Linear Regression

By

Romilly Djee Yin Hills

August 5, 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

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

1.1.2 Multiple Linear Regression

Here we examine the case where we desire a relation between multiple non-interacting variables and our value to predict

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k \tag{1.12}$$

with N data points of the form

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + e_i \tag{1.13}$$

where i ranges from one to N and e_i is some noise included in the data. We again are aiming to minimise the sum of squared errors (SSE) which we will restate as

$$SSE = \sum e_i^2. \tag{1.14}$$

To account for multiple varaibles and data points we can rewrite equation (1.13) with vectors and matrices

$$\vec{Y} = X\vec{B} + \vec{E} \tag{1.15}$$

where

$$\vec{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_N \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ 1 & x_{N1} & x_{N2} & ... & x_{Nk} \end{bmatrix} \qquad \vec{B} = \begin{bmatrix} b_0 \\ b_1 \\ b_k \end{bmatrix} \qquad \vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_N \end{bmatrix}. \tag{1.16}$$

We can then use these definitions in equation (1.14) to find

$$\sum e_i^2 = e^T \cdot e = \left(\vec{Y} - \vec{B}X\right)^T \cdot \left(\vec{Y} - X\vec{B}\right) \tag{1.17}$$

$$= \vec{Y}^T\vec{Y} - \vec{Y}^T X\vec{B} - \vec{B}^T X^T \vec{Y} + \vec{B}^T X^T X \vec{B} \tag{1.18}$$

where the superscript T denotes the transpose. From here onwards, I will use 1x2 vectors and 2x2 matrices to provide worked examples of key points, as opposed to a fully proven derivation. To which point we can show that

$$\vec{Y}^T X\vec{B} = [y_1, y_2] \begin{bmatrix} b_0 + b_1 x_{11} \\ b_0 + b_1 x_{21} \end{bmatrix} = y_1 (b_0 + b_1 x_{11}) + y_2 (b_0 + b_1 x_{21}) \tag{1.19}$$

$$\vec{B}^T X^T \vec{Y} = [b_0 + b_1 x_{11}, b_0 + b_1 x_{21}] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = y_1 (b_0 + b_1 x_{11}) + y_2 (b_0 + b_1 x_{21}) \tag{1.20}$$

and we can write

$$\sum e_i^2 = \vec{Y}^T\vec{Y} - 2\vec{Y}^T X\vec{B} + \vec{B}^T X^T X\vec{B}. \tag{1.21}$$

In order to find the minimum error with respect to $\vec{B}$, we find when the gradient is equal to zero

$$\frac{\partial}{\partial \vec{B}} e^T \cdot e = 0. \tag{1.22}$$

Here I will show a 2x2 example of this calculation; we know the expanded version of $\vec{Y}^T X\vec{B}$, so we now need to find $\vec{B}^T X^T X\vec{B}$

$$\vec{B}^T X^T = [b_0 + b_1 x_{11}, b_0 + b_1 x_{21}] \tag{1.23}$$

$$\vec{B}^T X^T X = \left[2b_0 + b_1 (x_{11} + x_{21}), b_0 (x_{11} + x_{21}) + b_1 \left(x_{11}^2 + x_{21}^2\right)\right] \tag{1.24}$$

$$\vec{B}^T X^T X\vec{B} = 2b_0^2 + 2b_0 b_1 (x_{11} + x_{21}) + b_1^2 \left(x_{11}^2 + x_{21}^2\right). \tag{1.25}$$

With these expanded expressions we can now find the partial derivitives in equation (1.22). The first term $\vec{Y}^T\vec{Y}$ has no dependency on $\vec{B}$ and so will always be zero. For the second term we have

$$\frac{\partial}{\partial b_0} \vec{Y}^T X\vec{B} = y_1 + y_2 \tag{1.26}$$

$$\frac{\partial}{\partial b_1} \vec{Y}^T X\vec{B} = y_1 x_{11} + y_2 x_{21} \tag{1.27}$$

which we can restate in matrix form as

$$\frac{\partial}{\partial \vec{B}} \vec{Y}^T X \vec{B} = X^T \vec{Y}. \tag{1.28}$$

For the second term we have

$$\frac{\partial}{\partial b_0} \vec{B}^T X^T X \vec{B} = 4b_0 + 2b_1 \left(x_{11} + x_{21}\right) \tag{1.29}$$

$$\frac{\partial}{\partial b_1} \vec{B}^T X^T X \vec{B} = 2b_0 \left(x_{11} + x_{21}\right) + 2b_1 \left(x_{11}^2 + x_{21}^2\right) \tag{1.30}$$

which, after recognising the expansion of $X^T X$ we can write these two partial derivites as

$$\frac{\partial}{\partial \vec{B}} \vec{B}^T X^T X \vec{B} = 2X^T X \vec{B}. \tag{1.31}$$

We can now combine these definitions to get the result of equation (1.22)

$$\frac{\partial}{\partial \vec{B}} e^T \cdot e = -2X^T \vec{Y} + 2X^T X \vec{B} = 0 \tag{1.32}$$

which can simply be rearraged to make $\vec{B}$ the subject

$$\vec{B} = \left(X^T X\right)^{-1} X^T \vec{Y}. \tag{1.33}$$

The above relation can be used to calculate the components of $\vec{B}$ and substituted into equation (1.12) to produce an equation for the multiple linear regression line.