STAT 306: Assignment

Emad Zadegan: 27452119
Michael Nguyen: 38223111
Chingiz Khalifazadeh: 21059134
Kevin Zhu: 38726139

April 7th, 2016

0.1 Abstract/Summary

The goal of this report is to form a prediction equation for the next day closing price of the US Dollar/Canadian Dollar exchange rate. We gathered data from investing.com as well as the MetaTrader trading platform. We used SQL to join the data across sources based on date, as well as form variables such as the volatility variables (named `... volatility` in R code) and the response variable, which is the US Dollar/Canadian Dollar exchange rate of the following day (called `response` in the R code). We omitted the details of joining and creation of variables code for sake of brevity for the reader.

0.2 Initial Analysis

Here we analyze the raw data to gain intuition on any scaling or transformations that may need to be performed before attempting model fitting.

```
> dat = read.table('matrix_final_5.csv', sep='\t', header=T)
> # remove the last row since it does not have a valid response entry
> # since we did not collect the US Dollar/Canadian Dollar exchange
> # rate of the  following date for the last entry
> dat=dat[-dim(dat)[1],]
> names(dat)
```

```
 [1] "Date"                      "response"
 [3] "EUR.USD.High"              "EUR.USD.Low"
 [5] "EUR.USD.volatility"        "EUR.USD.Close"
 [7] "Overnight.Rate"            "SP500.High"
 [9] "SP500.Low"                 "SP500.volatility"
[11] "SP500.Close"               "SPTSX.Close"
[13] "SPTSX.High"                "SPTSX.Low"
[15] "SPTSX.volatility"          "SPTSX.Volume.in.Millions"
[17] "USD.CAD.High"              "USD.CAD.Low"
[19] "USD.CAD.volatility"        "USD.CAD.Close"
[21] "wti.crude.oil.spot.close" "YUAN.USD.Close"
[23] "YUAN.USD.High"             "YUAN.USD.Low"
[25] "YUAN.USD.volatility"
```

0.2.1 Summary Statistics

We choose to only work with a subset of the data, due to a lot of variables in the original data set being intuitively correlated such as the high, low and close price of a particular index on a given day.

```
> subset_expl=dat[,c(2,5,6,7,10,11,12,15,16,19,20,21,22,25)]
> summary(subset_expl)
```

```
    response     EUR.USD.volatility EUR.USD.Close   Overnight.Rate
 Min.   :0.983   Min.   :0.00170    Min.   :1.049   Min.   :0.4932
 1st Qu.:1.046   1st Qu.:0.00650    1st Qu.:1.125   1st Qu.:0.7487
 Median :1.103   Median :0.00892    Median :1.300   Median :0.9987
 Mean   :1.145   Mean   :0.01003    Mean   :1.251   Mean   :0.8747
 3rd Qu.:1.248   3rd Qu.:0.01230    3rd Qu.:1.352   3rd Qu.:1.0016
 Max.   :1.458   Max.   :0.04652    Max.   :1.393   Max.   :1.0202
 SP500.volatility  SP500.Close    SPTSX.Close    SPTSX.volatility
 Min.   : 3.70    Min.   :1457   Min.   :11837   Min.   : 25.22
 1st Qu.:10.48    1st Qu.:1705   1st Qu.:12824   1st Qu.: 76.38
 Median :15.32    Median :1924   Median :13878   Median :111.50
 Mean   :18.25    Mean   :1879   Mean   :13866   Mean   :126.02
 3rd Qu.:22.53    3rd Qu.:2052   3rd Qu.:14765   3rd Qu.:153.68
 Max.   :98.14    Max.   :2131   Max.   :15658   Max.   :691.59
 SPTSX.Volume.in.Millions USD.CAD.volatility USD.CAD.Close
```

```
Min.   : 37.86           Min.   :0.001400   Min.   :0.983
1st Qu.:151.68           1st Qu.:0.004970   1st Qu.:1.046
Median :175.00           Median :0.007520   Median :1.103
Mean   :185.14           Mean   :0.008457   Mean   :1.145
3rd Qu.:207.20           3rd Qu.:0.010740   3rd Qu.:1.247
Max.   :795.34           Max.   :0.038800   Max.   :1.458
wti.crude.oil.spot.close YUAN.USD.Close  YUAN.USD.volatility
Min.   : 26.68           Min.   :6.041   Min.   :0.000000
1st Qu.: 51.41           1st Qu.:6.131   1st Qu.:0.004000
Median : 93.12           Median :6.203   Median :0.006500
Mean   : 78.78           Mean   :6.207   Mean   :0.008319
3rd Qu.: 99.60           3rd Qu.:6.233   3rd Qu.:0.010100
Max.   :110.62           Max.   :6.596   Max.   :0.081200
```

0.2.2 Description of Data

We collect data from various within the time range of January 1st 2013, to January 28th 2016. As explained before the original set of variables has been subsetted due to many variables being from the same source.

Variables	explanation of units
EUR/USD Close	Exchange rate of Euro to the US dollar
Overnight Rate	Overnight interbank lending rate
S&P500 Close	Standard & Poor's index of 500 top US companies by market capitalization.
S&P TSX Close	Standard & Poor's index of top Canadian companies by market capitalization.
S&P TSX Volume	Volume of contracts traded on the TSX during the trading day in millions of units
WTI Crude Oil Spot Price	Closing price for a contract specifying the price of a barrel of crude oil commodity.
Yuan/USD Close	Exchange rate of Chinese Yuan to US dollar
EUR/USD, S&P500, S&PTSX, USD/CAD Volatility	Difference of High & Low within a day, in market points

The next two figures cover visual aids for the data.

