Linear Regression Demystified

Linear regression is an important subject in statistics. In elementary statistics courses, formulae related to linear regression are often stated without derivation. This note intends to derive these formulae for students with more advanced math background.

1 Simple Regression: One Variable

1.1 Least Square Prescription

Suppose we have a set of data points (x_i, y_i) $(i = 1, 2, \cdots, n)$. The goal of linear regression is to find a straight line that best fit the data. In other words, we want to build a model

$$\hat{y}_i = \beta_0 + \beta_1 x_i \tag{1}$$

and tune the parameters β_0 and β_1 so that $\hat{y}_i$ is as close to y_i as possible for all $i = 1, 2, \cdots, n$.

Before we do the math, we need to clarify the problem. How do we judge the "closeness" of $\hat{y}_i$ and y_i for all i? If the data points (x_i, y_i) do not fall exactly on a straight line, y_i and $\hat{y}_i$ is not going to be the same for all i. The deviation of y_i from $\hat{y}_i$ is called the *residual* and is denoted by ϵ_i here. In other words,

$$y_i = \hat{y}_i + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{2}$$

The least square prescription is to find β_0 and β_1 that minimize the sum of the square of the residuals SSE:

$$SSE = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3}$$

If you are familiar with calculus, you will know that the minimization can be done by setting the derivatives of SSE with respect to β_0 and β_1 to 0, i.e. $\partial SSE/\partial\beta_0 = 0$ and $\partial SSE/\partial\beta_1 = 0$. The resulting equations are[1]

$$\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \epsilon_i = 0 \quad \text{and} \quad \sum_{i=1}^n x_i(y_i - \hat{y}_i) = \sum_{i=1}^n x_i\epsilon_i = 0. \tag{4}$$

We can combine the first and second equation to derive an alternative equation for the second equation:

$$\sum_{i=1}^n \epsilon_i = 0 \quad \Rightarrow \quad \sum_{i=1}^n \bar{x}\epsilon_i = 0$$

Subtracting this equation from the second equation of (4) yields

$$\sum_{i=1}^n (x_i - \bar{x})\epsilon_i = 0.$$

Equation (4) can be expressed as

$$\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \epsilon_i = 0 \quad \text{and} \quad \sum_{i=1}^n (x_i - \bar{x})(y_i - \hat{y}_i) = \sum_{i=1}^n (x_i - \bar{x})\epsilon_i = 0. \tag{5}$$

If you are not familiar with calculus, we can show you a proof using algebra. I actually prefer his proof since it is more rigorous. Before presenting the proof, let's remind you some concepts in statistics.

[1] Careful students may realize that this only gaurantees that SSE is stationary but not necessarily minimum. There is also a concern whether the resulting solution is a global minimum or a local minimum. Detail analysis (see Section 1.3) reveals that the solution in our case is a global minimum.

1.2 Mean, Standard Deviation and Correlation

The mean and standard deviation of a set of points $\{u_i\}$ are defined as

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i \quad , \quad SD_u = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2} \ . \tag{6}$$

The Z score of u_i is defined as

$$Z_{ui} = \frac{u_i - \bar{u}}{SD_u}. \tag{7}$$

The *correlation* of two set of points (with equal number of elements) $\{x_i\}$ and $\{y_i\}$ is defined as

$$r_{xy} = r_{yx} = \frac{1}{n}\sum_{i=1}^{n} Z_{xi}Z_{yi} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{SD_x SD_y} \tag{8}$$

The Cauchy-Schwarz inequality states that for two set of real numbers $\{u_i\}$ and $\{v_i\}$,

$$\left(\sum_{i=1}^{n} u_i v_i\right)^2 \le \left(\sum_{i=1}^{n} u_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right)$$

and the equality holds if and only if $v_i = ku_i$ for all i, where k is a constant. You can find the proof of the inequality in, e.g., Wikipedia.

It follows from the Cauchy-Schwarz inequality that the correlation $|r_{xy}| \le 1$, with $|r_{xy}| = 1$ if and only if x_i and y_i fall exactly on a straight line, i.e. $\epsilon_i = 0$ for all i. So the correlation r_{xy} measures how tightly the points (x_i, y_i) are clustered around a line.

1.3 Regression Equation

To derive the regression equation, we first rewrite SSE in equation (3) in terms of Z-scores. It follows from the definition of Z-scores that

$$x_i = \bar{x} + SD_x Z_{xi} \quad , \quad y_i = \bar{y} + SD_y Z_{yi}$$

and the SSE becomes

$$\begin{aligned} SSE &= \sum_{i=1}^{n}[\bar{y} + SD_y Z_{yi} - \beta_0 - \beta_1(\bar{x} + SD_x Z_{xi})]^2 \\ &= \sum_{i=1}^{n}(SD_y Z_{yi} - \beta_1 SD_x Z_{xi} + \bar{y} - \beta_0 - \beta_1\bar{x})^2 \end{aligned}$$

For simplicity, we denote $\tilde{\beta}_0 = \bar{y} - \beta_0 - \beta_1\bar{x}$. Then we have

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(SD_y Z_{yi} - \beta_1 SD_x Z_{xi} + \tilde{\beta}_0)^2 \\ &= SD_y^2\sum_{i=1}^{n} Z_{yi}^2 + \beta_1^2 SD_x^2 \sum_{i=1}^{n} Z_{xi}^2 + \sum_{i=1}^{n}\tilde{\beta}_0^2 - 2\beta_1 SD_x SD_y \sum_{i=1}^{n} Z_{xi}Z_{yi} \\ &\quad +2\tilde{\beta}_0 SD_y \sum_{i=1}^{n} Z_{yi} - 2\tilde{\beta}_0\beta_1 SD_x \sum_{i=1}^{n} Z_{xi} \end{aligned} \tag{9}$$

