Linear Regression

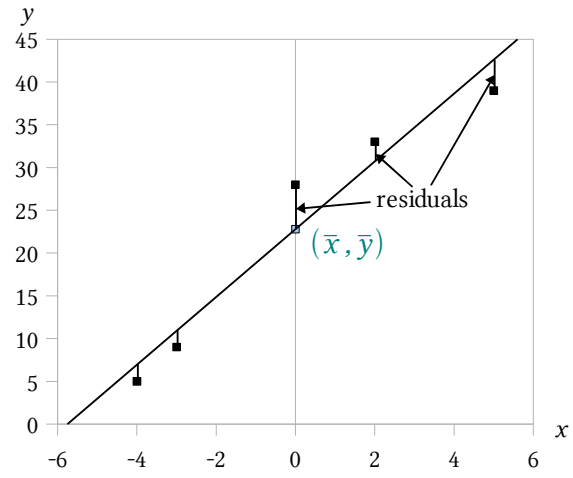
Philip Cho

This algorithm uses the method of least squares. That is, the linear model for which the sum of squared residuals has its minimum value is considered the best fit.

The regression line, the best-fit linear model, should represent the whole data. Therefore, the line should go through the point whose coordinates are the mean values of the two variables.

Algorithm:

Let's call each point (x_i, y_i).
Since the regression line should go through $(\bar{x}, \bar{y})$, the equation of the line is $y = m(x - \bar{x}) + \bar{y}$.



The sum of the squares of the residuals is as follows:

$$\sum \{m(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$
$$= \sum (m X_i - Y_i)^2 \qquad \text{(Let } x_i - \bar{x} = X_i \text{ and } y_i - \bar{y} = Y_i\text{.)}$$
$$= \sum (m^2 X_i^{\,2} - 2m X_i Y_i + Y_i^{\,2})$$
$$= m^2 \sum X_i^{\,2} - 2m \sum X_i Y_i + \sum Y_i^{\,2}$$
$$= \left(\sum X_i^{\,2}\right) m^2 - 2\left(\sum X_i Y_i\right) m + \sum Y_i^{\,2}$$

From the expression above, it can be inferred that the sum of the squares of the residuals has its minimum value when $m = \dfrac{\sum X_i Y_i}{\sum X_i^{\,2}} = \dfrac{\sum \{(x_i - \bar{x})(y_i - \bar{y})\}}{\sum (x_i - \bar{x})^2}$

Therefore, the slope and the y-intercept of the regression line are as follows:

Slope: $m = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

y-intercept: $n = m\bar{x} + \bar{y}$

Example:

Consider the following set of data:

x_i	-4	-3	0	2	5
y_i	5	9	28	33	39

$$\sum (x_i - \bar{x})^2 = 54$$
$$\sum (y_i - \bar{y})^2 = 900.8$$
$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 214$$

$x_i - \bar{x}$	-4	-3	0	2	5
$y_i - \bar{y}$	-17.8	-13.8	5.2	10.2	16.2

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{214}{54} \approx 3.963$$
$$n = m\bar{x} + \bar{y} = 3.963 \cdot 0 + 22.8 = 22.8$$

Thus, the equation of the regression line is $y = 3.963x + 22.8$.