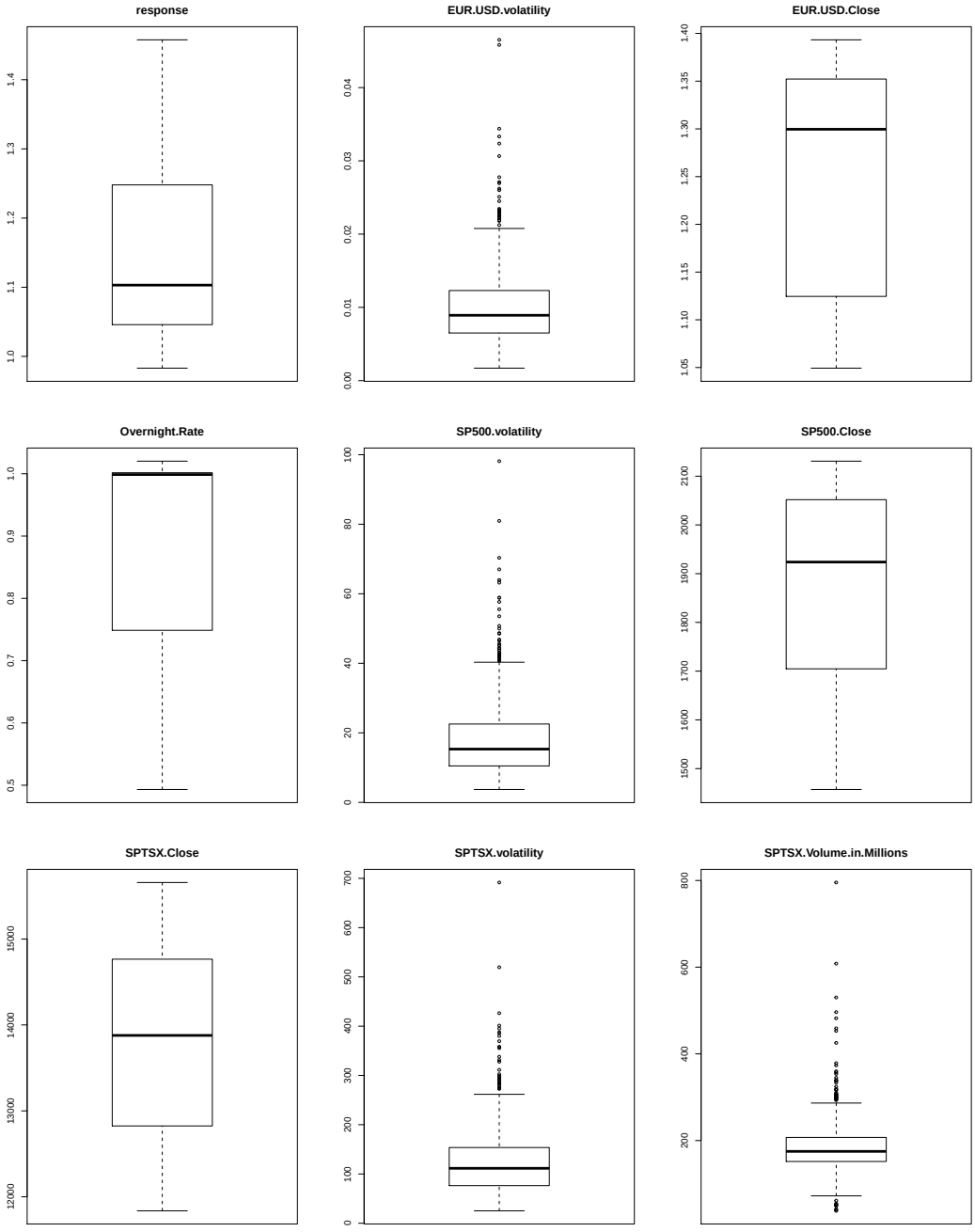


Figure 1: Box plot for the first response and the first 8 explanatory variables

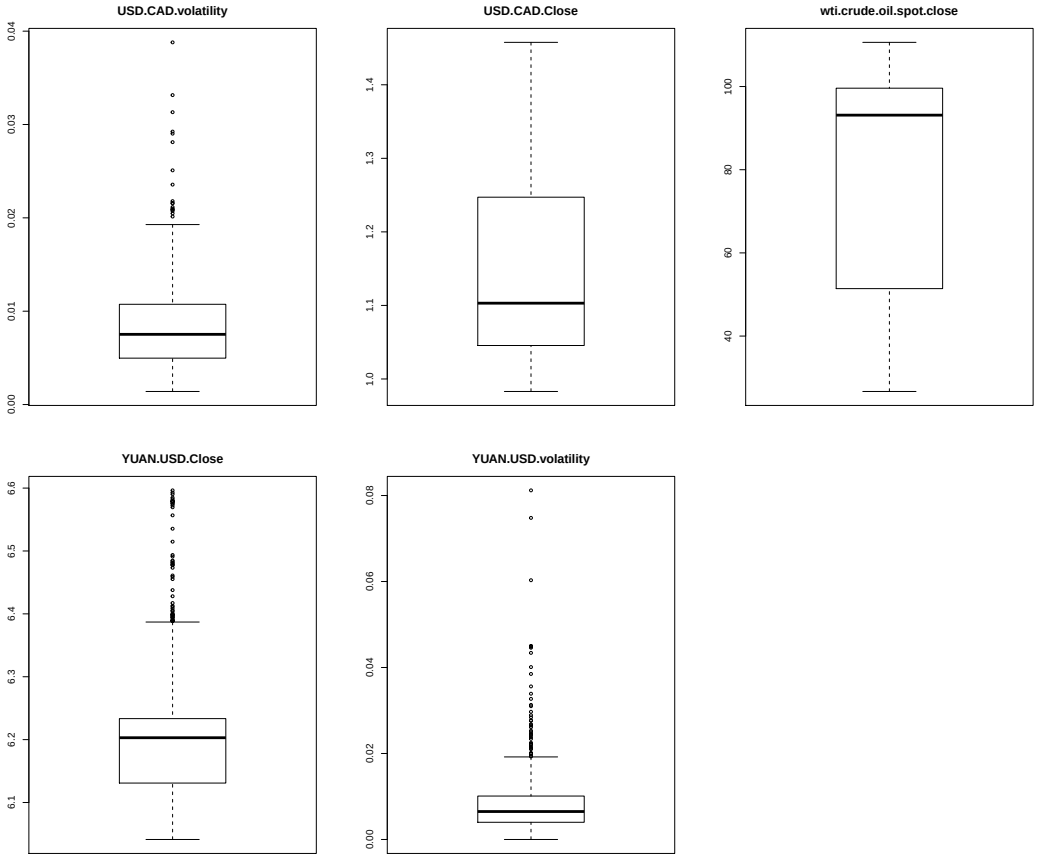


Figure 2: Boxplots for 5 explanatory variables

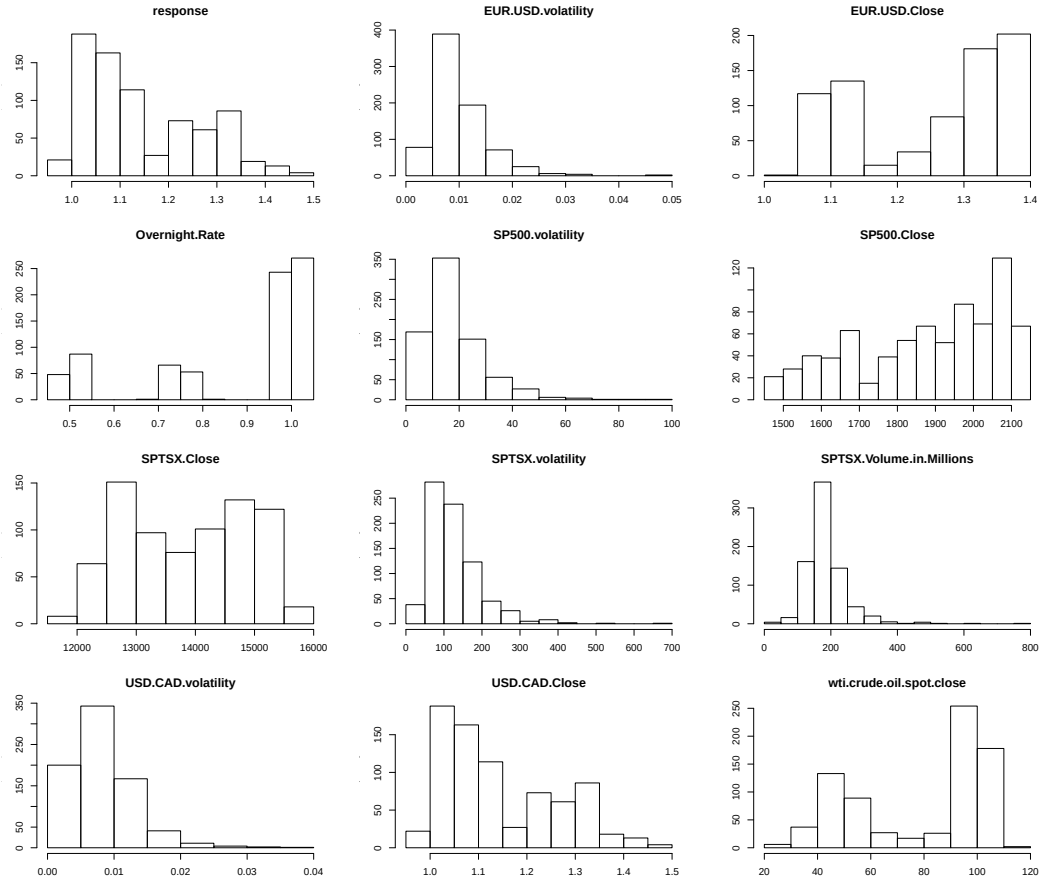


Figure 3: Histogram of response and all explanatory variables

0.2.3 Correlation Matrix

Correlation between response and all explanatory variables. We notice very high correlation between the response and USD/CAD closing exchange rate, this is explained from the response being created through a shifting of USD/CAD closing exchange rate. There is also strong negative correlation between oil prices (`wti.crude.oil.spot.close`) and the response; this could imply an inverse relationship between US/Canadian exchange rate and oil prices in US Dollars.

```
> mat=matrix(cor(subset_expl),ncol=dim(subset_expl)[2])
> mat
```

```
           [,1]        [,2]       [,3]        [,4]       [,5]       [,6]
 [1,]  1.0000000  0.23477148 -0.8870483 -0.92252918  0.4712699  0.7505648
 [2,]  0.2347715  1.00000000 -0.3755660 -0.24745509  0.3571020  0.1034859
 [3,] -0.8870483 -0.37556597  1.0000000  0.86841690 -0.3674406 -0.6545406
 [4,] -0.9225292 -0.24745509  0.8684169  1.00000000 -0.4028152 -0.5652675
 [5,]  0.4712699  0.35710197 -0.3674406 -0.40281516  1.0000000  0.1773545
 [6,]  0.7505648  0.10348593 -0.6545406 -0.56526754  0.1773545  1.0000000
 [7,]  0.2556492 -0.02364067 -0.1999367 -0.05599272 -0.1009093  0.7691865
 [8,]  0.4151010  0.31723094 -0.3818118 -0.32737189  0.7062335  0.2421481
 [9,]  0.3071224  0.25683726 -0.2863140 -0.23362094  0.4072460  0.1749299
[10,]  0.5786644  0.45475951 -0.5800125 -0.50463210  0.4331223  0.4091634
[11,]  0.9988795  0.23169308 -0.8873667 -0.92326415  0.4706371  0.7492881
[12,] -0.9284169 -0.34317108  0.9519876  0.87358498 -0.4577087 -0.6636018
[13,]  0.7605435  0.12605861 -0.6496489 -0.78457950  0.4342941  0.3263918
[14,]  0.2152695  0.08289132 -0.1270722 -0.16855513  0.2061758  0.1902816
            [,7]        [,8]        [,9]       [,10]      [,11]      [,12]
 [1,]  0.25564922  0.41510102  0.30712236  0.5786644  0.9988795 -0.9284169
