Introduction to Regression Analysis

Qiao Zhou*

August 8, 2016

Contents

*Haas School of Business, University of California, Berkeley. Email address: qiao_zhou@mfe.berkeley.edu.

1 Introduction

This is a technical note on regression analysis, with a focus on linear regressions. We will start with the introduction to classical linear regression (i.e., ordinary least squares regression), and move on to cover regularized regressions, such as Ridge, Lasso and Elastic Net. We will also relax certain classical assumptions and study the regression model's asymptotic behaviors when working with time-series financial data.

2 Ordinary Least Squares Regression

Let input $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathbb{R}^p$, outcome variable $y_i \in \mathbb{R}$, error term $\epsilon_i \sim N(0, \sigma^2)$, the linear regression formula can be expressed as

$$y = X\beta + \epsilon \tag{1}$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

The mean prediction of y conditional on observed X can be obtained by taking expectation on both sides of 1

$$E(y|X) = X\beta \tag{2}$$

2.1 OLS β Estimator

We find the β coefficients by minimizing the sum of squared errors (SSE)[4]:

$$\min_{\beta} SSE(\beta)$$

Where SSE can be expressed as

$$SSE(\beta) = \|y - X\beta\|_2 = (y - X\beta)^T(y - X\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

The first-order-condition (FOC) gives that

$$\begin{aligned} \frac{\partial SSE}{\partial \beta} &= \frac{\partial (y - X\beta)^T(y - X\beta)}{\partial \beta} \\ &= -2X^Ty + 2X^TX\beta = 0 \end{aligned}$$

Assuming that X has full column rank, then X^TX is positive definite (PD)[1]. The β estimator is

$$\hat{\beta} = (X^TX)^{-1}(X^Ty) \tag{3}$$

[1]In general, for randomly sampled real world data, this assumption is usually satisfied.

The estimator for outcome variable $\hat{y}$ is then

$$\hat{y} = X\hat{\beta} = X(X^TX)^{-1}(X^Ty) = Hy \tag{4}$$

Where H is known as the hat matrix[2] or projection matrix[3]:

$$H = X(X^TX)^{-1}X^T \tag{5}$$

Geometrically, $\hat{y}$[3] is the least square projection on the sub-space (hyper-plane) spanned by the feature vectors $\mathbf{x_j}$, $j = 1, 2, \ldots, p$ and $y - \hat{y}$ is the residual vector that is perpendicular to the sub-space.

2.1.1 Issues with Non-Full Rank

If columns of X are linearly dependent, then X^TX become singular, and β will not be uniquely defined in this case. As a result, the projection $\hat{y}$ will also not be unique. This colinearnarity issue can be resolved by dropping the dependent columns of X.

2.2 Properties of OLS Estimators

2.2.1 OLS Estimator is Unbiased

Under the assumption of exogenous errors $E[\epsilon|X] = 0$, it can be proven that the OLS estimator is unbiased[2]. First, we decompose $\hat{\beta}$ as

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}(X^Ty) \\ &= (X^TX)^{-1}X^T(X\beta + \epsilon) \\ &= \beta + (X^TX)^{-1}X^T\epsilon\end{aligned} \tag{6}$$

Then, substitute $\hat{\beta}$ with (6), we have

$$\begin{aligned}E[\hat{\beta} - \beta|X] &= E[\beta + (X^TX)^{-1}X^T\epsilon - \beta|X] \\ &= (X'X)^{-1}X'E[\epsilon|X] \\ &= 0\end{aligned}$$

Thus under the assumption of error exogeneity, $\hat{\beta}$ is an unbiased estimator:

$$E[\hat{\beta}] = E[E[\hat{\beta}|X]] = \beta \tag{7}$$

2.2.2 Variance Estimator for β and ϵ

We denote the variance-covariance matrix of $\hat{\beta}$ as V_β

$$\begin{aligned}V_\beta &= Var(\hat{\beta}|X) \\ &= Var(\beta + (X'X)^{-1}X'\epsilon) = Var((X'X)^{-1}X'\epsilon) \\ &= [(X'X)^{-1}X'\epsilon][(X'X)^{-1}X'\epsilon]' \\ &= (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} \\ &= (X'X)^{-1}X'DX(X'X)^{-1}\end{aligned} \tag{8}$$

[2]Because it operates on y and put a hat on it.

