Project 2

Data Analytics in R

MSc Data Analytics

Contents

List of Figures ii

List of Tables iii

1 Exploratory Data Analysis 1

2 Model Selection 3

3 Prediction 6

List of Figures

1.1 SMT Correlation 1
1.2 EXPN and SPX correlation 1
1.3 SSE and STJ Correlation 1
1.4 Correlation plot with VOD 2
2.1 Leaps and Bounds models 3
2.2 Model summary plot 5
2.3 Box Cox plot 5

List of Tables

2.1 Model comparisons 4

3.1 Expected VOD vs Predicted VOD values 6

Chapter 1

Exploratory Data Analysis

Beginning with the exploratory analysis, we start by getting the correlations in the dataset. There are groups of variables highly correlated with each other, SMT has the highest multicollinearity as most of the variables are correlated to it with a value greater than 0.87 (see Fig 1.1). Now considering 0.87 as the correlation point for high correlation, removing SMT to avoid complications in the dataset. Also, check individual variables correlated and removing variable SPX as it is highly correlated to EXPN (see Fig 1.2) and doing the same with SVT due to its high correlation with SSE (see Fig 1.3) as both these variables are comparatively less correlated with VOD than their highest correlated independent variables. Thus, cleaning the dataset will help avoid further complications in the linear model.

	VOD	SVT	STJ	SMT
VOD	1.000000000	0.59486286	0.08768426	-0.20520928
SVT	0.594862855	1.00000000	-0.43664085	0.25253502
STJ	0.087684263	-0.43664085	1.00000000	-0.50976440
SMT	-0.205209283	0.25253502	-0.50976440	1.00000000
MGGT	-0.343893582	0.15019040	-0.33209840	0.65086988
TSCO	-0.566789297	-0.47928635	-0.09856268	0.29238247
EXPN	-0.363714222	0.29145649	-0.55067735	0.87702927
SPX	-0.382599791	0.11684990	-0.45862311	0.94583810
AUTO	-0.168233775	0.49376488	-0.55327612	0.80137276
SSE	0.639246796	0.91148318	-0.43957040	0.25364058
AHT	-0.399493814	-0.02023152	-0.36353196	0.87481775
RTO	-0.460110253	-0.16978907	-0.23316971	0.74169765

Figure 1.1: SMT Correlation

	MGGT	TSCO	EXPN
VOD	-0.34389358	-0.56678930	-0.36371422
SVT	0.15019040	-0.47928635	0.29145649
STJ	-0.33209840	-0.09856268	-0.55067735
SMT	0.65086988	0.29238247	0.87702927
MGGT	1.00000000	0.25149136	0.71881390
TSCO	0.25149136	1.00000000	0.34944366
EXPN	0.71881390	0.34944366	1.00000000
SPX	0.77025302	0.33306945	0.89493729
AUTO	0.58164012	0.13705647	0.80795046
SSE	0.03691237	-0.32214524	0.22662072

Figure 1.2: EXPN and SPX correlation

	AUTO	SSE
VOD	-0.1682338	0.639246796
SVT	0.4937649	0.911483181
STJ	-0.5532761	-0.439570397

Figure 1.3: SSE and STJ Correlation

In order to build a model to fit the dataset, we will check the plots to see if any transformations are required. Firstly, creating a new data frame removing the date, month, year and weekday column to check the relation of the response variable VOD closing price with other companies closing prices in plot, since the dataset has many columns splitting the dataset to view the plots. Using Turkey's Ladder transformation as reference, as seen in Fig 1.4 variables SMIN, BA and PRU form a backward C shape pattern with respect to the response variable and thus taking the square of these variables to increase the value of x.