 [2,] -0.02364067  0.31723094  0.25683726  0.4547595  0.2316931 -0.3431711
 [3,] -0.19993673 -0.38181178 -0.28631397 -0.5800125 -0.8873667  0.9519876
 [4,] -0.05599272 -0.32737189 -0.23362094 -0.5046321 -0.9232641  0.8735850
 [5,] -0.10090929  0.70623345  0.40724603  0.4331223  0.4706371 -0.4577087
 [6,]  0.76918651  0.24214807  0.17492995  0.4091634  0.7492881 -0.6636018
 [7,]  1.00000000 -0.02460196 -0.02102552  0.1407235  0.2556861 -0.1728773
 [8,] -0.02460196  1.00000000  0.47162351  0.4091654  0.4145368 -0.4517857
 [9,] -0.02102552  0.47162351  1.00000000  0.4085876  0.3066406 -0.3661144
[10,]  0.14072346  0.40916541  0.40858755  1.0000000  0.5794168 -0.5956941
[11,]  0.25568610  0.41453684  0.30664058  0.5794168  1.0000000 -0.9277661
[12,] -0.17287727 -0.45178567 -0.36611436 -0.5956941 -0.9277661  1.0000000
[13,] -0.08354976  0.31431574  0.25362161  0.3639090  0.7615777 -0.7116930
[14,]  0.09084870  0.15845977  0.14058375  0.1658971  0.2153524 -0.1943676
           [,13]       [,14]
 [1,]  0.76054351  0.21526953
 [2,]  0.12605861  0.08289132
 [3,] -0.64964893 -0.12707221
 [4,] -0.78457950 -0.16855513
 [5,]  0.43429414  0.20617576
 [6,]  0.32639185  0.19028156
 [7,] -0.08354976  0.09084870
 [8,]  0.31431574  0.15845977
 [9,]  0.25362161  0.14058375
[10,]  0.36390904  0.16589712
[11,]  0.76157767  0.21535236
[12,] -0.71169300 -0.19436764
[13,]  1.00000000  0.24962867
[14,]  0.24962867  1.00000000
```

```
> # summary statistics of the matrix
> # removed values the diagonal values, which are the only values equal to 1
```

```
> mat_vals=c(mat)[which(c(mat) != 1)]
> summary(mat_vals)

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.92840 -0.33920  0.16590  0.04845  0.40830  0.99890
```

where each number represents, the ith column of the subset data set or again:

```
> names(subset_expl)

 [1] "response"                 "EUR.USD.volatility"
 [3] "EUR.USD.Close"            "Overnight.Rate"
 [5] "SP500.volatility"         "SP500.Close"
 [7] "SPTSX.Close"              "SPTSX.volatility"
 [9] "SPTSX.Volume.in.Millions" "USD.CAD.volatility"
[11] "USD.CAD.Close"            "wti.crude.oil.spot.close"
[13] "YUAN.USD.Close"           "YUAN.USD.volatility"
```