Recall that the Z-scores are constructed to have zero mean and unit standard deviation. It follows that

$$\sum_{i=1}^{n} Z_{xi} = \sum_{i} Z_{yi} = 0 \quad , \quad \frac{1}{n}\sum_{i=1}^{n} Z_{xi}^2 = \frac{1}{n}\sum_{i=1}^{n} Z_{yi}^2 = 1$$

and from the definition of correlation we have

$$\frac{1}{n}\sum_{i=1}^{n} Z_{xi}Z_{yi} = r_{xy},$$

Thus the SSE in equation (9) can be written as

$$\begin{aligned} SSE &= n(SD_y^2 + \beta_1^2 SD_x^2 + \tilde{\beta}_0^2 - 2\beta_1 r_{xy} SD_x SD_y) \\ &= n[\tilde{\beta}_0^2 + (SD_x\beta_1 - r_{xy}SD_y)^2 + (1-r_{xy}^2)SD_y^2] \\ &= n[(\bar{y}-\beta_0-\beta_1\bar{x})^2 + (SD_x\beta_1 - r_{xy}SD_y)^2] + n(1-r_{xy}^2)SD_y^2 \end{aligned} \tag{10}$$

Note that we have re-expressed $\tilde{\beta}_0$ in terms of β_0 and β_1 in the last step. Remember your goal is to find β_0 and β_1 to minimize SSE. The quantity inside the square bracket in equation (10) is a sum of two squares and contain the variables β_0 and β_1, whereas $n(1-r_{xy}^2)SD_y^2$ is a fixed quantity. Therefore, we conclude that

$$SSE \geq n(1-r_{xy}^2)SD_y^2$$

and the equality holds if and only if

$$\bar{y} - \beta_0 - \beta_1\bar{x} = 0 \quad \text{and} \quad SD_x\beta_1 - r_{xy}SD_y = 0.$$

Therefore, the values of β_0 and β_1 that minimize SSE is

$$\boxed{\beta_1 = r_{xy}\frac{SD_y}{SD_x} \quad , \quad \beta_0 = \bar{y} - \beta_1\bar{x}} \tag{11}$$

and the minimum SSE is

$$\boxed{SSE = n(1-r_{xy}^2)SD_y^2} \tag{12}$$

How does equations (11) related to equations (5) derived from calculus? We are going to prove that they are the same thing. Using equations (11) we have

$$\begin{aligned} \sum_{i=1}^{n}\epsilon_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ &= \sum_{i=1}^{n}(y_i - \bar{y} + \beta_1\bar{x} - \beta_1 x_i) \\ &= \sum_{i=1}^{n}(y_i - \bar{y}) - \beta_1\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= 0 \end{aligned}$$

If the last step is not obvious to you, here's a simple proof:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \sum_{i=1}^{n} x_i = n\bar{x} = \sum_{i=1}^{n}\bar{x} \quad \Rightarrow \quad \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

and similarly

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0.$$

Now let's look at the second equation of (5):

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}+\beta_1\bar{x}-\beta_1 x_i) \\ &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2\end{aligned}$$

From the definition of correlation and standard deviation, we have

$$r_{xy}=\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})(y_i-\bar{y})}{SD_x SD_y} \quad\Rightarrow\quad \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=nr_{xy}SD_xSD_y$$

$$SD_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad\Rightarrow\quad \sum_{i=1}^{n}(x_i-\bar{x})^2=nSD_x^2$$

Thus,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=nr_{xy}SD_xSD_y-\beta_1 nSD_x^2=nSD_x^2\left(r_{xy}\frac{SD_y}{SD_x}-\beta_1\right)=0.$$

Therefore, we have proved that equations (5) follows from equations (11). On the other hand, equations (5) are two linear equations for β_0 and β_1. We can solve for β_0 and β_1 from these equations, which we will do so in Section 1.7. The result is that β_0 and β_1 are given by equations (11) and so the two sets of equations are equivalent.

1.4 Interpretation

We will try to give you an intuition of the meaning of equations (5) below.

We can rewrite the first equation of (5) as

$$\bar{\epsilon}=0. \tag{13}$$

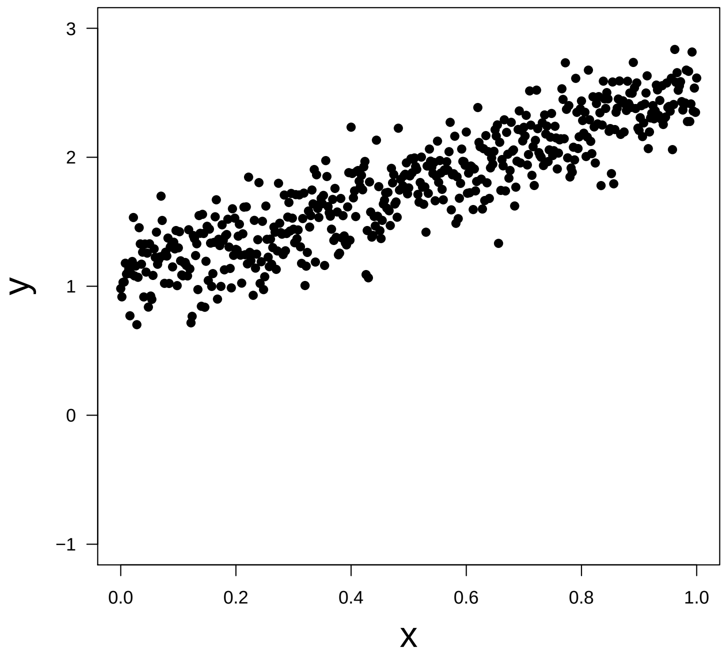
That is to say that the mean of ϵ_i vanishes. The second equation of (5) means that

$$\sum_{i=1}^{n}(x_i-\bar{x})\epsilon_i=\sum_{i=1}^{n}(x_i-\bar{x})(\epsilon_i-\bar{\epsilon})=0 \quad\Rightarrow\quad nSD_xSD_\epsilon r_{x\epsilon}=0 \quad\Rightarrow\quad r_{x\epsilon}=0.$$

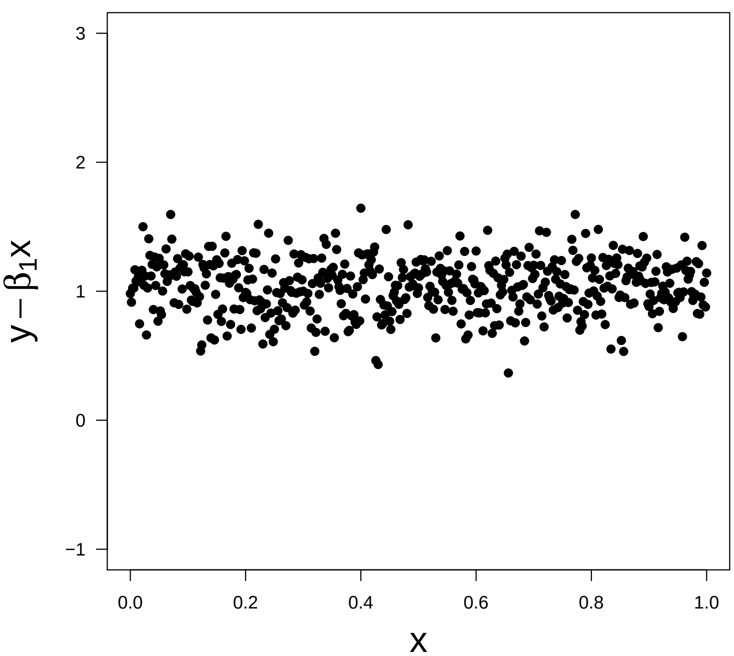
That is to say that x_i and ϵ_i are uncorrelated.

