

The line of best fit via transformations

David G Radcliffe

In this note, we will show how transformations can be used to obtain a radically simple derivation of the equation of the line of best fit. Our approach also gives a simple geometric interpretation of the Pearson correlation coefficient.

Given a sequence of n points in the plane $(X_1, Y_1), \dots, (X_n, Y_n)$ we seek the linear equation $y = a + bx$ that approximates the points as closely as possible, in the sense that the sum of the squared residuals $E = \sum_{i=1}^n (Y_i - a - bX_i)^2$ is minimized.

We assume that not all of the points lie on a single horizontal or vertical line. In that case, we can apply a *transformation* to the points so that $\sum x_i = \sum y_i = 0$ and $\sum x_i^2 = \sum y_i^2 = 1$. The transformation is defined by

$$x_i = \frac{X_i - \bar{X}}{\sqrt{\sum (X_i - \bar{X})^2}} \quad \text{and} \quad y_i = \frac{Y_i - \bar{Y}}{\sqrt{\sum (Y_i - \bar{Y})^2}}.$$

This transformation is linear, so it maps lines to lines. If we transform a line fitted to the data, the sum of squared residuals is multiplied by a positive constant factor. Therefore, the transformation preserves the line of best fit.

Let $r = \sum x_i y_i$. Then

$$\begin{aligned} E &= \sum (y_i - a - bx_i)^2 \\ &= \sum (y_i^2 + a^2 + b^2 x_i^2 - 2ay_i - 2bx_i y_i + 2abx_i) \\ &= \sum y_i^2 + \sum a^2 + \sum b^2 x_i^2 - \sum 2ay_i - \sum 2bx_i y_i + \sum 2abx_i \\ &= 1 + na^2 + b^2 - 2br \\ &= (1 - r^2) + na^2 + (b - r)^2. \end{aligned}$$

The sum is minimized when $a = 0$ and $b = r$, so the line of best fit is $y = rx$. What a simple equation! Unfortunately, the equation is a bit messier when expressed in terms of the original variables.

$$\begin{aligned} \frac{y - \bar{Y}}{\sqrt{\sum (Y_i - \bar{Y})^2}} &= \left(\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \right) \left(\frac{x - \bar{X}}{\sqrt{\sum (X_i - \bar{X})^2}} \right) \\ y - \bar{Y} &= \left(\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \right) (x - \bar{X}). \end{aligned}$$

Note that r is the Pearson correlation coefficient of the sample. This shows that the correlation coefficient can be interpreted geometrically as the slope of the line of best fit when the x and y values are standardized.