DSME5110
Statistical Analysis

Model to Predict Stock Price

Miao Yuhang 1155201284
Cai Yansheng 1155199262

Purpose

The project is to establish a model to predict the stock using indexes of the previous day.

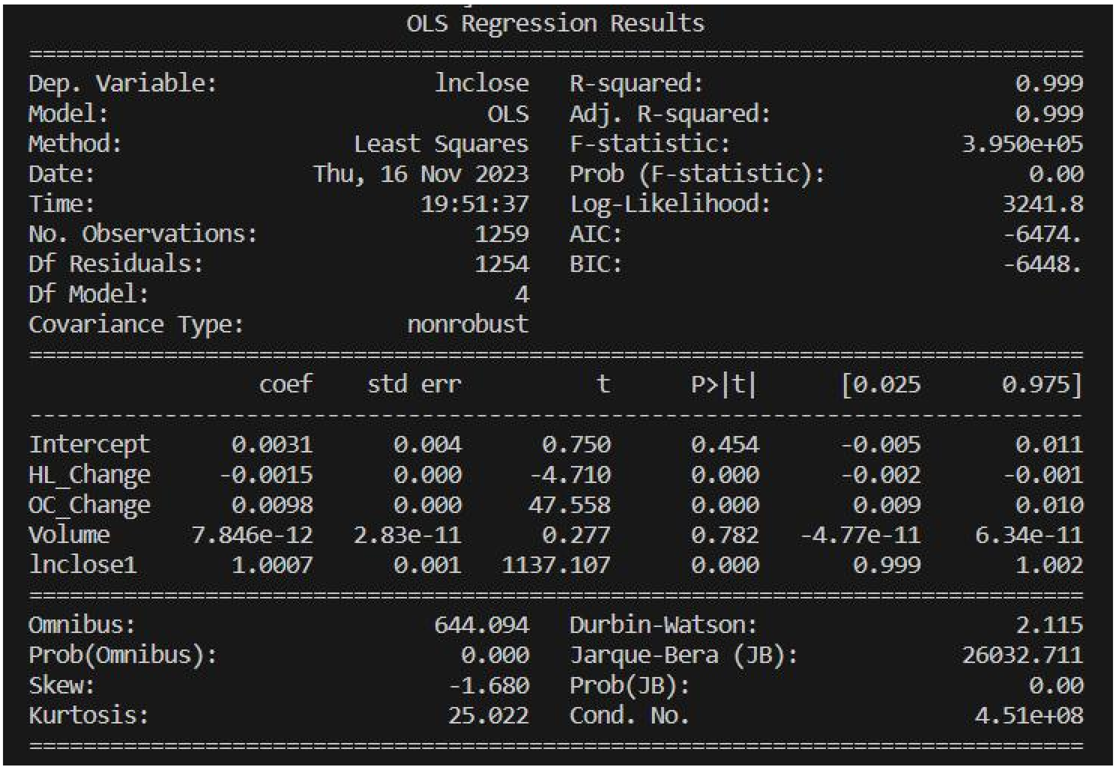
Data clean

Generally there won’t be null or duplicate value, the deletion of N/A and duplicate value is done.

OLS

Primarily an OLS model is established:

close = a + b*HL_Change + c*OC_Change + d*Volume + e*lnclose1



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                lnclose   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 3.950e+05
Date:                Thu, 16 Nov 2023   Prob (F-statistic):               0.00
Time:                        19:51:37   Log-Likelihood:                 3241.8
No. Observations:                1259   AIC:                            -6474.
Df Residuals:                    1254   BIC:                            -6448.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0031      0.004      0.750      0.454      -0.005       0.011
HL_Change     -0.0015      0.000     -4.710      0.000      -0.002      -0.001
OC_Change      0.0098      0.000     47.558      0.000       0.009       0.010
Volume      7.846e-12   2.83e-11      0.277      0.782   -4.77e-11    6.34e-11
lnclose1       1.0007      0.001   1137.107      0.000       0.999       1.002
==============================================================================
Omnibus:                      644.094   Durbin-Watson:                   2.115
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            26032.711
Skew:                          -1.680   Prob(JB):                         0.00
Kurtosis:                      25.022   Cond. No.                     4.51e+08
==============================================================================
```

HL_Change is (highest price – lowest price)/highest price, OC_Change is (open price – close price)/open price. Volume is the volume traded in that day and lnclose is ln(close price), and lnclose1 is ln(close price of previous day).

Accoding to the chart, the R-squared is 0.999 which is very high due to the lnclose1 is used to predict and is the base of the price in the second day. The HL_change has a negative and low impact which means the fluctuation of the price can lower the price. This may indicate that the fluctuation of price hurts the holders’ belief on the value of the company. OC_Change has a larger and positive impact which means if the previous the stock price descended for 1%, the ln(stock price) will ascend in the second day for 9.8*10^(-5). For the impact of ln(close1), the higher the previous day’s close price, so does the close price of the second day. This may result from the confidence of the holder or intended buyer on the stock. However, we can’t say the other two are different from 0 on 95% confidence interval.

Though the R-squared is high, it should be useless in this OLS case as the prediction is based on the close price of the previous day and the close price generally doesn’t change significantly.

Besides, the OLS can’t be tested using a test set which compare the prediction and the reality of a set of data unused in the regression, so we apply a machine learning model to the topic.

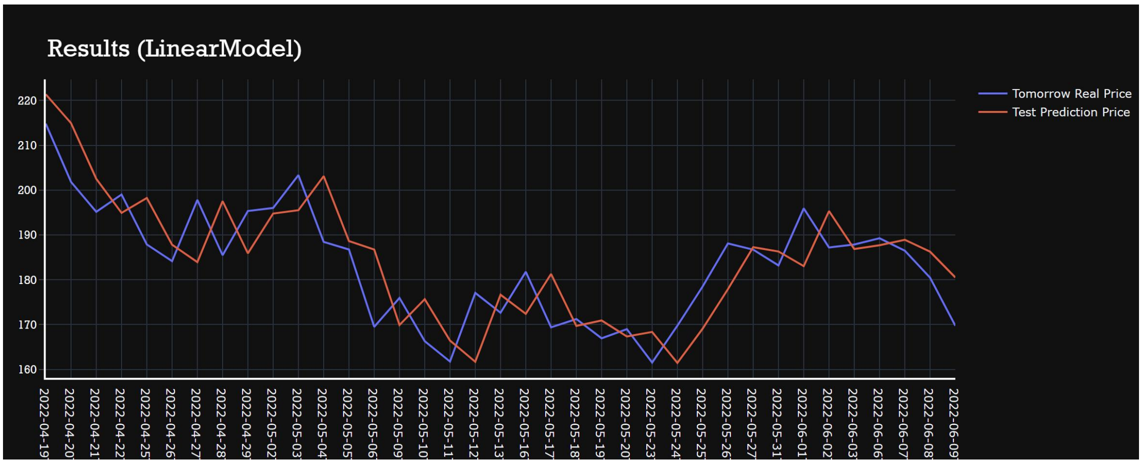
Linear fit

First is the training of the data. The termination conditions defined by the training and the functions is rendered below to prevent gradient explosion and gradient reduction, which are prevented by data compression and compensated by loss functions, the loss function maps random variable values to non-negative real numbers to reduce event risk and loss.

The termination condition is that the error exceeds a certain range or the number of times exceeds 3000, the training stops.

After training the model, it is necessary to process the data set, through the analysis of This stock price, three variables are defined, use the opening price, the highest price and the closing price in the original data, and obtain the change amount of the price by subtracting the opening price and the selling price, and obtain the change amount of one day by subtracting the highest price and the lowest price. At last, the four variables required are combined into a data frame.

Finally, the graph of this regression model is drawn according to time and compared with the actual graph, it can be seen that there is a certain lag in time. Although the linear fit is very good, there is a lag in time, because the data of the previous day is used to predict the next day, so there is a delay.



Therefore, for the defection in this model, some improvements can be made after searching the data. Through the time series model, a time fitting can be carried out by calculating the data of one week or one month ago. The advantages of using this model are as follows:Solving the problem of vanishing gradients, capture long-term dependencies and time series can be learned. This new model may be a good solution to the time lag and it is also the direction of improvement for this project.