Thus the least square prescription is to find β_0 and β_1 to make the residuals having zero mean and being uncorrelated with x_i. Let's illustrate how this can be done graphically.

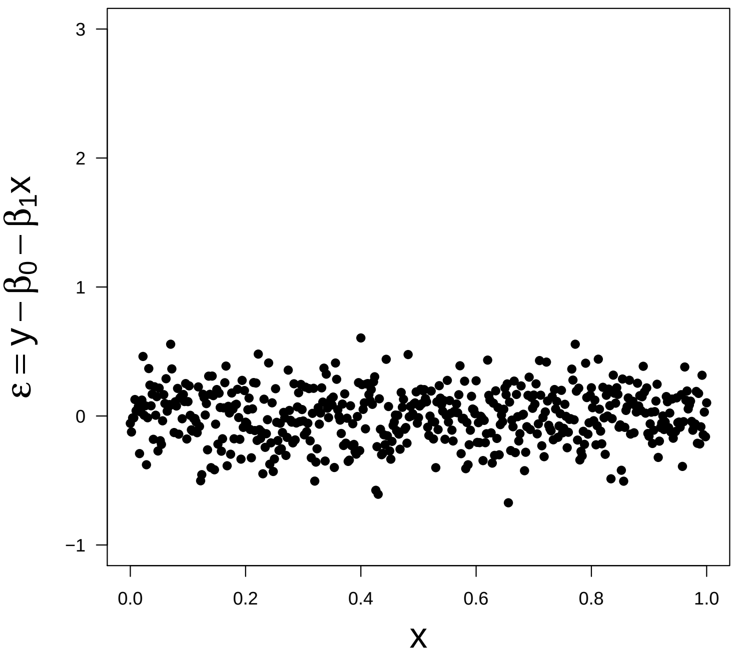
Suppose we have the following data set of x_i and y_i:



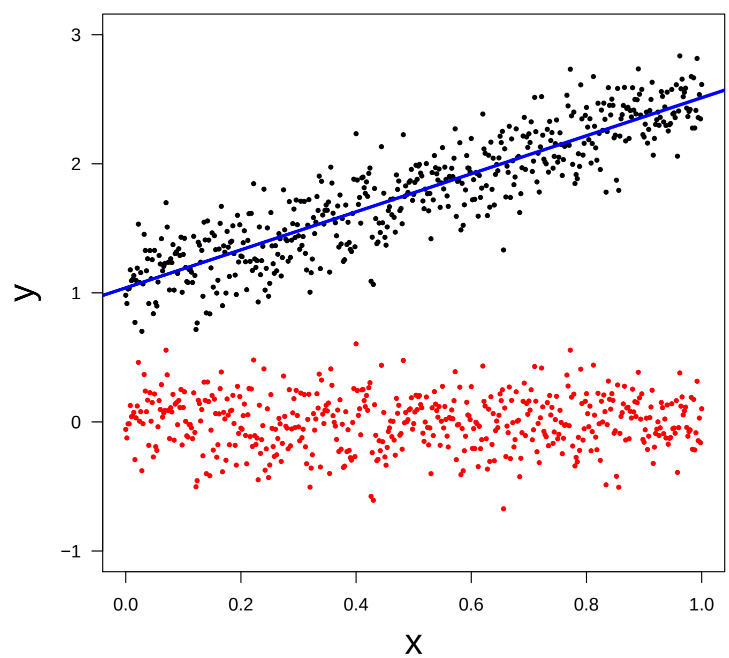
We want to fit a straight line $y = \beta_0 + \beta_1 x$ to the data points. We know the line has to make the resuduals having zero mean and being uncorrelated with x_i. First, let's imagine choosing various values of β_1 and see what the plot $y - \beta_1 x$ versus x look like. In general, we will have the plot similar to the one above but with different overall slope. If we pick the right value of β_1, we will see that the residual $y - \beta_1 x$ vs x is flat:



We know have achieved the first goal: by choosing the right value of β_1, the residual $y - \beta_1 x$ is uncorrelated with x. However, the residuals do not have zero mean. So next we want to choose β_0 so that $y - \beta_0 - \beta_1 x$ has zero mean. When the right value of β_0 is chosen, the residuals become this:



We have accomplished our goal. The residuals now have zero mean and are uncorrelated with x_i. The resulting regression line $y = \beta_0 + \beta_1 x$ is the straight line that is best fit to the data points. The graph below shows the regression line (blue), data points (black) and the residuals (red).



We have outlined the idea of linear regression. Next we introduce the the concept of a vector, which proves to be very useful for regression with more than one variables.

1.5 Vectors

An n-dimensional vector $\boldsymbol{V}$ contains n real numbers $v_1, v_2, \cdots, v_n$, often written in the form

$$\boldsymbol{V} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

The numbers $v_1, v_2, \cdots, v_n$ are called the *components* of $\boldsymbol{V}$. Here we adopt the convention to write vector variables in boldface to distinguish them from numbers. It is useful to define a special vector $\boldsymbol{X_0}$ whose components are all equal to 1. That is,

$$\boldsymbol{X_0} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{14}$$

The scalar product of two vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ is defined as

$$\boldsymbol{U} \cdot \boldsymbol{V} = \boldsymbol{V} \cdot \boldsymbol{U} = \sum_{i=1}^{n} u_i v_i,$$

where v_i and u_i are components of $\boldsymbol{V}$ and $\boldsymbol{U}$. Note that the result of the scalar product of two vectors is a number, not a vector. Two vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ are said to be *orthogonal* if and only if $\boldsymbol{U} \cdot \boldsymbol{V} = 0$.

