamos a encontrar las ecuaciones normales para el vector de parámetros B. Como $y = P_{Col(X)}(y) + e = XB + e$, donde $e \in Col(X)^{\perp}$, entonces el vector e es ortogonal a cada columna de X, y así, $X^T e = 0$, sustituyendo $e = y - XB$ se tiene $X^T(y - XB) = 0$, donde X^T es la transpuesta de X. De este modo obtenemos las ecuaciones normales para el vector de parámetros B.

$$X^T y = X^T X B$$

Y para obtener los b_i se soluciona dicho sistema de ecuaciones. Si existe la matriz inversa $(X^T X)^{-1}$, dicho sistema tendrá única solución, en otro caso, se suele usar pseudoinversa u otro método numérico para hallar una solución aproximada a B.

Contexto ciencia de datos: Tanto la regresión lineal como el método de mínimos cuadrados juegan un papel muy importante en el mundo del análisis de datos, en primer lugar esto se debe a que ayudan a determinar relaciones entre

variables, más específicamente entre la variable dependiente (o resultado) y una o más variables independientes (o explicativas) donde comúnmente se usan una o dos variables independientes para que combinado con herramientas de visualización faciliten el análisis de los resultados; además, la regresión lineal es uno de los métodos más sencillos y rápidos para modelar datos, por eso sirve como una excelente herramienta introductoria para aprender los conceptos claves en la ciencia de datos como la interpretación de coeficientes, análisis de residuos y validación de modelos.
Otro punto muy importante de estos métodos en el análisis de datos está cuando se quiere usar la analítica predictiva ya que el método de mínimos cuadrados minimiza la suma de los errores al cuadrado entre los datos observados y los predichos, esto permite generar modelos precisos y rápidos para tratar de predecir eventos, por ejemplo si tenemos el historial de llegada de pacientes a un hospital día a día, podemos tratar de predecir cuantos pacientes podrían llegar en un día futuro, esto ayudaría a preparar mejor los implementos y el personal del hospital para minimizar costos y mejorar la atención.

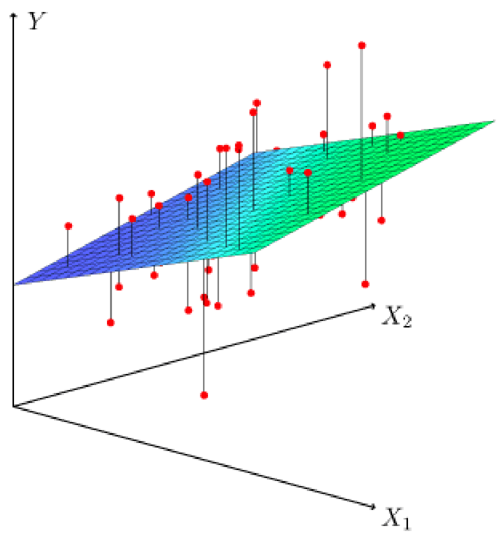

Figura 2: Regresión lineal con dos variables indepedendientes. [5]

Referencias

1. Datta B. N. *Numerical Linear Algebra and Applications* 2nd ed, (2010). Society for Industrial and Applied Mathematics.

2. Burden, R.L. and Faires, J.D. *Numerical Analysis*, (2011). 9th Edition, Brookscole, Boston.

3. Hoffman, J. D., Frankel, S. *Numerical methods for engineers and scientists*, (2018). CRC press.

4. Sheldon Axler. *Linear Algebra Done Right*, (2024). Springer

5. `https://statsandr.com/blog/multiple-linear-regression-made-simple/`