[3]Note that H transforms y to $\hat{y}$

Where the error covariance matrix is denoted as

$$D = E[\epsilon\epsilon^{'}|X] \tag{9}$$

In the homogeneous case,

$$V_{\beta} = (X'X)\sigma^2 \tag{10}$$

Usually, the error variance is not known and thus needs to be estimated from the data. An unbiased estimator of error variance σ^2 is

$$\hat{\sigma}_u^2 = \frac{1}{n-p}\sum_{i=1}^{n}\epsilon_i^2 \tag{11}$$

Thus, in the homoskedastic case, the variance-covariance matrix of $\hat{\beta}$ can be estimated as

$$\widehat{V}_{\widehat{\beta}}^{0} = (X'X)\hat{\sigma}_u^2 \tag{12}$$

In the heteroskedastic case, the error variance can be estimated as

$$\widehat{V}_{\widehat{\beta}}^{W} = (X'X)^{-1}(\sum_{i=1}^{n} x_i x_i^{'} \hat{e}_i{}^2)(X'X)^{-1} \tag{13}$$

Which is known as the heteroskedasticity-robust white estimator.

2.2.3 Distributional Properties of $\hat{\beta}$

For hypothesis testing purpose, we need to have the distribution of $\hat{\beta}$. Under the assumption of homoskedastic linear regression model and normally distributed errors $\epsilon_i|X \sim N(0, \sigma^2)$, we have that $\hat{\beta}$ is also normally distributed

$$\hat{\beta} - \beta \sim N(0, \sigma^2(X'X)^{-1}) \tag{14}$$

2.2.4 Distributional Properties of Error Variance

Given normally distributed errors $\epsilon_i|X \sim N(0, \sigma^2)$, it can be shown that

$$n\frac{\hat{\sigma}^2}{\sigma^2} = (n-k)\frac{\hat{\sigma}_u^2}{\sigma^2} \sim \chi^2_{n-p} \tag{15}$$

Where χ^2_{n-p} stands for chi-squared distribution with $n-p$ degree of freedom.

2.3 Hypothesis Testing and Confidence Intervals of $\hat{\beta}$

Under the distributional assumption of (14), we have for each $\hat{\beta}_j, \forall j \in [1, p]$

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p} \tag{16}$$

Where t_{n-p} stands for t-distribution with $(n-p)$ degrees of freedom (DF). Under the null hypothesis of $\beta_j = 0$, the following expression for t_j will follow t-distribution with DF of $(n-p)$[1]:

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{V_j^{1/2}\hat{\sigma}_u} =\sim t_{n-p} \tag{17}$$

Where $V_j^{1/2}$ is the j-th diagnoal element of $(X^T X)^{-1}$, $s.e.(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$. For a given significance level α, β_j is significantly different from 0 and the null hypothesis can thus be rejected if t_j exceeds the critical values[4].In matrix notation, the t-stats vector t can be represented as

$$t = diag(V_1^{1/2}\hat{\sigma}_u, V_2^{1/2}\hat{\sigma}_u, \ldots, V_p^{1/2}\hat{\sigma}_u)^{-1}\hat{\beta} \tag{18}$$

In addition, the confidence interval with significance interval $(1-\alpha)$ for $\hat{\beta}_j$ can be constructed as

$$(\beta_j - z^{1-\alpha/2}V_j^{1/2}\hat{\sigma}_u, \beta_j + z^{1-\alpha/2}V_j^{1/2}\hat{\sigma}_u) \tag{19}$$

Where $z^{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution.

[4]For a two-tailed t-test, the null hypothesis will be rejected if the test-statistics satisfies $t_j > t_{1-\alpha/2,n-p}$ or $t_j < -t_{1-\alpha/2,n-p}$

References

[1] Farinaz Koushanfar. Lecture 4,5- linear regression, rice elec 697. http://www.ece.rice.edu/~fk1/classes/ELEC697/Lec_4n5_LinearReg.pdf, 2006.

[2] Martin Lettau. Empirical methods in finance mfe230e, week 2: Asymptotic ols, estimation of ar processes.

[3] Jia Li. Linear, ridge regression, and principal component analysis.

[4] Xiaojin Zhu. Cs731 spring 2011 advanced artificial intelligence, linear regression. http://pages.cs.wisc.edu/~jerryzhu/cs731/regression.pdf, 2011.