It follows from the definition that

$$\boldsymbol{V} \cdot \boldsymbol{V} = \sum_{i=1}^{n} v_i^2 \geq 0,$$

with $\boldsymbol{V} \cdot \boldsymbol{V} = 0$ if and only if all components of $\boldsymbol{V}$ are 0. It follows from the Cauchy-Schwarz inequality that

$$(\boldsymbol{U} \cdot \boldsymbol{V})^2 \leq (\boldsymbol{U} \cdot \boldsymbol{U})(\boldsymbol{V} \cdot \boldsymbol{V})$$

with the equality holds if and only if $\boldsymbol{U} = k\boldsymbol{V}$ for some constant k.

Now that we have introduced the basic concept of a vector, we are now ready to express the regression equations in vector form.

1.6 Regression Equations in Vector Form

Equation (2) can be written as

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{\epsilon} = \beta_0 \boldsymbol{X_0} + \beta_1 \boldsymbol{X} + \boldsymbol{\epsilon}. \tag{15}$$

The sum of square of residuals SSE in (3) can be written as

$$SSE = \boldsymbol{\epsilon} \cdot \boldsymbol{\epsilon} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \cdot (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X}) \cdot (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X}) \tag{16}$$

To minimize SSE, we set the derivatives of SSE with respect to β_0 and β_1 to 0. The resulting equations are

$$\boldsymbol{X_0} \cdot (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X}) = 0 \quad , \quad \boldsymbol{X} \cdot (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X}) = 0$$

or

$$\boldsymbol{X_0} \cdot \boldsymbol{\epsilon} = \boldsymbol{X} \cdot \boldsymbol{\epsilon} = 0. \tag{17}$$

This is equation (4) written in vector form. We see that the least square prescription can be interpreted as finding β_0 and β_1 to make the residual vector $\boldsymbol{\epsilon}$ orthogonal to both $\boldsymbol{X_0}$ and $\boldsymbol{X}$.

The mean and standard deviation in (6) can be expressed as

$$\bar{u} = \frac{1}{n}\boldsymbol{X_0} \cdot \boldsymbol{U} \quad , \quad SD_u = \sqrt{\frac{1}{n}(\boldsymbol{U} - \bar{u}\boldsymbol{X_0}) \cdot (\boldsymbol{U} - \bar{u}\boldsymbol{X_0})} \tag{18}$$

and the Z score in (7) becomes

$$\boldsymbol{Z_u} = \frac{\boldsymbol{U} - \bar{u}\boldsymbol{X_0}}{SD_u} \tag{19}$$

It follows that

$$\boldsymbol{Z_u} \cdot \boldsymbol{Z_u} = n \tag{20}$$

The correlation r_{xy} in (8) is proportional to the scalar product of $\boldsymbol{Z_x}$ and $\boldsymbol{Z_y}$:

$$r_{xy} = \frac{1}{n}\boldsymbol{Z_x} \cdot \boldsymbol{Z_y} \tag{21}$$

The equation $r_{x\epsilon} = 0$ is equivalent to

$$\boldsymbol{Z_x} \cdot \boldsymbol{\epsilon} = 0 \tag{22}$$

We see that the equations in vector form are simpler and more elegant. We are now ready to solve the regression equations (15) and (17) for β_0 and β_1.

1.7 Regression Coefficients

In this section, we are going to use the vector algebra to solve for β_0 and β_1 to re-derive equations (11). This is a useful exercise because it can be easily generalized to multiple regression we consider later.

Start with equation (15). Taking the scalar product of (15) with $\boldsymbol{X_0}$, using equations (18), (17) and $\boldsymbol{X_0}\cdot\boldsymbol{X_0} = n$, we obtain

$$n\bar{y} = n\beta_0 + n\beta_1\bar{x} \quad \Rightarrow \quad \bar{y} = \beta_0 + \beta_1\bar{x}, \tag{23}$$

Multiplying both sides of the above equation by $\boldsymbol{X_0}$ gives

$$\bar{y}\boldsymbol{X_0} = \beta_0\boldsymbol{X_0} + \beta_1\bar{x}\boldsymbol{X_0}$$

Subtracting the above equation from (15) gives

$$\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = \beta_1(\boldsymbol{X} - \bar{x}\boldsymbol{X_0}) + \boldsymbol{\epsilon}$$

It follows from equation (19) that $\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = SD_y\boldsymbol{Z_y}$ and $\boldsymbol{X} - \bar{x}\boldsymbol{X_0} = SD_x\boldsymbol{Z_x}$. Hence we have

$$SD_y\boldsymbol{Z_y} = \beta_1 SD_x\boldsymbol{Z_x} + \boldsymbol{\epsilon}$$

Taking the scalar product with $\boldsymbol{Z_x}$ and using $\boldsymbol{Z_x} \cdot \boldsymbol{\epsilon} = 0$, we obtain

$$SD_y\boldsymbol{Z_x} \cdot \boldsymbol{Z_y} = \beta_1 SD_x\boldsymbol{Z_x} \cdot \boldsymbol{Z_x}$$

It follows from equation (21) and (20) that $\boldsymbol{Z_x} \cdot \boldsymbol{Z_y} = nr_{xy}$ and $\boldsymbol{Z_x} \cdot \boldsymbol{Z_x} = n$ and so the above equation reduces to

$$SD_y r_{xy} = \beta_1 SD_x \quad \Rightarrow \quad \beta_1 = r_{xy}\frac{SD_y}{SD_x}.$$

Combining the above equation with (23), we finally solve β_0 and β_1:

$$\boxed{\beta_1 = r_{xy}\frac{SD_y}{SD_x} \quad , \quad \beta_0 = \bar{y} - \beta_1\bar{x}} \tag{24}$$

The regression line is the straight line with slope $r_{xy}SD_y/SD_x$ passing through the point $(\bar{x}, \bar{y})$.

1.8 Root Mean Square Error

Having found a straight line that best fit the data points, we next want to know how good the fit is. One way to characterize the "goodness" of the fit is to calculate the mean square error RMSE, defined as

$$RMSE = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{\boldsymbol{\epsilon} \cdot \boldsymbol{\epsilon}}{n}} \tag{25}$$

It follows from equation (12) that $RMSE = \sqrt{1 - r_{xy}^2}\, SD_y$. Here we derive it again using the vector algebra since it can be easily generalized to multiple regression.

