Stat 151A: Linear Models Lecture Notes

Joseph Pagadora
Instructor: Oscar Padilla

Fall 2018

Contents

Chapter 1

Simple Linear Regression

Suppose we have only one explanatory variable X along with a response variable Y. The data is of the form $\{(x_1, y_1), ..., (x_n, y_n)\}$, where each $(x_i, y_i) \in \mathbb{R}^2$. With appropriate assumptions, we may model the relationship between the variables X and Y as follows:

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X, \text{ where } \beta_0, \beta_1 \in \mathbb{R}.$$

Alternatively, we can write

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \mathbb{E}[\epsilon_i|x_i] = 0, \text{ for } i = 1, ..., n.$$

Our strategy for finding good estimates for β_0, β_1 is to minimize the mean-squared error. That is, we solve

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

It is easy to see that this function is concave, and that its second derivative shows that the critical point is a minimum. When we set the partial derivatives to zero, we get the global minimum:

- $\displaystyle \frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$
- $\displaystyle \frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i$

Setting these to zero, we get

- $n b_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$
- $b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$

The equations above are known as the **normal equations**. Solving, we get

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \dfrac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{\sum_{i=1}^n (x_i - \overline{x})^2}, \\ \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \end{cases}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Observe that for these estimates to make sense, we need $\sum_{i=1}^{n}(x_i - \overline{x})^2 > 0$, that is, not all the x_i's are the same. The OLS regression line is given by $y = \hat{\beta}_0 + \hat{\beta}_1 x$.