Linear Regression

By

Romilly Djee Yin Hills

July 25, 2020

Chapter 1

Linear Regression

1.1 Ordinary Least Squares

Commonly referred to as a 'line of best fit', this is a method of fitting noisy data to linear model, $y = mx + c$ by minimising some additional error.

1.1.1 Bivariate Case

The data is desired to be of the form

$$y_i = b_0 + b_1 x_i + e_i \tag{1.1}$$

where b_0 is the y intercept, b_1 is the gradient, e_i is some noise included in the data and the subscript i denotes the ith data point up to N. To find the best line of best fit, minimise the sum of squared errors (SSE) with respect to the y intercept and the gradient. The sum of squared errors is expressed as

$$SSE = \sum e_i^2. \tag{1.2}$$

Then with Equation (1.1) the SSE can be expressed as

$$SSE = \sum \left(y_i - b_0 - b_1 x_i\right)^2. \tag{1.3}$$

The minimum with respect to b_0 is then found with

$$\frac{\partial}{\partial b_0} SSE = \sum -2\left(y_i - b_0 - b_1 x_i\right) = 0. \tag{1.4}$$

Likewise, the minimum with respect to b_1 is then found with

$$\frac{\partial}{\partial b_1} SSE = \sum -2x_i\left(y_i - b_0 - b_1 x_i\right) = 0. \tag{1.5}$$

Using Equation (1.4) and Equation (1.5) as a set of simultaneous equations, an expression for b_0 and b_1 can be obtained. Writing Equation (1.4) in the form

$$\sum y_i - \sum b_0 - \sum b_1 x_i = 0 \tag{1.6}$$

allows the substitution $\sum b_0 = N b_0$ as b_0 is constant for all data points. Then b_1 can be expressed as

$$b_0 = \frac{\sum y_i - b_1 \sum x_i}{N}. \tag{1.7}$$

Then to obtain b_1, Equation (1.5) should be expressed as

$$\sum x_i y_i - b_0 \sum x_i - b_1 \sum x_i^2 = 0. \tag{1.8}$$

With the expression for b_0 from Equation (1.7), this becomes

$$\sum x_i y_i - \frac{\sum y_i - b_1 \sum x_i}{N} \sum x_i - b_1 \sum x_i^2 = 0 \tag{1.9}$$

and b_1 can then be expressed as

$$b_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{N}}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}. \tag{1.10}$$

With the expression for b_0 from equation (1.7) and the expression for b_1 from equation (1.10) we can create a linear model of the form $y = mx + c$ with the substitutions $m = b_1$ and $c = b_0$. We finaly have

$$y = b_1 x + b_0 \tag{1.11}$$