Mathematics behind single linear regression and multiple linear regression

Method of least squares:

Imagine that we now have a set of n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Now we want to find the most accurate linear function $y = ax + b$ that best represents these data points. In this sense, we want every y on this function to be closest to the original data points. Now for each data point y_{di}, the deviation of linear function from this data point is:

$$|\,(a\,x_{di} + b) - y_{di}|$$

The total deviation of the linear function from data points is:

$$\varepsilon(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

Now we want to minimize ε w.r.t a and b respectively.

Therefore, we set the partial derivatives of ε w.r.t a and b for minimization.

$$\frac{\partial \varepsilon}{\partial \mathrm{a}} = 0 \text{ and } \frac{\partial \varepsilon}{\partial b} = 0$$

$$\frac{\partial \varepsilon}{\partial \mathrm{a}} = 2\left(a \sum x_i^2 + b \sum x_i - \sum x_i\, y_i\right) = 0 - (1)$$

$$\frac{\partial \varepsilon}{\partial \mathrm{b}} = 2(a \sum x_i + bn - \sum y_i) = 0 - (2)$$

With these two equations, we now want to express a and b in terms of x and y.

$$a \sum x_i^2 + b \sum x_i = \sum x_i y_i - (3)$$

$$a \sum x_i + bn = \sum y_i \; - (4)$$

By Cramer's rule, we know that two linear equations in the form $pa + qb = r$ and $sa + tb = u$ can be readily solved by the following results:

$$a = \frac{rt - uq}{pt - sq}$$

$$b = \frac{pu - sr}{pt - sq}$$

Now, we can express 3 and 4 as:

$a = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$ – (5)

$b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$ – (6)

Note that the mean value $\bar{y}$ is equal to $\frac{1}{n}\sum y_i$.

With some arrangements of equation (5) and (6), and including $\bar{y}$, we get

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

A least squares line, or else called regression line, is thus found by substituting a and b in to $y = ax + b$.

Error sum of squares (SSE), total sum of squares (SST), regression sum of squares (SSR) and coefficient of determination, r^2

To understand more about linear regression, we define the following three quantities:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

A brief explanation on the three quantities:

- SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line,$\hat{y}_i$, is from the horizontal "no relationship line," the sample mean or $\bar{y}$.
- SSE is the "error sum of squares" and quantifies how much the data points,y_i, vary around the estimated regression line,$\hat{y}_i$.
- SSTO is the "total sum of squares" and quantifies how much the data points, y_i, vary around their mean, $\bar{y}$.

Note that SSTO = SSR + SSE, which is evident from the graph below.

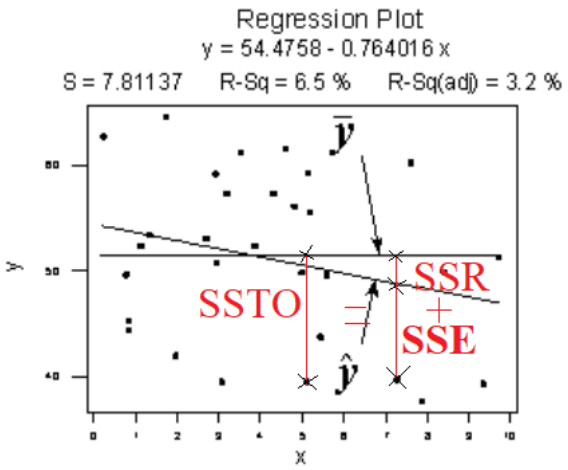

Figure 1, an example of regression plot with highly uncorrelated data

Now, we define another term, called coefficient of determination (or r-squared value), denoted r^2, to be the regression sum of squares divided by the total sum of squares. Mathematically, it is:

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Here are some basic characteristics of r^2:

- Since is a proportion, it is always a number between 0 and 1.
- If $r^2 = 1$, all of the data points fall perfectly on the regression line. The predictor x accounts for all of the variation in y_i
- If $r^2 = 0$, the estimated regression line is perfectly horizontal. The predictor x accounts for none of the variation in y_i
- If r^2 is between 0 and 1, we say that " r^2 ×100 percent of the variation in y is 'explained by' the variation in predictor x."

To interpret the last statement, take a look on the following plot:

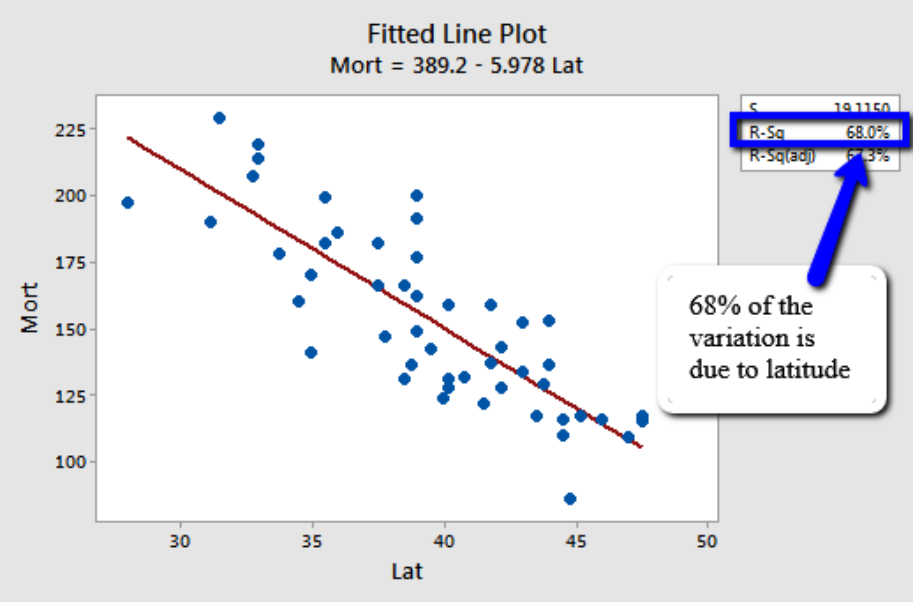

Figure 2, a plot of mortality rate of skin cancer versus latitude collected from https://online.stat.psu.edu/onlinecourses/sites/stat501/files/data/skincancer.txt

In this case, we can say that 68% of the variation in skin cancer mortality is 'due to' or is 'explained by' latitude. However, correlation does not mean causation. Hence, the term ‘due to’ merely refers to the correlation between mortality and latitude, and by no means indicating that latitude causes skin cancer mortality.