Lecture Notes - 03: Linear Regression

Dihui Lai

March 17, 2020

Contents

1 Single Variate Linear Regression

A simple linear model can be formulated by assuming the target variable is dependent only on one predictor i.e.

$$\hat{y} = \beta_0 + \beta_1 x$$

In order to have our estimate as close as to the actual value of y, we want to find the βs that minimize the sum squared error function

$$\epsilon = \sum_{i=1}^{n}(y^i - \hat{y}^i)^2 = \sum_{i=1}^{n}(y^i - \beta_0 - \beta_1 x^i)^2 \tag{1}$$

i.e.

$$\begin{cases} \frac{\partial \epsilon}{\partial \beta_1} = 0 \\ \frac{\partial \epsilon}{\partial \beta_0} = 0 \end{cases} \tag{2}$$

$$\Rightarrow \begin{cases} \sum\limits_{i=1}^{n}(y^i - \beta_0 - \beta_1 x^i)x^i = 0 \\ \sum\limits_{i=1}^{n}(y^i - \beta_0 - \beta_1 x^i) = 0 \end{cases} \tag{3}$$

Sorting the equations to get the solution for β_0 and β_1

$$\Rightarrow \begin{cases} \beta_0 \sum\limits_{i=1}^{n} x^i = \sum\limits_{i=1}^{n} y^i x^i - \beta_1 \sum\limits_{i=1}^{n} x^i x^i \\ \sum\limits_{i=1}^{n} \beta_0 = \sum\limits_{i=1}^{n} y^i - \beta_1 \sum\limits_{i=1}^{n} x^i \end{cases}$$

$$\Rightarrow \begin{cases} \beta_1 = \frac{\sum\limits_{i=1}^{n} y^i x^i - \beta_0 \sum\limits_{i=1}^{n} x^i}{\sum\limits_{i=1}^{n} x^i x^i} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} y^i x^i - \beta_0 \bar{x}}{\frac{1}{n}\sum\limits_{i=1}^{n} x^i x^i} \\ \beta_0 = \frac{1}{n}(\sum\limits_{i=1}^{n} y^i - \beta_1 \sum\limits_{i=1}^{n} x^i) = (\bar{y} - \beta_1 \bar{x}) \end{cases} \tag{4}$$

Substitute β_0 in to the first equation in equation set (4). We have

$$\beta_1 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} y^i x^i - (\bar{y}\bar{x} - \beta_1 \bar{x}\bar{x})}{\frac{1}{n}\sum\limits_{i=1}^{n} x^i x^i}$$

Solving for β_1 we have

$$\beta_1 \frac{1}{n}\sum_{i=1}^{n} x^i x^i = \frac{1}{n}\sum_{i=1}^{n} y^i x^i - \bar{y}\bar{x} + \beta_1 \bar{x}\bar{x}$$

$$\Rightarrow \beta_1 \frac{1}{n}\sum_{i=1}^{n} x^i x^i = \frac{1}{n}\sum_{i=1}^{n} y^i x^i - \bar{y}\bar{x} + \beta_1 \bar{x}\bar{x}$$

$$\Rightarrow \beta_1 (\frac{1}{n}\sum_{i=1}^{n} x^i x^i - \bar{x}\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} y^i x^i - \bar{y}\bar{x}$$

Thus we get the solution of β_1

$$\beta_1 = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} y^i x^i - \bar{y}\bar{x}}{\frac{1}{n}\sum\limits_{i=1}^{n} x^i x^i - \bar{x}\bar{x}} \tag{5}$$

Take a close look it is not hard to find that the numerator is the co-variance of X and Y. The denominator is the variance of X.
Therefore β_1 can also be written as

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \rho_{XY}\frac{\sigma_X}{\sigma_Y} \tag{6}$$

2 Multivariate Linear Regression

Assume y is a linear superposition of multiple xs, the model for y is then formulated as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

or simply

$$\hat{y} = \sum_{j=1}^{m} \beta_j x_j = \vec{\beta} \cdot \vec{x}$$

To estimate βs that best fit the data, we need to minimize the error

$$\epsilon = \sum_{i=1}^{n} (y^i - \hat{y}^i)^2 \tag{7}$$

$$= \sum_{i=1}^{n} (y^i - \vec{x}^i \cdot \vec{\beta})^2 \tag{8}$$

Writing in matrix notation, we have

$$\begin{aligned}\epsilon &= (y-\hat{y})^T(y-\hat{y})\\ &= (y-X\beta)^T(y-X\beta)\end{aligned}$$

here β is a $m \times 1$ matrix defined as

$$\beta = \begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_m\end{bmatrix}$$

To minimize the error ϵ we want to the βs satisfy the following equation set:

$$\frac{\partial\epsilon}{\partial\beta_j} = 0, j = 1, 2, 3, 4...m$$

Using equation (7), we have

$$\sum_{i=1}^{n}\frac{\partial}{\partial\beta_j}(y^i-\hat{y}^i) = 0$$

$$\sum_{i=1}^{n}(y^i-\hat{y}^i)\frac{\partial\hat{y}^i}{\partial\beta_j} = 0 \tag{9}$$

$$\Rightarrow \sum_{i=1}^{n}(y^i-\hat{y}^i)x_j^i = 0 \tag{10}$$

Going from equation (9) to equation (10), we use the fact that $\hat{y} = \vec{x}\cdot\vec{\beta} = \sum_{l=1}^{n} x_l^i\beta_l$ and

$$\frac{\partial\hat{y}^i}{\partial\beta_j} = \frac{\partial}{\partial\beta_j}\sum_{l=1}^{n}x_l^i\beta_l = x_j^i \tag{11}$$

Write equation (10) in matrix format we have

$$(y-X\beta)^TX = 0$$

or after transposing

$$X^Ty - X^TX\beta = 0$$

Therefore the β that minimize ϵ has to satisfy the following equation

$$\beta = (X^TX)^{-1}X^Ty \tag{12}$$

Loosely speaking, we can interpreting equation (12) as composed two components the covariance related term X^Ty and a term that is related to the variance of X i.e. X^TX