Predicting stock returns using lags with regression models

Linzi Guan

2/18/2022

Introduction

We have always been wondering how the lagged stock returns explain the real stock return so in this report, supervised machine learning models with Ridge regressions are come up with to make predictions of stock returns using 5 previous-day stock returns. The entire model building process starts with a data overview, and is followed by a simple stock model built with only one stock and finally a general stock model built with several stocks after taking consideration of the fixed effect inside each stock through one-hot encoding. 5-fold cross validation is used for selecting the optimal parameter in the loss function. And this report will be structured by discussing data in the Data section, model in the Model section, results in the Result section, and limitations in the Limitation Section.

Data

The package "rugarch" in R and the data set "dji30ret" are used in this report for analysis, which contains the closing value log returns for the Dow Jones 30 constituents from 1987-03-16 to 2009-02-03. A brief summary of the data set is as follows:

```
##        AA                  AXP                  BA
##  Min.   :-0.2745595   Min.   :-0.3034304   Min.   :-0.1938568
##  1st Qu.:-0.0114593   1st Qu.:-0.0109291   1st Qu.:-0.0098007
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0001608   Mean   : 0.0001687   Mean   : 0.0003058
##  3rd Qu.: 0.0116377   3rd Qu.: 0.0114812   3rd Qu.: 0.0105709
##  Max.   : 0.2087337   Max.   : 0.1712035   Max.   : 0.1439727
##       BAC                   C                   CAT
##  Min.   :-0.3420588   Min.   :-0.3056056   Min.   :-0.244156
##  1st Qu.:-0.0093365   1st Qu.:-0.0111602   1st Qu.:-0.010575
##  Median : 0.0000000   Median : 0.0000000   Median : 0.000000
##  Mean   : 0.0001149   Mean   : 0.0000796   Mean   : 0.000378
##  3rd Qu.: 0.0100293   3rd Qu.: 0.0116803   3rd Qu.: 0.011141
##  Max.   : 0.2698774   Max.   : 0.4572902   Max.   : 0.137371
##       CVX                   DD                  DIS
##  Min.   :-0.1812526   Min.   :-0.2018984   Min.   :-0.3426451
##  1st Qu.:-0.0082829   1st Qu.:-0.0093255   1st Qu.:-0.0102932
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0004538   Mean   : 0.0001774   Mean   : 0.0002886
##  3rd Qu.: 0.0095239   3rd Qu.: 0.0095836   3rd Qu.: 0.0105821
##  Max.   : 0.1894765   Max.   : 0.1086964   Max.   : 0.1756133
##        GE                   GM                   HD
##  Min.   :-0.1947441   Min.   :-0.3727220   Min.   :-0.3386365
```

```
##  1st Qu.:-0.0083683   1st Qu.:-0.0119502   1st Qu.:-0.0115889
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0002751   Mean   :-0.0002715   Mean   : 0.0006922
##  3rd Qu.: 0.0093459   3rd Qu.: 0.0115692   3rd Qu.: 0.0127656
##  Max.   : 0.1275967   Max.   : 0.3009365   Max.   : 0.1315251
##       HPQ                  IBM                 INTC
##  Min.   :-0.2263815   Min.   :-0.268161   Min.   :-0.2488610
##  1st Qu.:-0.0125577   1st Qu.:-0.009243   1st Qu.:-0.0139915
##  Median : 0.0000000   Median : 0.000000   Median : 0.0000000
##  Mean   : 0.0003792   Mean   : 0.000249   Mean   : 0.0005553
##  3rd Qu.: 0.0135366   3rd Qu.: 0.009469   3rd Qu.: 0.0158734
##  Max.   : 0.1591410   Max.   : 0.123635   Max.   : 0.2265276
##       JNJ                  JPM                  AIG
##  Min.   :-0.2043813   Min.   :-0.3234769   Min.   :-0.9362581
##  1st Qu.:-0.0078036   1st Qu.:-0.0111299   1st Qu.:-0.0089217
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0004993   Mean   : 0.0002586   Mean   :-0.0002978
##  3rd Qu.: 0.0084695   3rd Qu.: 0.0111094   3rd Qu.: 0.0094814
##  Max.   : 0.1153126   Max.   : 0.2239172   Max.   : 0.3585320
##        KO                   MCD                  MMM
##  Min.   :-0.2828628   Min.   :-0.1827990   Min.   :-0.2257926
##  1st Qu.:-0.0080020   1st Qu.:-0.0093024   1st Qu.:-0.0074074
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0004333   Mean   : 0.0004514   Mean   : 0.0003322
##  3rd Qu.: 0.0089027   3rd Qu.: 0.0099834   3rd Qu.: 0.0082499
##  Max.   : 0.1791113   Max.   : 0.1030806   Max.   : 0.1049975
##       MRK                  MSFT                 PFE
##  Min.   :-0.3119154   Min.   :-0.379490   Min.   :-0.1892420
##  1st Qu.:-0.0090772   1st Qu.:-0.010643   1st Qu.:-0.0096386
##  Median : 0.0000000   Median : 0.000000   Median : 0.0000000
##  Mean   : 0.0003447   Mean   : 0.000787   Mean   : 0.0003885
##  3rd Qu.: 0.0100307   3rd Qu.: 0.012848   3rd Qu.: 0.0107528
##  Max.   : 0.1224923   Max.   : 0.178465   Max.   : 0.0989399
##        PG                    T                   UTX
##  Min.   :-0.3598917   Min.   :-0.1352280   Min.   :-0.3029024
##  1st Qu.:-0.0074074   1st Qu.:-0.0087377   1st Qu.:-0.0083770
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0004917   Mean   : 0.0003364   Mean   : 0.0004517
##  3rd Qu.: 0.0085107   3rd Qu.: 0.0096619   3rd Qu.: 0.0098064
##  Max.   : 0.1980376   Max.   : 0.1505242   Max.   : 0.1278841
##        VZ                   WMT                  XOM
##  Min.   :-0.1931448   Min.   :-0.1249721   Min.   :-0.2676996
##  1st Qu.:-0.0087802   1st Qu.:-0.0100137   1st Qu.:-0.0078818
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0002848   Mean   : 0.0004985   Mean   : 0.0004964
##  3rd Qu.: 0.0089366   3rd Qu.: 0.0104713   3rd Qu.: 0.0090419
##  Max.   : 0.1365130   Max.   : 0.1146918   Max.   : 0.1653925
```