We introduce two quantities SST and SSM. The total sum square SST is defined as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \cdot (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \tag{26}$$

It follows from the definition of standard deviation [see equation (6) or (18)] that

$$SST = nSD_y^2 \tag{27}$$

The sum square predicted by the linear model SSM is defined as

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \tag{28}$$

Using $\hat{\boldsymbol{Y}} = \beta_0 \boldsymbol{X_0} + \beta_1 \boldsymbol{X}$ and $\beta_0 = \bar{y} - \beta_1 \bar{x}$, we have

$$\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0} = (\bar{y} - \beta_1 \bar{x})\boldsymbol{X_0} + \beta_1 \boldsymbol{X} - \bar{y}\boldsymbol{X_0} = \beta_1(\boldsymbol{X} - \bar{x}\boldsymbol{X_0}) = \beta_1 SD_x \boldsymbol{Z_x}, \tag{29}$$

where we have used equation (19) to obtain the last equality. Combining (28), (29), (20) (24) and (27), we obtain

$$SSM = n\beta_1^2 SD_x^2 = n\left(r_{xy}\frac{SD_y}{SD_x}\right)^2 SD_x^2 = nr_{xy}^2 SD_y^2 = r_{xy}^2 SST \tag{30}$$

Next we want to prove an identity that $SST = SSM + SSE$. To see that, we start with the identity

$$\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) = \boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})$$

We then "square" both sides by taking the scalar product with itself:

$$(\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \cdot (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) = [\boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})] \cdot [\boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})]$$

The left hand side is SST, the right hand side is the sum of SSE, SSM and a cross term:

$$\begin{aligned} SST &= \boldsymbol{\epsilon} \cdot \boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) + 2\boldsymbol{\epsilon} \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \\ &= SSE + SSM + 2\boldsymbol{\epsilon} \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \end{aligned}$$

It follows from (29) and $\boldsymbol{Z_x} \cdot \boldsymbol{\epsilon} = 0$ that the cross term vanishes:

$$\boldsymbol{\epsilon} \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) = \beta_1 SD_x \boldsymbol{Z_x} \cdot \boldsymbol{\epsilon} = 0$$

Another way of seeing this is to note that $\boldsymbol{\epsilon}$ is orthogonal to both $\boldsymbol{X_0}$ and $\boldsymbol{X}$, as required by the least square prescription (17). So ϵ is orthogonal to any linear combination of $\boldsymbol{X_0}$ and $\boldsymbol{X}$. Since $\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}$ is a linear combination of $\boldsymbol{X_0}$ and $\boldsymbol{X_1}$, it is orthogonal to $\boldsymbol{\epsilon}$.

We have just proved that

$$\boxed{SST = SSM + SSE} \tag{31}$$

We have calculated above that $SSM = r_{xy}^2$ and $SST = nSD_y^2$. Using the identity $SST = SSM + SSE$, we obtain

$$SSE = SST - SSM = (1 - r_{xy}^2)SST = n(1 - r_{xy}^2)SD_y^2.$$

Combining this equation with the definition of $RMSE$ in (25), we find

$$\boxed{RMSE = \sqrt{1 - r_{xy}^2}\, SD_y} \tag{32}$$

2 Multiple Regression

Now we want to generalize the results of simple regression to multiple regression. We will first consider the case with two variables in Section 2.1, because simple analytic expressions exist in this case and the derivation is relatively straightforward. Then we will turn to the more general case with more than two variables in Section 2.2. Finally, we will derive a general result for the RMSE calculation in Section 2.4.

2.1 Two Variables

Suppose we want to fit the data points $\{y_i\}$ with two variables $\{x_{1i}\}$ and $\{x_{2i}\}$ by a linear model

$$y_i = \hat{y}_i + \epsilon_i \quad , \quad i = 1, 2, \cdots, n$$

with

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

The least square prescription is again to find β_0, β_1 and β_2 to minimize the sum square of the residuals:

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

As in the case of simple regression, it is more convenient to rewrite the above equations in vector form as follows:

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{\epsilon} = \beta_0 \boldsymbol{X_0} + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \boldsymbol{\epsilon} \tag{33}$$

$$\hat{\boldsymbol{Y}} = \beta_0 \boldsymbol{X_0} + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} \tag{34}$$

$$SSE = \boldsymbol{\epsilon} \cdot \boldsymbol{\epsilon} = (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X_1} - \beta_2 \boldsymbol{X_2}) \cdot (\boldsymbol{Y} - \beta_0 \boldsymbol{X_0} - \beta_1 \boldsymbol{X_1} - \beta_2 \boldsymbol{X_2}) \tag{35}$$

To employ the least square prescription, we set the derivatives of SSE with respect to β_0, β_1 and β_2 to 0. The resulting equations are

$$\boldsymbol{X_0} \cdot \boldsymbol{\epsilon} = \boldsymbol{X_1} \cdot \boldsymbol{\epsilon} = \boldsymbol{X_2} \cdot \boldsymbol{\epsilon} = 0. \tag{36}$$

That is to say that $\boldsymbol{\epsilon}$ is orthogonal to $\boldsymbol{X_0}$, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. In other words, $\hat{\boldsymbol{Y}}$ is the vector $\boldsymbol{Y}$ projected onto the vector space spanned by $\boldsymbol{X_0}$, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$.