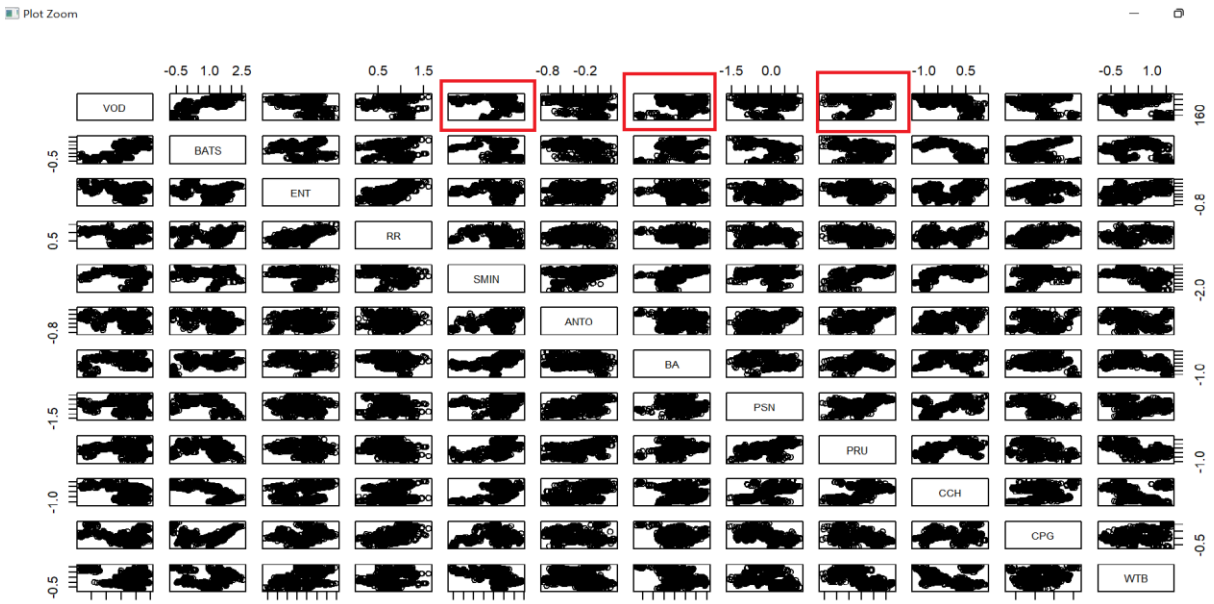


Figure 1.4: Correlation plot with VOD

The same pattern can be seen for variables SDR and STJ (Appendix Fig 1.1, Appendix Fig 1.2) and adding the squared values of these columns in the dataset.

Chapter 2

Model Selection

Once the transformations are applied, we will now try to build a model by adding enough variables that would explain the trend in the data while not adding many variables so as not to overfit. Starting with leaps and bounds to fit every possible model and pick the best models. The main method used in leaps and bounds is building set of models of different sizes and calculating using selection criteria Cp or adjusted R squared. We are applying adjusted R squared technique to select top 5 best models of leaps and bound. As seen in Fig 2.1 few models are obtained with respect to size, looking at the curve here we would probably pick the best model of sizes 8, 9, and 10 which is roughly where the plot looks like it's reaching an upper limit.

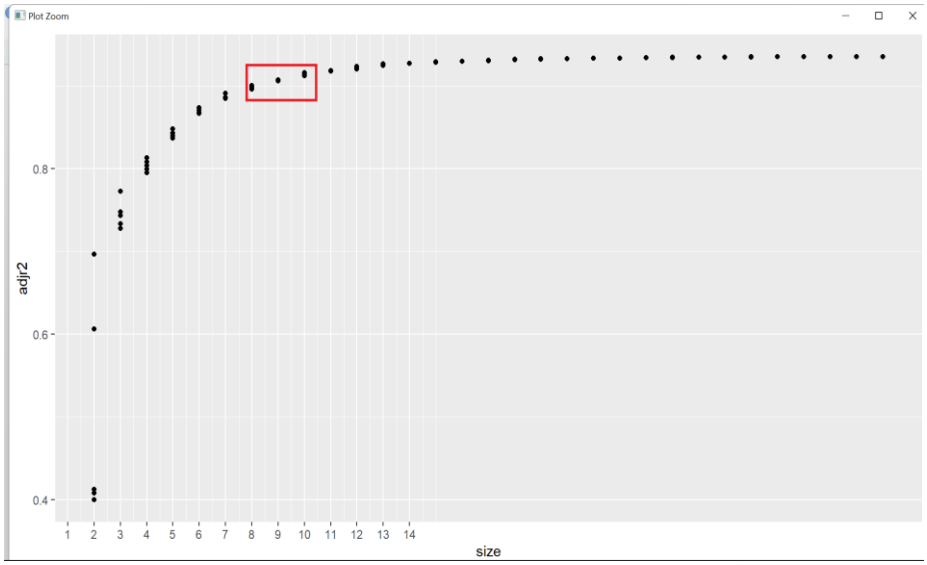


Figure 2.1: Leaps and Bounds models

Now, we have models leaps8, leaps9 and leaps10 with us. Going ahead fitting the full model to perform automatic variable selection procedures: forward selection, backward selection and stepwise selection to further build few more models to choose from. Forward selection starts with nothing in the model and gradually add variables until best model is obtained, backward selection does the opposite starting with a full model with all variables, removing each variable to only keep important ones whereas stepwise selection iterates between both forward and backward until a best model is obtained. These methods mainly use AIC or hypothesis testing to add or remove variables from the model.