There are 5521 observations and 30 stocks inside the data set.

Model

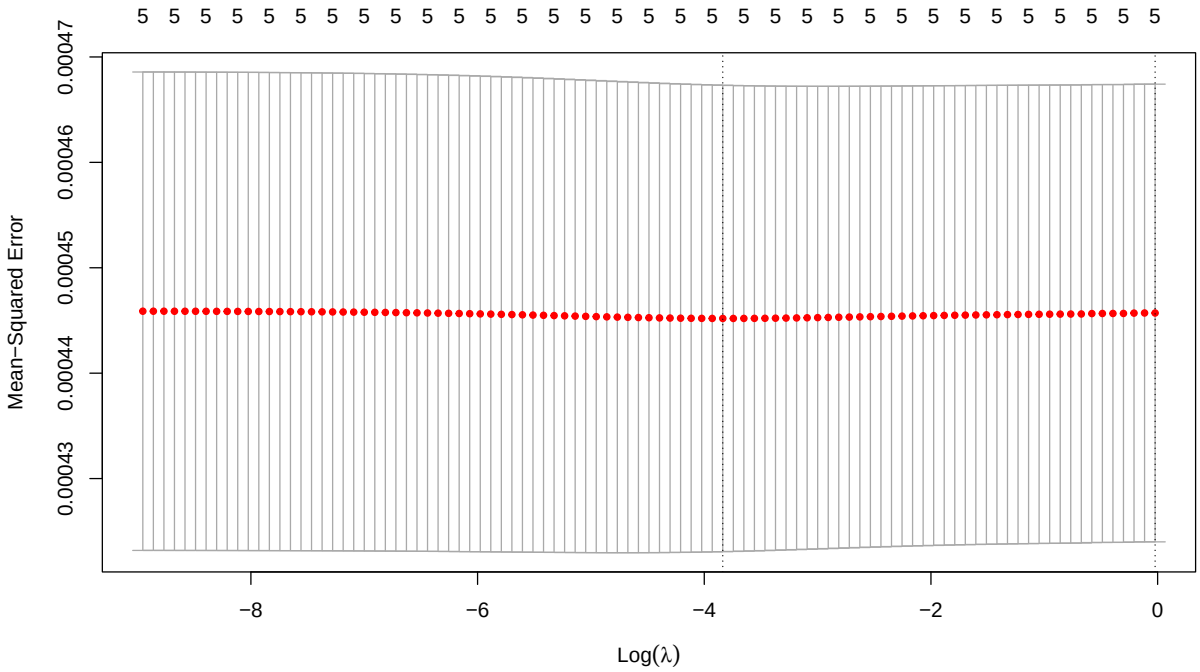
The simple model using only "AA" stock

To start a simple model, we take the Alcoa Corporation stock as an instance first, which is denoted by "AA" and we create 5 lagged stock returns as inputs for our model and use the real stock return as the output. A brief summary of the variables of the model is as follows:

```
##    AA_return             lag_1              lag_2
##  Min.   :-0.2745595   Min.   :-0.274560   Min.   :-0.2745595
##  1st Qu.:-0.0114625   1st Qu.:-0.011463   1st Qu.:-0.0114461
##  Median : 0.0000000   Median : 0.000000   Median : 0.0000000
##  Mean   : 0.0001576   Mean   : 0.000149   Mean   : 0.0001552
##  3rd Qu.: 0.0116305   3rd Qu.: 0.011618   3rd Qu.: 0.0116305
##  Max.   : 0.2087337   Max.   : 0.208734   Max.   : 0.2087337
##      lag_3                lag_4                lag_5
##  Min.   :-0.2745595   Min.   :-0.2745595   Min.   :-0.2745595
##  1st Qu.:-0.0114343   1st Qu.:-0.0114295   1st Qu.:-0.0114343
##  Median : 0.0000000   Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0001697   Mean   : 0.0001807   Mean   : 0.0001733
##  3rd Qu.: 0.0116305   3rd Qu.: 0.0116378   3rd Qu.: 0.0116305
##  Max.   : 0.2087337   Max.   : 0.2087337   Max.   : 0.2087337
```

The data to date 2002-12-31 are used as the training data and the data from date 2003-01-02 to date 2009-02-03 are used as the testing data. The training set is fit into a ridge regressing to predict the target variable using the features(the lagged five returns).

