Lecture 28: Application – Data Line of best fit

Given a data set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ find

$$\text{the line of best fit} \quad y = a_{\mathrm{best}} x + b_{\mathrm{best}}.$$

Let

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} a \\ b \end{pmatrix}$$

so that

$$\mathbf{y} - A\mathbf{u} = \begin{pmatrix} y_1 - (a + bx_1) \\ \vdots \\ y_n - (a + bx_n) \end{pmatrix} \qquad \textit{PICTURE}$$

Goal: Minimize $\|\mathbf{y} - A\mathbf{u}\|$.

Let

$$W = \{A\mathbf{u} \mid \mathbf{u} = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2\} \text{ and}$$

$$\vec{s} \in \mathbb{R}^2 \quad \text{such that} \quad A\vec{s} = \text{proj}_W(\mathbf{y}).$$

PICTURE

Then $\|\mathbf{y} - A\vec{s}\|$ will be minimal if $\mathbf{y} - A\vec{s}$ is perpendicular to W.

So we want if $\mathbf{u} \in \mathbb{R}^2$ then $\langle \mathbf{y} - A\vec{s}, A\mathbf{u} \rangle = 0$.

So we want if $\mathbf{u} \in \mathbb{R}^2$ then $\mathbf{u}^t A^t(\mathbf{y} - A\vec{s}) = 0$.

So we want $A^t\mathbf{y} - A^tA\vec{s} = 0$.

So we want $A^tA\vec{s} = A^t\mathbf{y}$.

So we want

$$\vec{s} = (A^tA)^{-1}A^t\mathbf{y} = \begin{pmatrix} a_{\text{best}} \\ b_{\text{best}} \end{pmatrix}$$

and the line of best fit is $y = a_{\text{best}}x + b_{\text{best}}$.

Example IP14 Follow the above procedure. Given the data set $D = \{(-1,1),(1,1),(2,3)\}$ then

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}.$$

Then

$$A^t A = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix} \quad \text{and}$$

$$A^t \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix} \quad \text{and} \quad (A^t A)^{-1} = \tfrac{1}{14} \begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}.$$

So

$$\vec{s} = \frac{1}{14} \begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \frac{1}{14} \begin{pmatrix} 18 \\ 8 \end{pmatrix} = \begin{pmatrix} \frac{9}{7} \\ \frac{4}{7} \end{pmatrix}.$$

So the line of best fit is

$$y = \tfrac{9}{7} + \tfrac{4}{7} x.$$