Predicting the Closing Stock Price Using Linear Regression

Sitare Arslantürk

Introduction

The dynamic nature of stock market makes it difficult for the investors to predict the prices considering the external factors such as political situations or the public image of the company. In order to predict the closing prices of a company, for example Apple Inc.'s, statistical modelling was used. Linear Regression technique was performed for finding the close values from the open, high and low values.

Hypothesis

The main null hypothesis states that there is no relationship between the open, high, low variables and the output close values. The alternative hypothesis states that dependent variable close has a linear relationship to the independent variables of open, high, low.

Data Collection and Pre-processing

I have used Yahoo Finance to retrieve the stock price values of the Apple Inc. company from 2019 to 2020. The data consisted of Open, High, Low and Close attributes, each having its own value. Total data of 252 entries was collected as input for the linear regression model and saved into a csv file. The data collected was modified by selecting only the required columns and removing the unwanted data. Afterwards, the data was checked for invalid entries such as NaN values or null (empty) values. The invalid entries were removed by list-wise deletion. The test file written performs tests to validate the data input.

Linear Regression Model

The linear regression tries to produce the best possible predictions for the dataset by estimating the coefficients of the linear regression model, and then using these coefficients to find the prediction values for closing price. The standard error and credible intervals of 95% is also calculated. Table1 illustrates the code-generated resulting regression table. Figure 2 shows the Excel analysis of the data.

The mathematical calculation was derived as follows:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y},$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X'X})^{-1},$$

$$\sigma^2 = \frac{\boldsymbol{e'e}}{n-k-1},$$

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}},$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}.$$

Conclusion

The coefficients lay inside the range that the 95% credible (confidence) interval proposes. By looking at the credible intervals and also validating the results by checking the P-value associated with F statistics, I reject the null hypothesis.

B_hat	Standard Error	Lower 95%	Upper 95%
-0.625701	0.451992	-1.515935	0.264532
0.757666	0.053441	0.652410	0.862921
0.719667	0.048873	0.623408	0.815926
-0.474085	0.058895	-0.590083	-0.358086

Table 1: Regression Table

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.99946479					
R Square	0.998929866					
Adjusted R Square	0.998916921					
Standard Error	1.136683544					
Observations	252					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	299107.7631	99702.59	77166.23033	0	
Residual	248	320.4282708	1.292049			
Total	251	299428.1914				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.625701396	0.451992453	-1.38432	0.167505387	-1.515934728	0.264531935
High	0.757665684	0.053440693	14.17769	8.12E-34	0.652410196	0.862921171
Low	0.719667037	0.048873139	14.72521	1.08452E-35	0.623407692	0.815926383
Open	-0.474084541	0.058895097	-8.04964	3.46104E-14	-0.590082891	-0.358086191

Table 2: Excel Analysis