

Least Squares Regression

Suppose you observe n pairs of values x_i and y_i ($i = 1, \dots, n$) and want to fit a straight line to predict y values from the x values:

- x_i = value of the explanatory variable for individual i .
- y_i = actual value of the response variable for individual i .
- $\hat{y}_i = a + bx_i$ = fitted value (estimated mean value of the response) for individual i .

The *least square* fit chooses a and b to minimize the sum of the squared prediction errors (SSE). That is, it minimizes

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

You can derive the least squares estimates using calculus. It turns out they depend only on the sample means, standard deviations and correlation coefficient ($\bar{x}, \bar{y}, s_x, s_y, r$). A nice representation is as follows:

$$b = r \frac{s_y}{s_x}; \quad a = \bar{y} - b\bar{x}.$$

Notice that it matters which variable you specify as y and which you specify for x . The fitted slope translates units of x to units of y by including the ratio of the standard deviations (r is unitless). The fitted intercept ensures that the prediction for someone with an average x value will be the average y value:

$$\bar{y} = a + b\bar{x}.$$

We can also compare the standardized values of the explanatory variable and the fitted y value to see the “regression to the mean” effect:

$$\frac{\hat{y}_i - \bar{y}}{s_y} = r \left(\frac{x_i - \bar{x}}{s_x} \right).$$

This implies that, unless there is a perfect correlation ($r = \pm 1$), the fitted y_i values are fewer standard deviations from the mean than are the corresponding x_i values. That is, the predictions are “regressed towards the mean.” Galton first noticed this effect in the late 1800’s when considering the adult heights of sons and fathers. He noticed that, on average, the sons of tall fathers are taller than average, but not as tall, on average, as the fathers. Similarly, the sons of short fathers were, on average, shorter than the mean height, but not as short, on average, as their fathers. The explanation of this is that father’s height explains only a part of the son’s height. When considering averages, the other factors that affect height tend to average out, leading to an average height closer to the mean height of all men (the same phenomenon is observed for fathers and daughters, mothers and sons, and mothers and daughters).

Cautions: The fitted line is merely a descriptive tool, and shouldn’t be treated as the “truth” about an association. If the association appears linear, the least squares line will give reasonably good predictions within the range of the x values (“interpolation”). It is risky to try to make predictions outside of this range (“extrapolation”) because there is no data to confirm that the linear association continues to hold.