For convenience, we denote $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ as the Z-score vectors associated with $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, respectively. That is,

$$\boldsymbol{Z_1} = \frac{\boldsymbol{X_1} - \bar{x}_1 \boldsymbol{X_0}}{SD_1} \quad , \quad \boldsymbol{Z_2} = \frac{\boldsymbol{X_2} - \bar{x}_2 \boldsymbol{X_0}}{SD_2} \quad , \tag{37}$$

where SD_1 and SD_2 are the standard deviation of $\{x_{1i}\}$ and $\{x_{2i}\}$, respectively. We see that $\boldsymbol{Z_1}$ is a linear combination of $\boldsymbol{X_0}$ and $\boldsymbol{X_1}$; $\boldsymbol{Z_2}$ is a linear combination of $\boldsymbol{X_0}$ and $\boldsymbol{X_2}$. Since $\boldsymbol{\epsilon}$ to orthogonal to $\boldsymbol{X_0}$, $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, it is orthogonal to $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ as well:

$$\boldsymbol{Z_1} \cdot \boldsymbol{\epsilon} = \boldsymbol{Z_2} \cdot \boldsymbol{\epsilon} = 0. \tag{38}$$

Recall that the mean of a vector $\boldsymbol{U}$ is $\bar{u} = \boldsymbol{X_0} \cdot \boldsymbol{U}/n$, and the correlation between $\boldsymbol{U}$ and $\boldsymbol{V}$ is $r_{uv} = \boldsymbol{Z_u} \cdot \boldsymbol{Z_v}/n$. Thus the orthogonality conditions of $\boldsymbol{\epsilon}$ mean that (1) $\boldsymbol{\epsilon}$ has zero mean, $\bar{\epsilon} = 0$, and (2) $\boldsymbol{\epsilon}$ is uncorrelated with both $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$: $r_{\epsilon 1} = r_{\epsilon 2} = 0$.

To solve the regression equations (36), we take the scalar product of equation (33) with $\boldsymbol{X_0}$, resulting in the equation

$$\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 \tag{39}$$

This means that the regression line passes through the average point $(\bar{x}_1, \bar{x}_2, \bar{y})$. Multiplying equation (39) by $\boldsymbol{X_0}$ yields

$$\bar{y} \boldsymbol{X_0} = \beta_0 \boldsymbol{X_0} + \beta_1 \bar{x}_1 \boldsymbol{X_0} + \beta_2 \bar{x}_2 \boldsymbol{X_0}$$

Subtracting the above equation from (33) gives

$$\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = \beta_1(\boldsymbol{X_1} - \bar{x}_1\boldsymbol{X_0}) + \beta_2(\boldsymbol{X_2} - \bar{x}_2\boldsymbol{X_0}) + \boldsymbol{\epsilon}$$

Using the definition of the Z score vector, we can write the above equation as

$$SD_y\boldsymbol{Z_y} = \beta_1 SD_1\boldsymbol{Z_1} + \beta_2 SD_2\boldsymbol{Z_2} + \boldsymbol{\epsilon} \tag{40}$$

Taking the scalar product of the above equation with $\boldsymbol{Z_1}$ gives

$$SD_y r_{y1} = \beta_1 SD_1 + \beta_2 SD_2 r_{12}, \tag{41}$$

where $r_{y1} = \boldsymbol{Z_y}\cdot\boldsymbol{Z_1}/n$ is the correlation between $\boldsymbol{Y}$ and $\boldsymbol{X_1}$, $r_{12} = \boldsymbol{Z_1}\cdot\boldsymbol{Z_2}/n$ is the correlation between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. Equation (41) can be written as

$$\beta_1 = r_{y1}\frac{SD_y}{SD_1} - \beta_2\left(r_{12}\frac{SD_2}{SD_1}\right) = \beta_{y1} - \beta_2\beta_{21}, \tag{42}$$

where $\beta_{y1} = r_{y1}SD_y/SD_1$ is the slope in the simple regression for predicting $\boldsymbol{Y}$ from $\boldsymbol{X_1}$, and $\beta_{21} = r_{12}SD_2/SD_1$ is the slope in the simple regression for predicting $\boldsymbol{X_2}$ from $\boldsymbol{X_1}$.

Before deriving the final expression for β_1, let's point out two things. First, if $r_{12} = 0$ (or equivalently $\boldsymbol{X_1}\cdot\boldsymbol{X_2} = 0$) then $\beta_1 = \beta_{y1}$. That is to say that if $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are uncorrelated (orthogonal), the slope β_1 in the multiple regression is exactly the same as the slope β_{y1} in the simple regression for prediction $\boldsymbol{Y}$ from $\boldsymbol{X_1}$. The addition of the $\boldsymbol{X_2}$ does not change the slope. Second, if $r_{12} > 0$ and $\beta_2 > 0$, then $\beta_1 < \beta_{y1}$. The slope β_1 decreases if $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are positively correlated and if the slope of $\boldsymbol{X_2}$ in the multiple regression is positive.

Let's go back to equation (42). If $r_{12} \neq 0$, the equation of β_1 involves β_2. So β_1 and β_2 has to be solved together. Since $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are symmetric, the equation for β_2 can be obtained by simply exchanging the index between 1 and 2 of the β_1 equation:

$$\beta_2 = \beta_{y2} - \beta_1\beta_{12}, \tag{43}$$

where

$$\beta_{y2} = r_{y2}\frac{SD_y}{SD_2} \quad \text{and} \quad \beta_{12} = r_{12}\frac{SD_1}{SD_2}. \tag{44}$$

Substituting β_2 from (43) into (42), we obtain

$$\beta_1 = \beta_{y1} - (\beta_{y2} - \beta_1\beta_{12})\beta_{21} = \beta_{y1} - \beta_{y2}\beta_{21} + \beta_1\beta_{12}\beta_{21} \tag{45}$$

Note that the product

$$\beta_{12}\beta_{21} = \left(r_{12}\frac{SD_1}{SD_2}\right)\left(r_{12}\frac{SD_2}{SD_1}\right) = r_{12}^2.$$

Thus equation (45) becomes

$$\beta_1 = \beta_{y1} - \beta_{y2}\beta_{21} + r_{12}^2\beta_1 \quad \Rightarrow \quad (1 - r_{12}^2)\beta_1 = \beta_{y1} - \beta_{y2}\beta_{21}$$

or

$$\beta_1 = \frac{\beta_{y1} - \beta_{y2}\beta_{21}}{1 - r_{12}^2}$$

Using the expressions for β_{y1}, β_{y2} and β_{12}, we find

$$\beta_1 = b_1\frac{SD_y}{SD_1}, \tag{46}$$

where

$$b_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

may be interpreted as the adjusted correlation between $\boldsymbol{Y}$ and $\boldsymbol{X_1}$ taking into account the presence of $\boldsymbol{X_2}$. The expression for β_2 is obtained by exchanging the index between 1 and 2:

$$\beta_2 = b_2 \frac{SD_y}{SD_2} \quad , \quad b_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}.$$