The optimal lambda of the Ridge regression is found via 5-fold cross validation and a plot of the lambda parameter vs. the Mean Squared Error is shown as follows:



The optimal lambda parameter is 0.02156627 and the model after fitting the Ridge regression using the optimal parameter chosen above (use the entire Training Set) is summarized as follows:

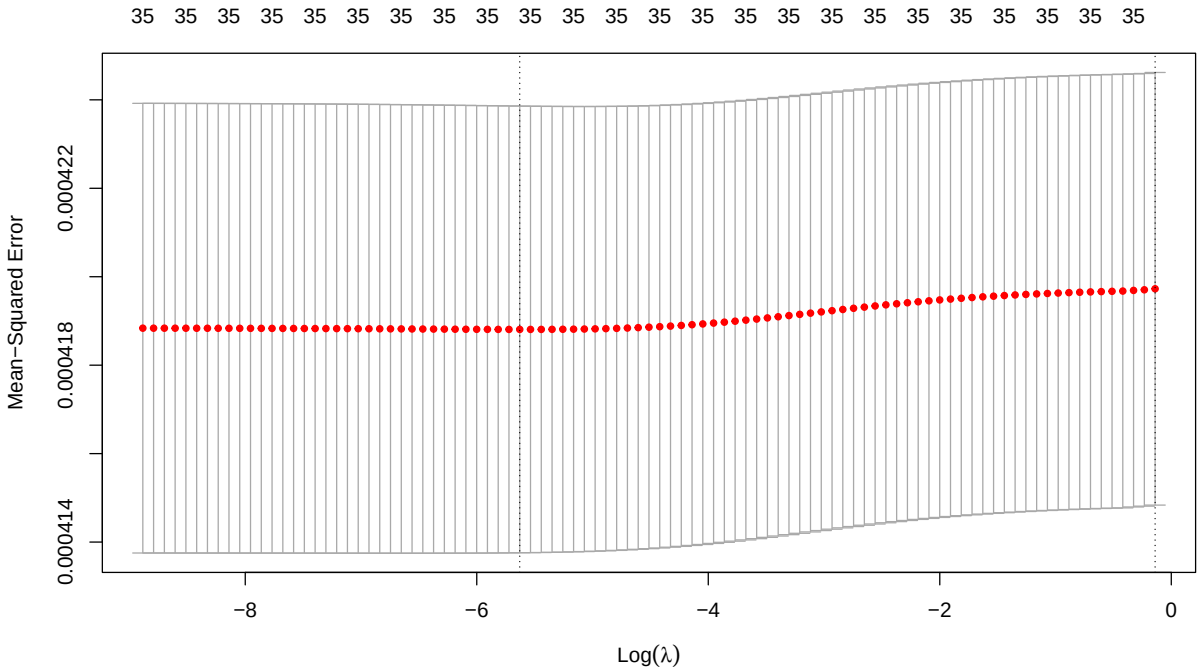
```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      5      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

The general model using all the stocks

Similarly, 5 lagged stock returns are added as input variables. Besides this, dummy variables are added to realize one-hot encoding to get rid of the fix effects in different stocks. After preprocessing, the data frame has 165480 observations with 36 columns(5 lagged values, 1 true return and 30 dummy columns).

Also, the data to date 2002-12-31 are used as the training data and the data from date 2003-01-02 to date 2009-02-03 are used as the testing data. The training set is fit into a ridge regressing to predict the target variable using the features(the lagged five returns and dummy variables).

The optimal lambda of the Ridge regression is found via 5-fold cross validation and a plot of the lambda parameter vs. the Mean Squared Error is shown as follows:



The optimal lambda chosen in this model is 0.003593454 and after fitting the Ridge regression using the optimal parameter chosen above (use the entire Training Set), a summary of the fitted model is as follows:

```
##           Length Class     Mode
```

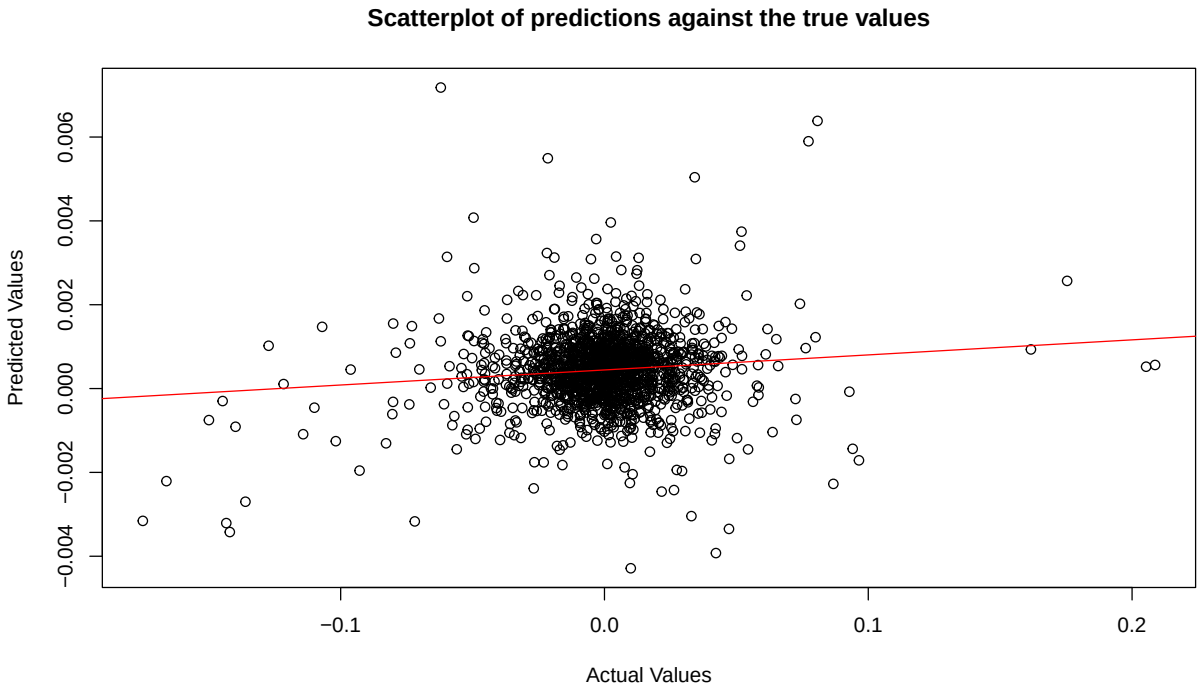
```
## a0         1     -none-    numeric
## beta      35     dgCMatrix S4
## df         1     -none-    numeric
## dim        2     -none-    numeric
## lambda     1     -none-    numeric
## dev.ratio  1     -none-    numeric
## nulldev    1     -none-    numeric
## npasses    1     -none-    numeric
## jerr       1     -none-    numeric
## offset     1     -none-    logical
## call       5     -none-    call
## nobs       1     -none-    numeric
```

Result

The simple model using only "AA" stock

The mean absolute error using the fitted model above to predict the returns of AA in the Test Set is 0.01772329.

The scatterplot of predictions against the true values is as follows:

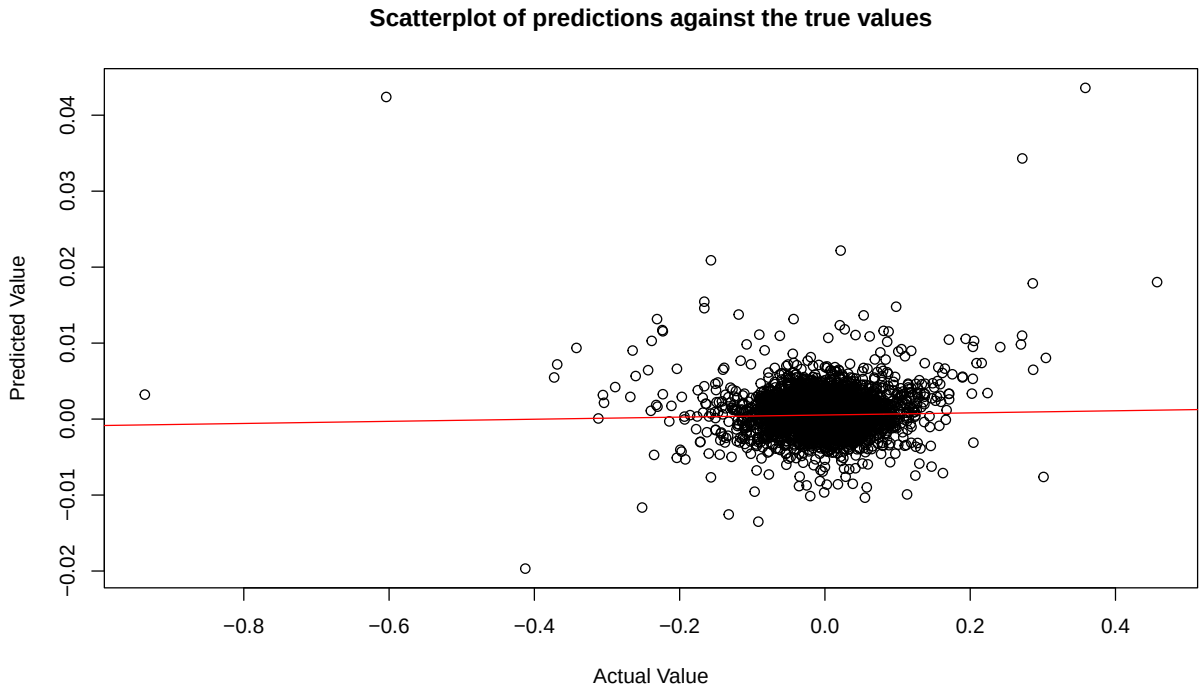


Conclusions can be drawn from the scatterplot that the model does not fit quite well from the scatterplot as the predicted values are not that close to actual values (the line is not close to a regressed diagonal line)

The general model using all the stocks

The mean absolute error using the fitted model above to predict the returns of all stocks in the Test Set is 0.0127128.

The scatterplot of predictions against the true values is as follows:



Compared to the previous predicted value vs actual value graph, this model does not have a significant improvement from the above although MAE decreases since the predicted value range is still much narrower than the actual value range and the line is not regressed diagonally.

Limitations and next steps

There are several limitations in the models. One is multicollinearity and next step could be setting a baseline of the variables and dropping one dummy variable to avoid multicollinearity. Furthermore, if we consider using dummy variables to improve the model, we could incorporate factors that impact stock behaviors such as the size of the company(whether it is a large company or small company indicated by P/E ratio), whether the company is a value company or growth company, and so on. We can add dummy variables in these aspects and train our model to see whether there would be some improvements for the fitted model.

Reference

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).