Universidad Carlos III de Madrid

Mínimos cuadrados y análisis de regresión

Para entender el problema de la regresión por medio del ajuste por mínimos cuadrados, veamos un ejemplo concreto y muy sencillo, para luego extender el resultado a una situación más general.

Dados los puntos con coordenadas (1, 2) ; (2, 4) y (3, 4), que, evidentemente, no están alineados, pero se pretenden "trazar" una recta que se ajuste lo mejor posible a ellos, cuya ecuación será:

$$y = a_0 + a_1 x$$

y de la que se pretende hallar el valor de los parámetros a_0 y a_1.

El planteamiento del problema mediante un sistema de ecuaciones, sería el siguiente:

$$\begin{cases} a_0 + 1a_1 = 2 \\ a_0 + 2a_1 = 4 \\ a_0 + 3a_1 = 4 \end{cases}$$

que escrito en forma matricial y vectorial será:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 4 \end{bmatrix} \Leftrightarrow a_0 c_1 + a_1 c_2 = b \Leftrightarrow Ax = b$$

Evidentemente este sistema será incompatible (en caso general también), a no ser que los tres puntos estuvieran alineados (pero en ese caso no tendría sentido la regresión).

Desde un punto de vista puramente algebraico, podemos entender el problema, diciendo que el vector columna de los términos independientes, b, de $\mathbb{R}^3$, no pertenece al subespacio vectorial generado por los dos vectores columna de la matriz A, que suele designarse por *ColA*:

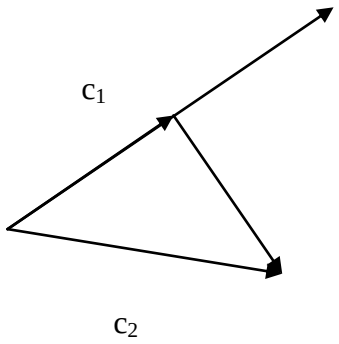
$$b \notin Gen(c_1, c_2) = ColA$$

La búsqueda de la "mejor" solución (aproximada) del problema, se basa en la consideración de que "mejor" significará proyectar el vector b sobre el espacio *ColA*, que designaremos por $\tilde{b}$, es decir:

$$\tilde{b} = proy_{ColA}(b)$$

Y eso dará lugar a nuevo sistema compatible con una solución aproximada x, pero la mejor aproximación, en el sentido "*mínimos cuadrados*", esta última frase se denomina así, por minimizar la norma habitual (euclídea), que como se conoce desde siempre es la raíz de una suma de cuadrados (pero no tendría por qué ser así necesariamente).

Así que el problema se reduce a hallar proyecciones, pero no se trata de un problema directo ya que la proyección sobre un subespacio requiere la obtención previa de una base ortogonal. En nuestro caso los dos vectores columna no son ortogonales($\{c_1, c_2\}$ no es ortogonal, es decir, así se necesita una ortogonalización previa, que puede ser tal y como muestra la figura siguiente, dando lugar a los vectores:



$\{e_1, e_2\}$ que tendrán la siguiente relación con $\{c_1, c_2\}$

$$e_1 = c_1$$
$$e_2 = c_2 - proy_{c_1}\{c_2\}$$

Si denominamos por B a la matriz formada por las columnas de componentes los nuevos vectores, tendremos una base ortogonal, para luego proyectar el vector b sobre ella y sumando resultados obtener $\tilde{b}$, es decir:

$$\tilde{b} = proy_{ColB}(b) = proy_{e_1}(b) + proy_{e_2}(b)$$

Vayamos a los cálculos en nuestro caso concreto:

$$proy_{c_1}\{c_2\} = \left(\frac{c_1.c_2}{c_1.c_1}\right)c_1 = \frac{1.1+1.2+1.3}{1.1+1.1+1.1}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}2\\2\\2\end{bmatrix}$$

$$e_2 = c_2 - proy_{c_1}\{c_2\} = c_2 - e_1 = \begin{bmatrix}1\\2\\3\end{bmatrix} - \begin{bmatrix}2\\2\\2\end{bmatrix} = \begin{bmatrix}-1\\0\\1\end{bmatrix}$$

$$proy_{e_1}\{b\} = \left(\frac{e_1.b}{e_1.e_1}\right)e_1 = \frac{1.2+1.4+1.4}{1.1+1.1+1.1}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}10/3\\10/3\\10/3\end{bmatrix}$$

$$proy_{e_2}\{b\} = \left(\frac{e_2.b}{e_2.e_2}\right)e_2 = \frac{-1.2+0.4+1.4}{2}\begin{bmatrix}-1\\0\\1\end{bmatrix} = \begin{bmatrix}-1\\0\\-1\end{bmatrix}$$ luego

$$\tilde{b} = proy_{ColB}(b) = proy_{e_1}(b) + proy_{e_2}(b) = \begin{bmatrix}7/3\\10/3\\13/3\end{bmatrix}$$

Así que el sistema (ahora compatible) que debemos resolver será:

$$\begin{cases} a_0 + 1a_1 = 7/3 \\ a_0 + 2a_1 = 10/3 \\ a_0 + 3a_1 = 13/3 \end{cases}$$

Cuya solución es $a_0 = 4/3$, $a_1 = 1$, lo que da lugar a la recta de regresión siguiente:

$$y = \frac{4}{3} + x$$

Conviene aclarar inmediatamente, que este no es el procedimiento estándar, por ser demasiado laborioso. Después de una carga teórica importante, se puede deducir que ese resultado se obtiene de resolver el sistema siguiente (que será compatible):

$$\left(A^T A\right) x = A^T b$$

Y, por supuesto, ésto también es válido para cualquier otra regresión posible, por ejemplo, el ajuste de una serie de datos (n) a una parábola:

$$y = a_0 + a_1 x + a_2 x^2$$

con planteamiento similar, pero:

$$\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_n^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$