Statistic Learning Protfolio: Linear Regression Analysis

Yu Yangcheng
yuyc23@mails.tsinghua.edu.cn

Tsinghua University— February 25, 2025

Introduction

This document is written for self-revision of the course Linear Regression Analysis and recognize my learning process. Therefore, it may not be a encyclopedic reference for this subject, but a personal record of my learning. Now, it's time for our journey to begin!

Info: Considering my learning background, in these documents, I will focus more on the math explanation of the concepts, and try to make the concepts more intuitive. I will also try to connect the concepts with the real world, and make the concepts more vivid.

1 History of Regression

1.1 Galton's Discovery

Linear regression is a statistical method that models the linear relationship between two variables. The word "Regression", was originate from Galton, Darwin's cousin, who first used it to describe the phenomenon that the offspring of parents with extreme characteristics tend to have characteristics that are closer to the average.

Galton's discovery comes from Galton board, also called "quincunx" in PPT, which is a device consist of different layers of pins. When a ball is dropped from the top, it will bounce off the pins and finally fall into the bottom. The final distribution of the balls is a normal distribution.

By observing this ball's distribution between different layers, Galton found that the balls tend to fall into the middle, and the distribution tends to be more and more distracted in the falling process. This leads to a paradox: if the offspring of parents with extreme characteristics would inherit their parents' characteristics, how could species keep stable? All in all, a species is recognized by some steady characteristics.

Galton explain this by introducing the concept of regression. He found that the offspring of parents with extreme characteristics tend to have characteristics that are closer to the average. It doesn't mean that the inheritance is not exist, taller parents' children may still be more possible to be tall, but the bias will be smaller. For example, if the father is 2 meters tall(like Yao ming), the son may be 1.9 meters tall(Yao ming's daughter seems to be very tall, too). In long-term development, the regression effect will contain the accumulation of the variation.

1.2 Regression & Correlation

Regression and Correlation are two closely related concepts. When you collected a series of data, you may want to draw the regression line to describe the relationship between the variables, and calculate the slope of the regression line. The slope of the regression line is always between -1 and 1, which is called the correlation coefficient. The slope's value have 3 cases:

- If the slope is 0, it means that x,y does't have linear relationship. **But** x,y could have non-linear relationship, eg. $y = x^2$ (prove it!).

- If the slop $\in (-1,1) \setminus \{0\}$, it means that x,y could have linear relationship. The closer the slope is to 1, the stronger the linear relationship is. The sign of the slope indicates the direction of the relationship.
- If the slop is ± 1, it means that x,y are completely linear related.

Notice: No matter what data you choose, you will always get a correlation coefficient between -1 and 1. This is **a game of mathematics**, having nothing to do with genes, magic, or anything else. In Galton's research, he studied the relationship between brother's height, and found the regression phenomenon as well. As we all know, it's unfair to clarify that brothers could inherit each other's excellent genes...

1.3 Math Explanation of Regression

If F is the n-1 th. generation's distribution, S is the n th. generation and X_n is the variation in the n th. turn. We assume that the species is stable, meaning that $Var(S) = Var(F)$. Let $\rho(F,S)$ be the correlation coefficient of two variables F and S. Then we have

$$\rho(F,S) = \frac{Cov(F,S)}{\sqrt{Var(F)Var(S)}} = \frac{Cov(F,F+X_n)}{Var(F)Var(S)} = \frac{Var(F)+Cov(F,X_n)}{Var(F)Var(S)} = 1 + \frac{Cov(F,X_n)}{Var(F)} < 1$$

The final step is based on the assumption made in front. The conclusion is that the relation between F and X_n is negative. This is the math explanation of regression.

1.4 Regression & Bivariate Normal Distribution

The bivariate normal distribution is a generalization of the normal distribution to two dimensions. It has two random variables, X and Y, which are normally distributed and have a correlation coefficient ρ. The joint probability density function of X and Y is given by

$$f(x,y) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1} \exp[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)}{\sigma_2^2}\right)]$$

Ususally, we use $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ to denote the bivariate normal distribution. μ and σ^2 means the mean/variation of x/y, and ρ means the correlation coefficient between x and y. In regression analysis, if we assume that x,y's marginal distribution is normal, then the joint distribution of x and y is bivariate normal. Conversely, the regression model can describe the linear property of bivariate distribution.

2 Simple Linear Regression

2.1 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.

2.2 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.

2.2.1 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 $e_1, e_2 \dots e_n$ 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.

2.2.2 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.