0.2.4 Initial Analysis Summary

After summary statistics and intial analysis the following scaling was done to avoid small regression coefficients. We attempted to scale all variables to be in the range 0 and 10.

```
> tdat=subset_expl
> tdat$EUR.USD.volatility=subset_expl$EUR.USD.volatility * 100
> tdat$SP500.volatility=subset_expl$SP500.volatility / 10
> tdat$SP500.Close=subset_expl$SP500.Close / 1000
> tdat$SPTSX.volatility=subset_expl$SPTSX.volatility / 100
> tdat$SPTSX.Close=subset_expl$SPTSX.Close / 10000
> tdat$SPTSX.Volume.in.Millions=subset_expl$SPTSX.Volume.in.Millions / 100
> tdat$USD.CAD.volatility=subset_expl$USD.CAD.volatility * 100
> tdat$wti.crude.oil.spot.close=subset_expl$wti.crude.oil.spot.close / 100
> tdat$YUAN.USD.Close=subset_expl$YUAN.USD.Close / 10
> tdat$YUAN.USD.volatility=subset_expl$YUAN.USD.volatility * 100
```

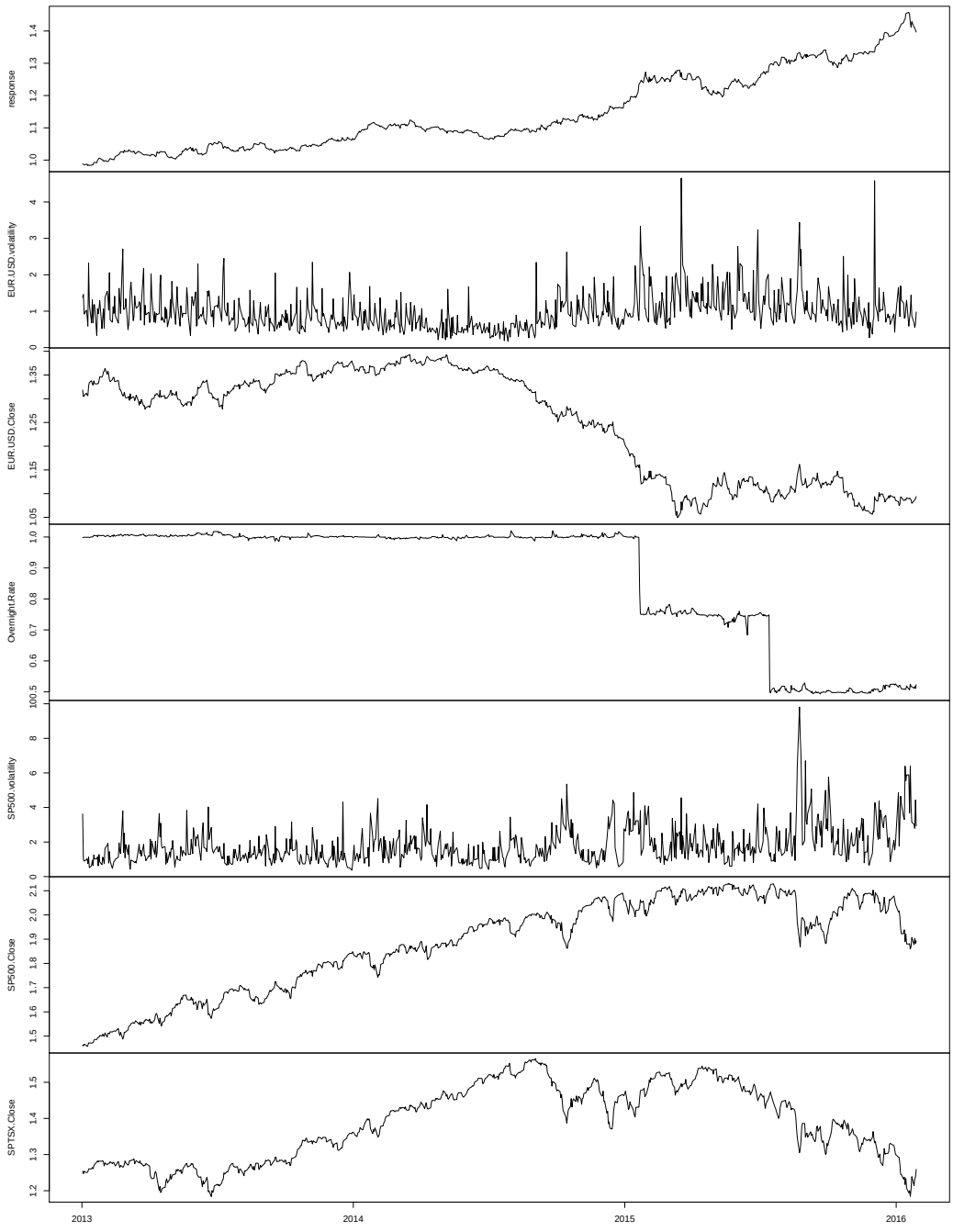
As a result, new summary statistics:

