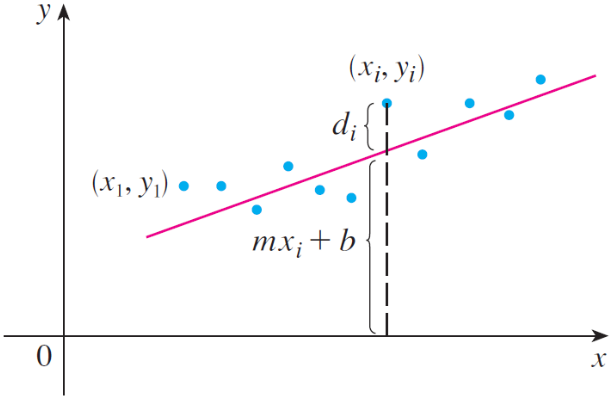
Math 1800 Project 7: Ordinary Linear Regression[*]

Subhadip Chowdhury

Suppose that a scientist has reason to believe that two quantities x and y are related linearly, that is, $y = mx + b$, at least approximately, for some values of m and b. The scientist performs an experiment and collects data in the form of points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, and then plots these points. The points dont lie exactly on a straight line, so the scientist wants to find constants m and b so that the line $y = mx + b$ fits the points as well as possible (see the figure).



1. For each data point (x_i, y_i), show that the corresponding point directly above or below it on the best fit line has y-coordinate $b + mx_i$.

2. Let d_i be the vertical distance between each data point and the corresponding point on the straight line found in part (1) (see the figure). For each data point (x_i, y_i), show that the d_i is $(y_i - (b + mx_i))$. We can think of these as error measurement of each data point.

3. The method of Ordinary Linear Regression tries to minimize the sum of the squares of the errors. Form the function $f(b, m)$ which is the sum of all of the n squared distances found in part (2).

   That is,

$$f(b, m) = \sum_{i=1}^{n} (y_i - (b + mx_i))^2$$

   Our goal is to find b and m that minimizes $f(b, m)$.

4. Show that the partial derivatives $\frac{\partial f}{\partial b}$ and $\frac{\partial f}{\partial m}$ are given by

$$\frac{\partial f}{\partial b} = -2 \sum_{i=1}^{n} (y_i - (b + mx_i))$$

   and

$$\frac{\partial f}{\partial m} = -2 \sum_{i=1}^{n} (y_i - (b + mx_i)) \cdot x_i$$

[*] Source: Hughes-Hallett, Stewart

5. Show that the critical point equations $\frac{\partial f}{\partial b} = 0$ and $\frac{\partial f}{\partial m} = 0$ lead to a pair of simultaneous linear equations in b and m:

$$\begin{aligned} nb + \left(\sum x_i\right) m &= \sum y_i \\ \left(\sum x_i\right) b + \left(\sum x_i^2\right) m &= \sum x_i y_i \end{aligned}$$

6. Solve the equations in part (d) for b and m, getting

$$\begin{aligned} b &= \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ m &= \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{aligned}$$

7. Find the line of best fit for the following data points: $(1,1)$, $(2,1)$, and $(3,3)$.