Gathering all the results, we conclude that the regression coefficients are given by

$$\boxed{\beta_1 = b_1 \frac{SD_y}{SD_1} \quad , \quad \beta_2 = b_2 \frac{SD_y}{SD_1} \quad , \quad \beta_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2} \tag{47}$$

with

$$\boxed{b_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2} \quad , \quad b_2 = \frac{r_{y2} - r_{y1} r_{12}}{1 - r_{12}^2}} \tag{48}$$

In Section 1.8, we see that the RMSE in simple regression is related to SD_y by $RMSE = \sqrt{1 - r_{xy}^2}\, SD_y$. In multiple regression, we will show in Section 2.4 that the formula is generalized to

$$RMSE = \sqrt{1 - R^2}\, SD_y,$$

where R is the correlation between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$:

$$R = \frac{\boldsymbol{Z_y} \cdot \boldsymbol{Z_{\hat{y}}}}{n} \ .$$

We will defer the calculation of R to Section 2.3. In the case of multiple regression with two variables considered here, R is given by

$$\boxed{R = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 r_{y1} r_{y2} r_{12}}{1 - r_{12}^2}}} \tag{49}$$

2.2 More Than Two Variables

Suppose we now want to fit $\{y_i\}$ with p variables $\{x_{1i}, x_{2i}, \cdots, x_{pi}\}$. The regression equation has $p+1$ parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$. It can be written in the vector form as

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{\epsilon} = \sum_{j=0}^{p} \beta_j \boldsymbol{X_j} + \boldsymbol{\epsilon} \tag{50}$$

$$\hat{\boldsymbol{Y}} = \sum_{j=0}^{p} \beta_j \boldsymbol{X_j} \tag{51}$$

$$SSE = \boldsymbol{\epsilon} \cdot \boldsymbol{\epsilon} = \left(\boldsymbol{Y} - \sum_{j=0}^{p} \beta_j \boldsymbol{X_j} \right) \cdot \left(\boldsymbol{Y} - \sum_{j=0}^{p} \beta_j \boldsymbol{X_j} \right) \tag{52}$$

To minimize SSE, we set the derivatives of SSE with respect to β_j ($j = 0, 1, \cdots, p$) to 0. The resulting equations can be written as

$$\boldsymbol{X_j} \cdot \boldsymbol{\epsilon} = 0 \quad , \quad j = 0, 1, \cdots, p. \tag{53}$$

This means that $\boldsymbol{\epsilon}$ is orthogonal all $\boldsymbol{X_0}, \boldsymbol{X_1}, \cdots, \boldsymbol{X_p}$, or equivalently, the mean of $\boldsymbol{\epsilon}$ is 0 and $\boldsymbol{\epsilon}$ is uncorrelated with any of the variables we are trying to fit.

To find the regression coefficients, we follow the same procedures as before. First, take the scalar product of equation (50) with $\boldsymbol{X_0}$. The result is

$$\bar{y} = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_j \tag{54}$$

As before, this means that the regression line passes through the point of average. Next, we multiple the above equation by $\boldsymbol{X_0}$:

$$\bar{y}\boldsymbol{X_0} = \beta_0 \boldsymbol{X_0} + \sum_{j=1}^{p} \beta_j \bar{x}_j \boldsymbol{X_0}$$

and then subtract it from equation (50):

$$\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = \sum_{j=1}^{p} \beta_j (\boldsymbol{X_j} - \bar{x}_j \boldsymbol{X_0}) + \boldsymbol{\epsilon} \ .$$

Using the definition of the Z score vector (19), we can write the above equation as

$$SD_y \boldsymbol{Z_y} = \sum_{j=1}^{p} \beta_j SD_j \boldsymbol{Z_j} + \boldsymbol{\epsilon}.$$

Dividing both sides by SD_y results in

$$\boldsymbol{Z_y} = \sum_{j=1}^{p} b_j \boldsymbol{Z_j} + \frac{\boldsymbol{\epsilon}}{SD_y}, \tag{55}$$

where

$$b_j = \beta_j \frac{SD_j}{SD_y}. \tag{56}$$

Taking the scalar product of equation (55) with $\boldsymbol{Z_i}$ ($i = 0, 1, 2, \cdots, p$), we obtain

$$\sum_{j=1}^{p} r_{ij} b_j = r_{yi} \quad , \quad i = 1, 2, \cdots, p. \tag{57}$$

This is a system of linear equations for $b_1, b_2, \cdots, b_p$. Written them out, they look like

$$\begin{aligned}
b_1 + r_{12}b_2 + r_{13}b_3 + \cdots + r_{1p}b_p &= r_{y1} \\
b_1 r_{21} + b_2 + r_{23}b_3 + \cdots + r_{2p}b_p &= r_{y2} \\
b_1 r_{31} + b_2 r_{32} + b_3 + \cdots + r_{3p}b_p &= r_{y3} \\
\vdots \quad &\vdots \quad \vdots \\
b_1 r_{p1} + b_2 r_{p2} + b_3 r_{p3} + \cdots + b_p &= r_{yp}
\end{aligned}$$

If all of the variables $\boldsymbol{X_j}$ are uncorrelated (i.e. $r_{ij} = 0$ if $i \neq j$), the solution is $b_j = r_{yj}$ and the slopes are $\beta_j = r_{yj} SD_y / SD_j$. This is exactly the same as the slopes in simple regression for predicting $\boldsymbol{Y}$ from $\boldsymbol{X_j}$.

There are no simple analytic expressions for b_j in general, but there are several well-known procedures to obtain the solution by successive algebraic operations, but we will not discuss the methods here.

Suppose all the b's have been solved using one of those procedures, the slopes are given by equation (56) as

$$\beta_j = b_j \frac{SD_j}{SD_y} \quad , \quad j = 1, 2, \cdots, p \tag{58}$$

and the intercept β_0 is given by equation (54) as

$$\beta_0 = \bar{y} - \sum_{j=1}^{p} \beta_j \bar{x}_j. \tag{59}$$

Finally, we should mention that the term "linear" in linear regession refers to a model being linear in the fitting parameters $\beta_0, \beta_1, \cdots, \beta_p$. For example, we can fit y_i by the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \sqrt{x_i} + \beta_3 x_i^2 + \beta_4 \ln x_i + \beta_5 \frac{x_i^3}{1+2^{x_i}} \tag{60}$$

using the technique of multiple linear regression since the model is linear in $\beta_0, \beta_1, \cdots, \beta_5$. We simply label

$$x_{1i} = x_i \quad , \quad x_{2i} = \sqrt{x_i} \quad , \quad x_{3i} = x_i^2 \quad , \quad x_{4i} = \ln x_i \quad , \quad x_{5i} = \frac{x_i^3}{1+2^{x_i}}$$

and equation (60) can be written in vector form as

$$\boldsymbol{Y} = \beta_0 \boldsymbol{X_0} + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \beta_3 \boldsymbol{X_3} + \beta_4 \boldsymbol{X_4} + \beta_5 \boldsymbol{X_5} \ ,$$

which is equation (50) with $p = 5$. The key is to note that the least square prescription is to minimize SSE by varying the paremeters β's, not the independent variables x's.