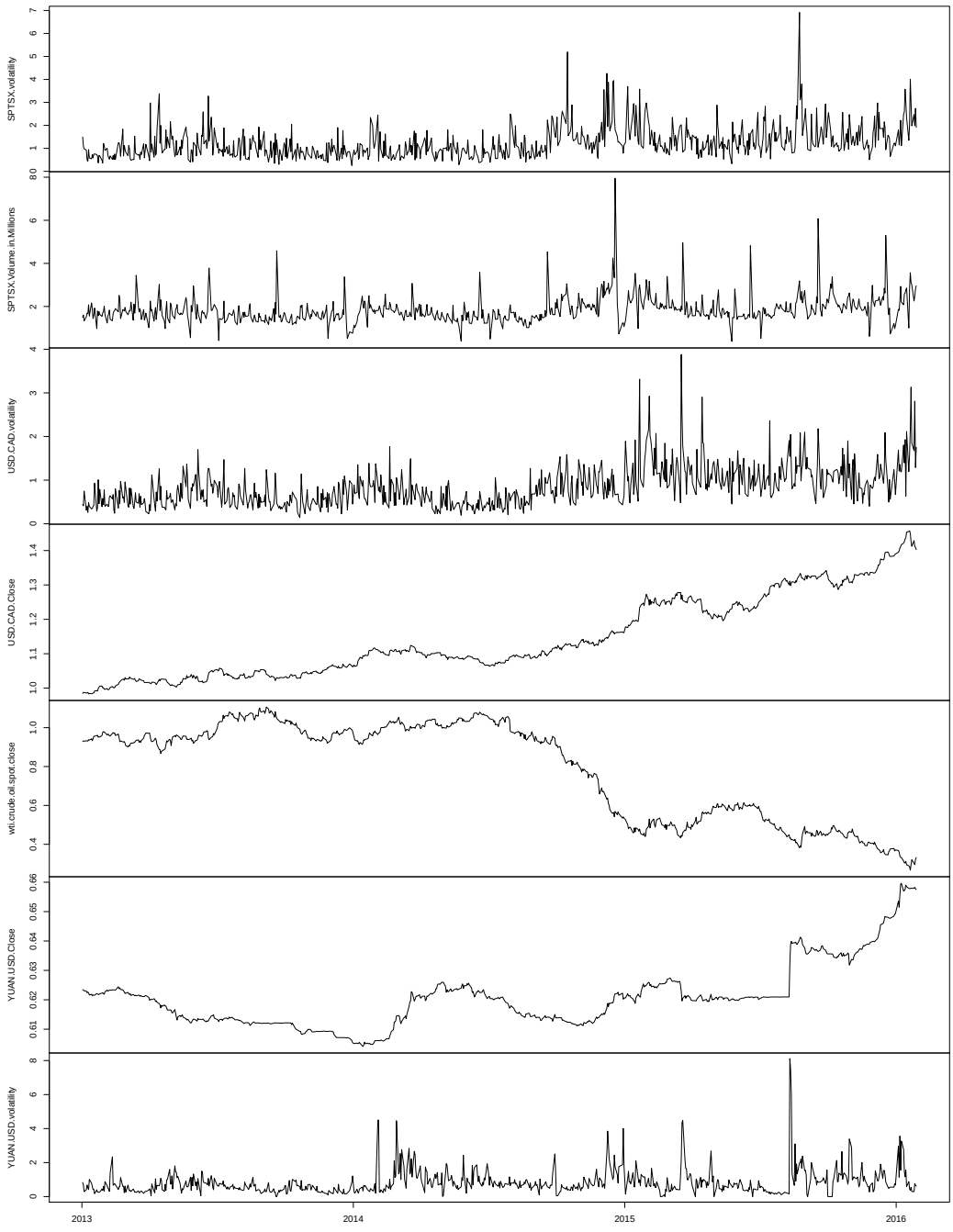
```
> summary(tdat)

        Date        response     EUR.USD.volatility EUR.USD.Close
 2013-01-02:  1   Min.   :0.983   Min.   :0.170      Min.   :1.049
 2013-01-03:  1   1st Qu.:1.046   1st Qu.:0.650      1st Qu.:1.125
 2013-01-04:  1   Median :1.103   Median :0.892      Median :1.300
 2013-01-07:  1   Mean   :1.145   Mean   :1.003      Mean   :1.251
 2013-01-08:  1   3rd Qu.:1.248   3rd Qu.:1.230      3rd Qu.:1.352
 2013-01-09:  1   Max.   :1.458   Max.   :4.652      Max.   :1.393
 (Other)   :763
 Overnight.Rate   SP500.volatility  SP500.Close     SPTSX.Close
 Min.   :0.4932   Min.   :0.370    Min.   :1.457   Min.   :1.184
 1st Qu.:0.7487   1st Qu.:1.048    1st Qu.:1.705   1st Qu.:1.282
 Median :0.9987   Median :1.532    Median :1.924   Median :1.388
 Mean   :0.8747   Mean   :1.825    Mean   :1.879   Mean   :1.387
 3rd Qu.:1.0016   3rd Qu.:2.253    3rd Qu.:2.052   3rd Qu.:1.477
 Max.   :1.0202   Max.   :9.814    Max.   :2.131   Max.   :1.566

 SPTSX.volatility SPTSX.Volume.in.Millions USD.CAD.volatility USD.CAD.Close
 Min.   :0.2522   Min.   :0.3786           Min.   :0.1400     Min.   :0.983
 1st Qu.:0.7638   1st Qu.:1.5168           1st Qu.:0.4970     1st Qu.:1.046
 Median :1.1150   Median :1.7500           Median :0.7520     Median :1.103
```

```
Mean   :1.2602   Mean   :1.8514          Mean   :0.8457     Mean   :1.145
3rd Qu.:1.5368   3rd Qu.:2.0720          3rd Qu.:1.0740     3rd Qu.:1.247
Max.   :6.9159   Max.   :7.9534          Max.   :3.8800     Max.   :1.458

wti.crude.oil.spot.close YUAN.USD.Close   YUAN.USD.volatility
Min.   :0.2668           Min.   :0.6041   Min.   :0.0000
1st Qu.:0.5141           1st Qu.:0.6131   1st Qu.:0.4000
Median :0.9312           Median :0.6203   Median :0.6500
Mean   :0.7878           Mean   :0.6207   Mean   :0.8319
3rd Qu.:0.9960           3rd Qu.:0.6233   3rd Qu.:1.0100
Max.   :1.1062           Max.   :0.6596   Max.   :8.1200
```

0.2.5 Line Charts

To gain intuition on the trend of the explanatory variables against date, we use line charts.


response
EUR.USD.volatility
EUR.USD.Close
Overnight.Rate
SP500.volatility
SP500.Close
SPTSX.Close
2013
2014
2015
2016



0.3 Variable Selection

We perform an exhaustive search to find the best model, based on linear order.

