Least Squares Regression Line

We want to find the values of a and b that minimise

$$\begin{aligned} D &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \\ &= \sum_{i=1}^{n} ((y_i - ax_i) - b)^2 \end{aligned} \tag{1}$$

Letting $u_i = y_i - ax_i$,

$$\begin{aligned} D &= \sum_{i=1}^{n} (u_i - b)^2 \\ &= \sum_{i=1}^{n} (u_i^2 - 2u_i b + b^2) \\ &= \sum_{i=1}^{n} u_i^2 - 2b\sum_{i=1}^{n} u_i + nb^2 \\ &= nb^2 - 2b\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} u_i^2 \end{aligned} \tag{2}$$

This equation is a quadratic in terms of b. Since $n > 0$, D is minimised when

$$\begin{aligned} b &= \frac{-(-2\sum u_i)}{2n} \\ &= \frac{1}{n}\sum_{i=1}^{n} u_i \\ &= \frac{1}{n}\sum_{i=1}^{n} (y_i - ax_i) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right) \\ &= \bar{y} - a\bar{x} \end{aligned} \tag{3}$$

Substituting this expression for b into the original expression for D, we obtain:

$$\begin{aligned} D &= \sum_{i=1}^{n} \left[y_i - (ax_i + (\bar{y} - a\bar{x}))\right]^2 \\ &= \sum_{i=1}^{n} \left[y_i - ax_i - \bar{y} + a\bar{x}\right]^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - a(x_i - \bar{x})\right]^2 \end{aligned} \tag{4}$$

$$\begin{aligned} D &= \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 - 2a(x_i - \bar{x})(y_i - \bar{y}) + a^2(x_i - \bar{x})^2\right] \\ &= a^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2a \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned} \tag{5}$$

This equation is a quadratic in terms of a. Since $\sum(x_i - \bar{x})^2 > 0$, D is minimised when

$$\begin{aligned} a &= \frac{-(-2\sum(x_i - \bar{x})(y_i - \bar{y}))}{2\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \end{aligned} \tag{6}$$

$$\therefore \quad a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b = \bar{y} - a\bar{x} \tag{7}$$