2.3 Correlation Between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$

To generalize the $RMSE$ expression in Section 1.8 for multiple regression, we consider the quantity R defined as the correlation between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$:

$$R = \frac{\boldsymbol{Z_y} \cdot \boldsymbol{Z_{\hat{y}}}}{n} \ . \tag{61}$$

We first calculate the average of $\hat{\boldsymbol{Y}}$:

$$\bar{\hat{y}} = \frac{1}{n} \hat{\boldsymbol{Y}} \cdot \boldsymbol{X_0} = \frac{1}{n} \left(\beta_0 \boldsymbol{X_0} + \sum_{j=1}^{p} \beta_j \boldsymbol{X_j} \right) \cdot \boldsymbol{X_0} = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_j = \bar{y},$$

where we have used equation (51) for $\hat{\boldsymbol{Y}}$ and (54) for $\bar{y}$. From the definition of the Z score vector (19) and $\bar{\hat{y}} = \bar{y}$, we can write (61) as

$$\begin{aligned} R &= \frac{(\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})}{nSD_y SD_{\hat{y}}} \\ &= \frac{(\hat{\boldsymbol{Y}} + \boldsymbol{\epsilon} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})}{nSD_y SD_{\hat{y}}} \\ &= \frac{(\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})}{nSD_y SD_{\hat{y}}} \\ &= \frac{SD_{\hat{y}}}{SD_y} \end{aligned} \tag{62}$$

where we have used $\boldsymbol{\epsilon} \cdot \boldsymbol{X_0} = 0$ and $\boldsymbol{\epsilon} \cdot \hat{\boldsymbol{Y}} = 0$ (since $\hat{\boldsymbol{Y}}$ is a linear combination of $\boldsymbol{X_0}, \cdots \boldsymbol{X_p}$ and all are orthogonal to $\boldsymbol{\epsilon}$). We have also used the definition of the standard deviation (18) to obtain the last line. Hence we have

$$R^2 = \frac{SD_{\hat{y}}^2}{SD_y^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \ , \tag{63}$$

which is interpreted as the fraction of the variance of $\boldsymbol{Y}$ explained by the linear model.

To compute R, we use equation (51) for $\hat{\boldsymbol{Y}}$ and (59) for β_0, and write

$$\begin{aligned}
SD_{\hat{y}}^2 &= \frac{1}{n}(\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \\
&= \frac{1}{n}\left[(\beta_0 - \bar{y})\boldsymbol{X_0} + \sum_{i=1}^{p} \beta_i \boldsymbol{X_i}\right] \cdot \left[(\beta_0 - \bar{y})\boldsymbol{X_0} + \sum_{j=1}^{p} \beta_j \boldsymbol{X_j}\right] \\
&= \frac{1}{n}\left[\sum_{i=1}^{p} \beta_i (\boldsymbol{X_i} - \bar{x}_i \boldsymbol{X_0})\right] \cdot \left[\sum_{j=1}^{p} \beta_j (\boldsymbol{X_j} - \bar{x}_j \boldsymbol{X_0})\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^{p} SD_i \beta_i \boldsymbol{Z_i}\right) \cdot \left(\sum_{j=1}^{p} SD_j \beta_j \boldsymbol{Z_j}\right) \\
&= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{p} SD_i SD_j \beta_i \beta_j \boldsymbol{Z_i} \cdot \boldsymbol{Z_j} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} SD_i SD_j \beta_i \beta_j r_{ij}
\end{aligned}$$

$$R^2 = \frac{SD_{\hat{y}}^2}{SD_y^2} = \sum_{i=1}^{p}\sum_{j=1}^{p} b_i b_j r_{ij},$$

where we have used the definition of b_j in equation (56). The sum can be simplified by noting that b_j satisfy equation (57), and thus we obtain

$$\boxed{R = \sqrt{\sum_{i=1}^{p} b_i r_{yi}}} \tag{64}$$

When the solution of b_i is obtained, R can be calculated using the above equation. In the case of multiple regression with two variables ($p = 2$), b_1 and b_2 are given by equation (48). Plugging them into equation (64), we obtain equation (49).

2.4 Root Mean Square Error

As in Section 1.8, we define SST and SSM as

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \cdot (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0})$$

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})$$

It follows from the definition of standard deviation that

$$SST = nSD_y^2 \tag{65}$$

and it follows from (63) that

$$SSM = R^2 SST. \tag{66}$$

The identity $SST = SSM + SSE$ still holds in muiltiple linear regression. The proof is almost exactly the same as in Section 1.8.

We start with the identity

$$\boldsymbol{Y} - \bar{y}\boldsymbol{X_0} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) = \boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}).$$

Take the scalar product with itself:

$$\begin{aligned}(\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) \cdot (\boldsymbol{Y} - \bar{y}\boldsymbol{X_0}) &= [\boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})] \cdot [\boldsymbol{\epsilon} + (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})] \\ SST &= SSE + SSM + 2\boldsymbol{\epsilon} \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0})\end{aligned}$$

Since $\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}$ is a linear combination of $\boldsymbol{X_0}, \boldsymbol{X_1}, \cdots, \boldsymbol{X_p}$ and $\boldsymbol{\epsilon}$ is orthogonal to all these vectors, $\epsilon \cdot (\hat{\boldsymbol{Y}} - \bar{y}\boldsymbol{X_0}) = 0$ and so we have

$$\boxed{SST = SSM + SSE} \tag{67}$$

Combining equation (65), (66) and (67), we have $SSE = n(1 - R^2)SD_y^2$. It follows from the definition of $RMSE = \sqrt{SSE/n}$ that

$$\boxed{RMSE = \sqrt{1 - R^2}\, SD_y} \tag{68}$$

which is the generalization of equation (32).