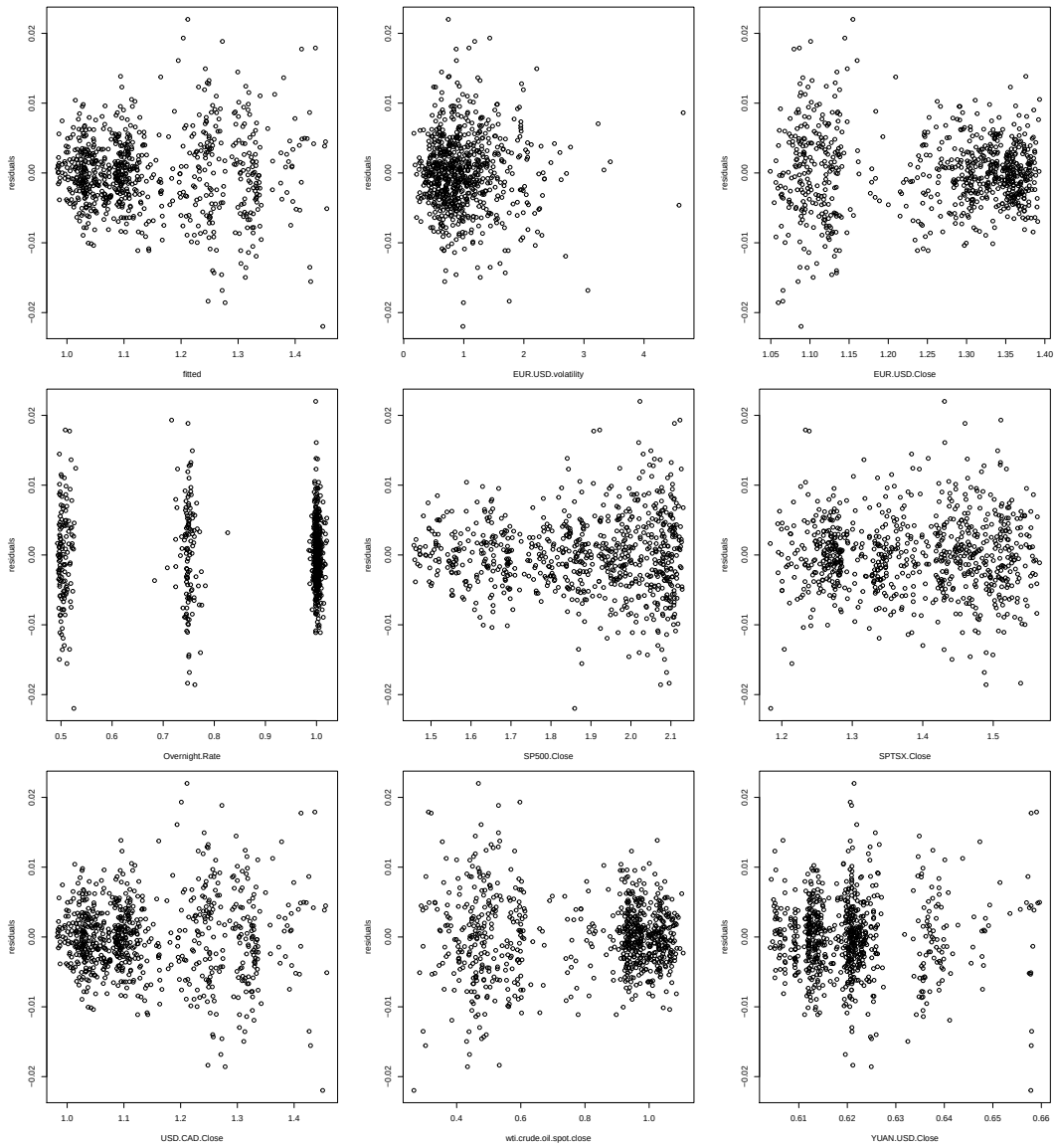
```
> # full model adjR2
> summary(lm(tdat$response~.,data=tdat))$adj.r.squared

[1] 0.9978366

> library(leaps)
> s1<- regsubsets(tdat$response~., data=tdat, method="exhaustive")
> ss1 <- summary(s1)
> # adrR2 and cp of best model based on adjR2
> ss1$adjr2[which.max(ss1$adjr2)]
```

```
[1] 0.9978426

> ss1$cp[which.max(ss1$adjr2)]

[1] 6.90572

> # here we extract the model which gave the highest adjusted R^2
> # then in the same line we draw the indices of the explanatory variables of this model
> # then we draw a vector of the names from the indices for easy human interpretation
> modeladj=names(tdat)[c(which(ss1$which[which.max(ss1$adjr2),] %in% TRUE))]
> modeladj

[1] "response"                 "EUR.USD.volatility"
[3] "EUR.USD.Close"            "Overnight.Rate"
[5] "SP500.Close"              "SPTSX.Close"
[7] "USD.CAD.Close"            "wti.crude.oil.spot.close"
[9] "YUAN.USD.Close"

> # adrR2 and cp of best model based on cp
> ss1$cp[which.min(ss1$cp)]

[1] 6.90572

> ss1$adjr2[which.min(ss1$cp)]

[1] 0.9978426

> modelcp=names(tdat)[c(which(ss1$which[which.min(ss1$cp),] %in% TRUE))]
> modelcp

[1] "response"                 "EUR.USD.volatility"
[3] "EUR.USD.Close"            "Overnight.Rate"
[5] "SP500.Close"              "SPTSX.Close"
[7] "USD.CAD.Close"            "wti.crude.oil.spot.close"
[9] "YUAN.USD.Close"

> # adjusted R^2 of full model
> summary(lm(tdat$response~.,data=tdat))$adj.r.squared

[1] 0.9978366

> # adjusted R^2 with a quadratic term
> summary(lm(tdat$response~.+tdat$USD.CAD.Close^2,data=tdat))$adj.r.squared

[1] 0.9978366
```

We find that based on C_p and adjusted R^2, both agree on a common best model. Also we did not find adding quadratic terms to the models could increase the adjusted R^2.

0.3.1 Residual Analysis

We plot residuals against fitted, as well as residuals against all explanatory variables.



We can see an issue with the homoscedasticity on the S&P 500 Close and residuals, but even with log, square root and log of log transforms we are unable to resolve the issue.

0.3.2 Prediction Intervals

We perform a test analagous to leave one out cross validation. We various training sizes (100,200,300,400,500,600) in order fit a model that is used to test and perform a prediction interval on the first day following the last day of the training set. We plot lines representing the actual price of the USD/CAD exchange (red), our prediction for that day (blue) and a 95% confidence interval for the prediction (black). As well we have a MSE, similar to the CVRMSE of cross validation, calculated as the square root of the sum of the errors divided by the number of predictions. Here, the errors are defined as the difference between the prediction and the price of the actual USD/CAD exchange on the date.

In summary, our prediction system performs *one regression for each prediction*. We experimented with various training sizes. The following plots show the predictions/CVRMSE the same time period (2015-11-10 to 2016-01-25) for various training sizes. The coefficients of regression are unique for each prediction, and the function *looper* (see appendix for source code), keeps track of the coefficients. We toyed with the idea of analyzing these regression coefficients as a time series, and found that they are of an autoregressive nature.

```
[1] "EUR.USD.volatility"       "EUR.USD.Close"
[3] "Overnight.Rate"           "SP500.Close"
[5] "SPTSX.Close"              "USD.CAD.Close"
[7] "wti.crude.oil.spot.close" "YUAN.USD.Close"
```

