Statistic Learning Protfolio: Linear Regression Analysis

Yu Yangcheng
yuyc23@mails.tsinghua.edu.cn

Tsinghua University— February 25, 2025

1 Simple Linear Regression

1.1 Building the structure

The simple linear regression(SLE) model's formula is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

Where Y_i is the response variable, X_i is the predictor variable, β_0 is the intercept, β_1 is the slope, and ε_i is the error term. The goal of simple linear regression is to estimate the coefficients β_0 and β_1 such that the error term ε_i is minimized. The error term ε_i is the difference between the observed value of the response variable and the value predicted by the model. The least squares method is used to estimate the coefficients. The least squares estimates(LSE) minimizes the sum of the squared errors. The sum of the squared errors is the sum of the squared differences between each observed value and the predicted value. Simple linear regression is built under 4 assumptions(LINE):

1. Linear Function: The mean of the response($E(Y_i)$), at each value of the predictor(X_i), is a linear function of X_i.
2. Independent: The errors(ε_i) are independent.
3. Normally Distributed: The errors(ε_i) are normally distributed.
4. Equal Variance: The errors(ε_i) have Equal variances.

Why should we have these assumptions? Firstly, Taylor expansion tells us that most of the functions can be approximated by a linear function, as long as the range is small enough. Secondly, independent errors make writting the joint distribution of the errors easier(We just need to do production). Thirdly, if we assume the errors are normally distributed, we can use the "3σ" principle to calculate the confidence interval. Finally, equal variance can reduce the complexity of the model.

> **Info:** Review what we just have learnt: SLR's LINE assumptions(**L**inear, **I**ndepent, **N**ormally distributed, **E**qual variance). In the following sections, we will take them as granted.

1.2 Estimate the coefficients

In SLR, we have **3** coefficients to estimate: β_0, β_1, and ε_i. The last one, ε_i, helps us to evaluate the model's performance. (Obviously, the smaller ε_i is, the better the model is.)
To estimate them, we will have 2 methods: **1.** least squares estimates(LSE), and **2.** Maximum Likelihood Estimates(MLE). The LSE method minimizes the sum of the squared errors, while the MLE method maximizes the likelihood function.

1.2.1 Least Squares Estimates

In the LSE method, we want to choose β_0 and β_1 to minimize

$$Q = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

Let's do calculus:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$
$$\frac{\partial Q}{\partial \beta_1} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) X_i = 0$$

Solve the equations, we get:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Replace $\hat{\beta}_0$ and $\hat{\beta}_1$ with b_0, b_1, we find:

$$\sum_{i=1}^{n} e_i = 0, \quad \sum_{i=1}^{n} X_i e_i = 0$$

Warning: The degress of freedom (df) of $e_1, e_2 \ldots e_n$ is $n-2$, because every time we add a constraint, we reduce the df by 1. Therefore,

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-1} = \frac{SSE}{df_E} = MSE$$

SSE is the short of Sum of Squared Errors, and MSE is the short of Mean Squared Errors.

1.2.2 Maximum Likelihood Estimates

In the MLE method, we want to maximize the likelihood function:

$$L(\beta_0, \beta_1, \sigma^2 | X_i, Y_i) = \prod_{i=1}^{n} f(Y_i | X_i; \beta_0, \beta_1, \sigma^2)$$

Where $f(Y_i|X_i; \beta_0, \beta_1, \sigma^2)$ is the probability density function of Y_i given X_i. We assume that Y_i is normally distributed, so we have:

$$f(Y_i|X_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

To get the MLE, we need to find $argmax_{\beta_0, \beta_1, \sigma^2} L(\beta_0, \beta_1, \sigma^2)$, that's to say, $argmax_{\beta_0, \beta_1, \sigma^2} - \log L(\beta_0, \beta_1, \sigma^2)$ Use calculus, we get:

$$\hat{\beta}_0^{ml} = b_0, \quad \hat{\beta}_1^{ml} = b_1,$$
$$\hat{\sigma}^{2^{ml}} = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

Why do we get different MSE(mean squared errors) in LSE and MLE? Because in MLE, we assume the normal distribution of Y_i at the beginning, which allow us to get n df. This also results in: ML estimates rely on distributional assumptions, while LS estimates do not.