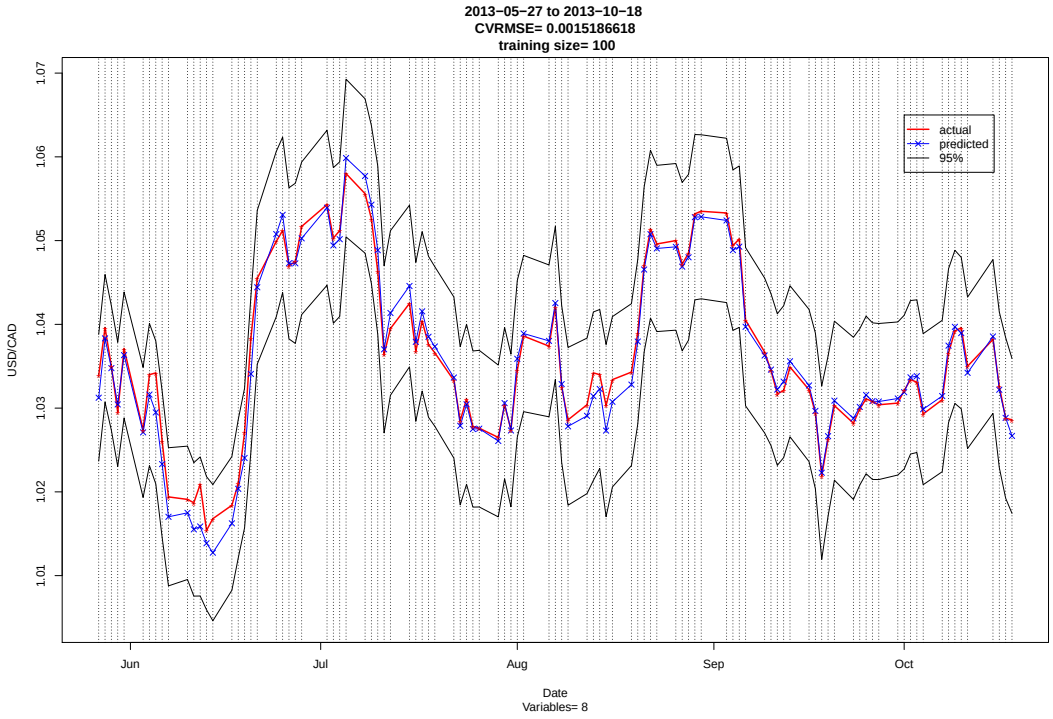


Figure 4: Plot of prediction vs time on top of actual vs time of exhaustive search best model with training size as 100

```
 [1] "EUR.USD.volatility"        "EUR.USD.Close"
 [3] "Overnight.Rate"            "SP500.volatility"
 [5] "SP500.Close"               "SPTSX.Close"
 [7] "SPTSX.volatility"          "SPTSX.Volume.in.Millions"
 [9] "USD.CAD.volatility"        "USD.CAD.Close"
[11] "wti.crude.oil.spot.close" "YUAN.USD.Close"
[13] "YUAN.USD.volatility"
```

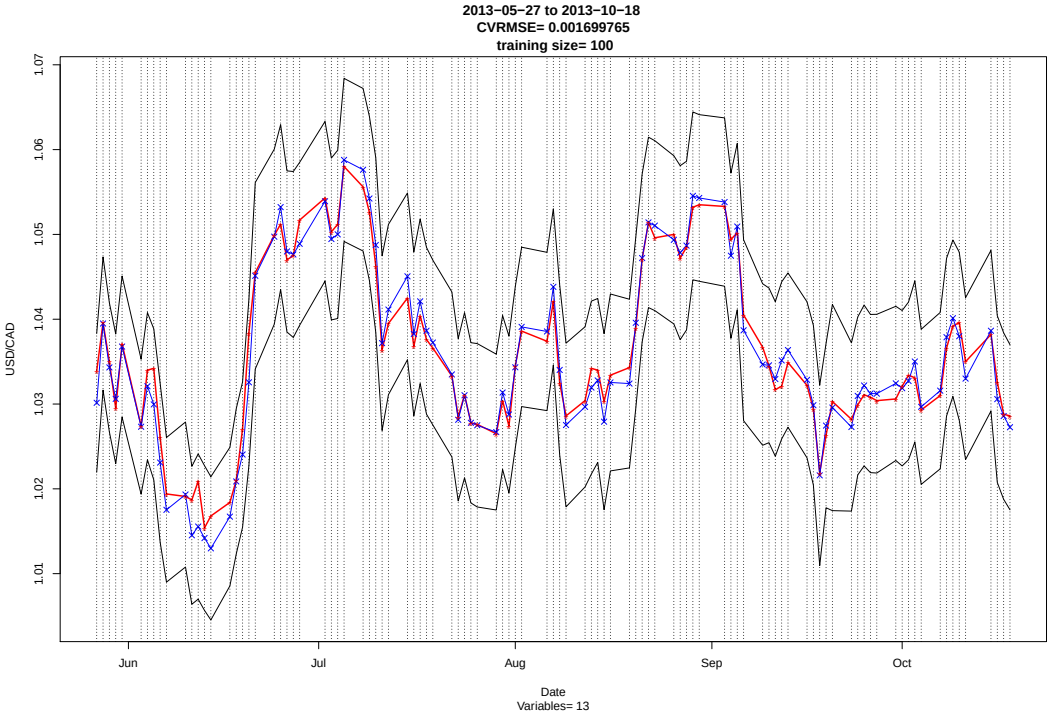


Figure 5: Plot of prediction vs time on top of actual vs time of full model with training size as 100

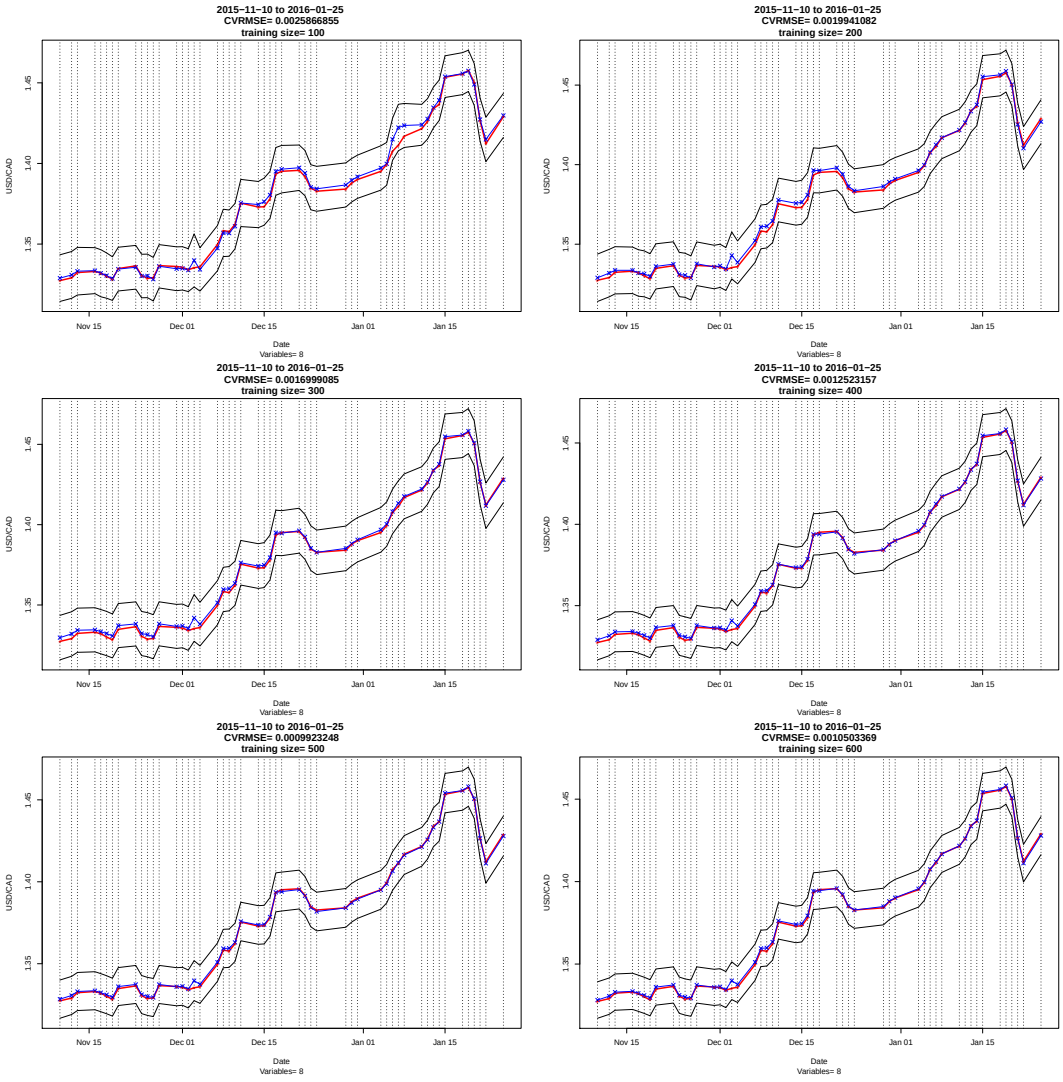


Figure 6: Plot of prediction vs time on top of actual vs time of for various training sizes cross validation exhaustive search best model, blue line: prediction, red line: actual prices, black lines: future prediction interval

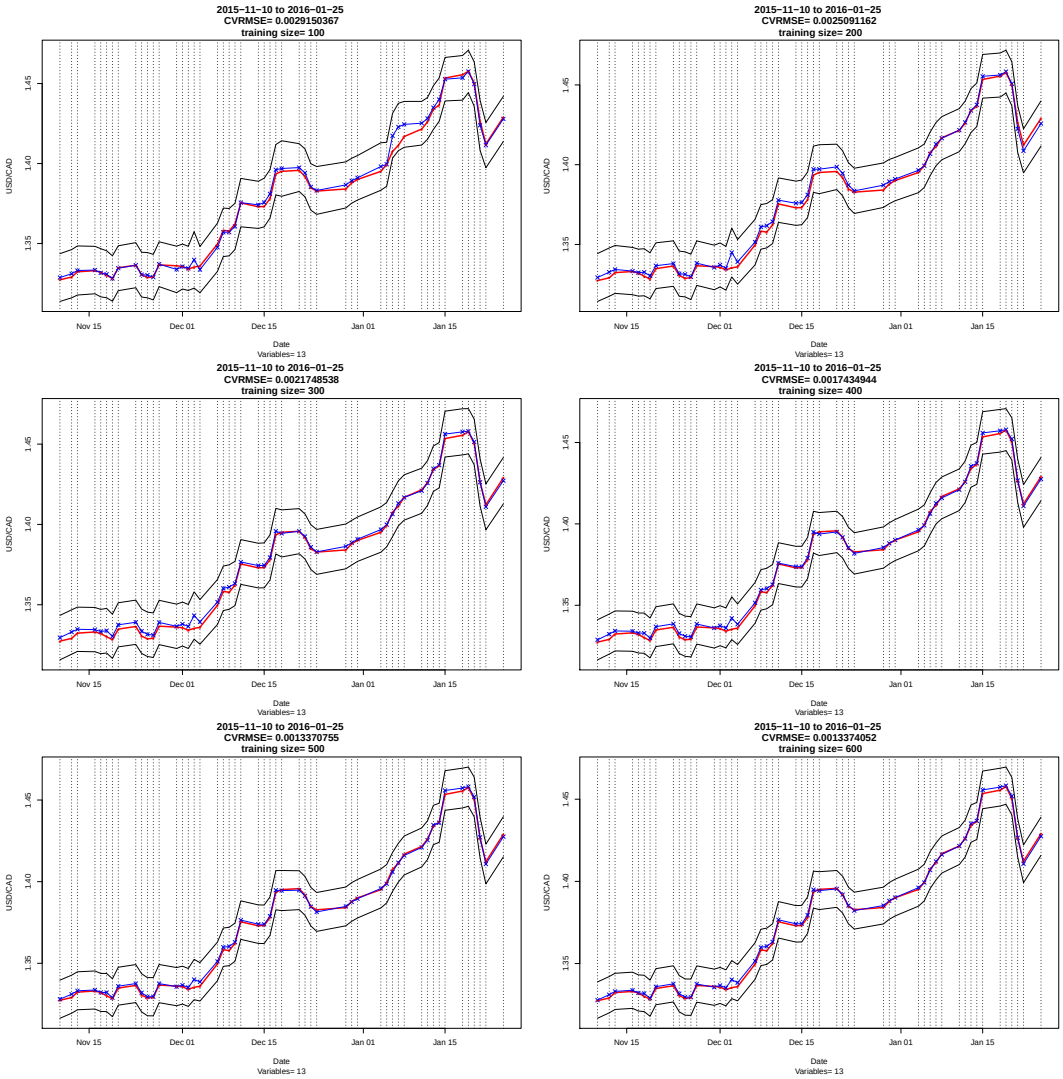


Figure 7: Plot of prediction vs time on top of actual vs time of for various training sizes cross validation for full model, blue line: prediction, red line: actual prices, black lines: future prediction interval

0.4 Summary

We performed a form k-fold cross-validation on the full model with all explanatory variables and on the model chosen by an exhaustive search. We found the model with the exhaustive search had the best adjusted R^2, the lowest C_p and the lower CVRMSE. Therefore this is the best model, with 8 linear explanatory variables.

```
> modeladj

[1] "response"                  "EUR.USD.volatility"
[3] "EUR.USD.Close"             "Overnight.Rate"
[5] "SP500.Close"               "SPTSX.Close"
[7] "USD.CAD.Close"             "wti.crude.oil.spot.close"
[9] "YUAN.USD.Close"
```