We now have all our models, leaps8, leaps9, leaps10, forwards, backwards and stepwise we need to compare the models and chose the best model to fit our data. To do that we will start with calculating Mallow's Cp commonly used for variable selection to directly compare the models. Mathematically calculated as follows:

$$\frac{RSS_p}{RSS_q/(n-q-1)} + 2(p+1) - n$$

p refers to the smaller of the two models being considered (which has p independent variables), and q refers to the larger of the two models with q independent variables. The model with lowest Cp value can be considered as a better model with respect to others.

To make a better comparison we will also calculate Press value for each model which will also tell us how model will perform on unseen data.

Table 2.1: Model comparisons

Model	p	adjR2	Cp	PRESS
Leaps - Size 8	8	0.9008	282.6736	37340.31
Leaps - Size 9	9	0.9079	210.0501	34667.14
Leaps - Size 10	10	0.9165	123.2582	31550.34
Forward	25	0.9273	26.67501	28035.94
Backward	24	0.9274	25.31669	28008.12
Stepwise	23	0.9273	24.41322	27941.48

Also considering adjusted R squared for better comparison, higher the adjusted R squared value better is the model. It can be noted that not a great difference can be seen in adjusted R squared values. Comparing all the models, I would choose Stepwise model as my final model as it has the lowest Cp which tells us how model performs on the available data along with having lowest PRESS which further indicates it as a good predictor. Although the adjusted R square is not highest, it is only slightly less than the highest value so going ahead with the stepwise model.

We would now go ahead and verify of normality and assumptions by plotting the graph for our selected stepwise model (Fig 2.2). if we take a look at the normality on the QQ plot, the residuals are almost linear, although it appears to be a bit skewed at the ends, we can say that our model is not great but a good fit model. Further checking the independence of the residuals, we can see residual vs fitted plot shows no particular pattern and also nothing much going on with the outliers.

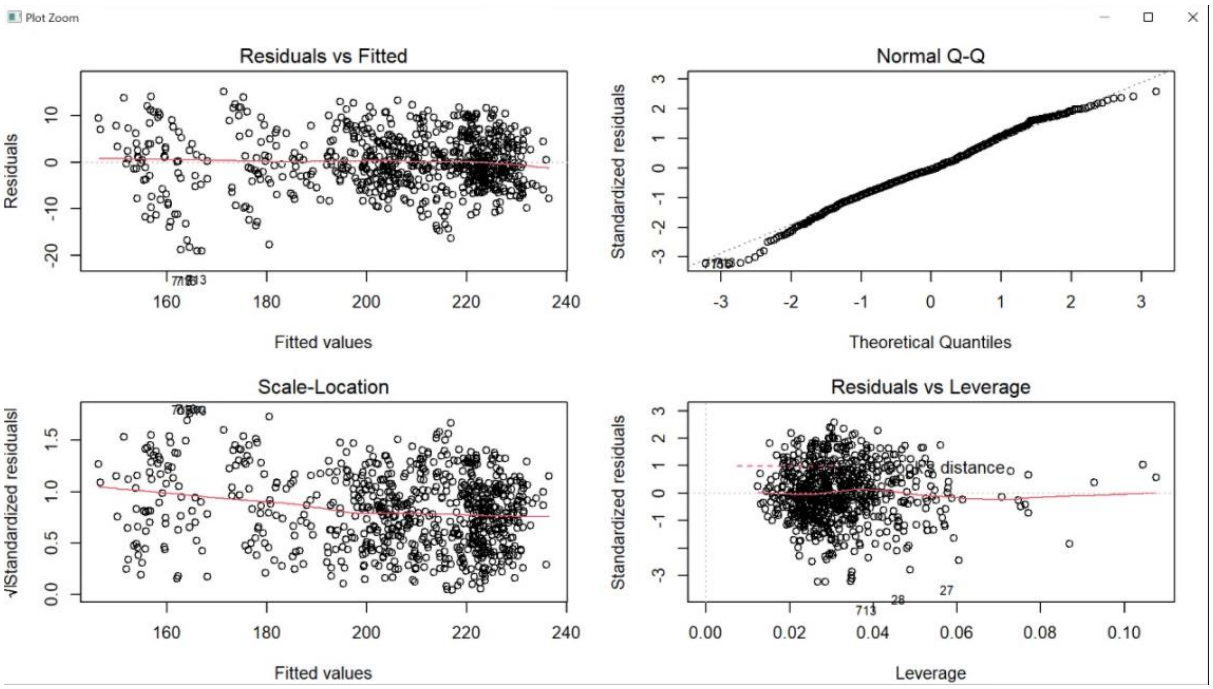


Figure 2.2: Model summary plot

To further analyse if we need to perform any transformations to make the model better, we will use Box-Cox to determine if any transformation is required. The lamda value suggested by Box-Cox (Fig 2.3) giving 95% confidence interval of fitted values is greater than 2 thereby stating no further transformation is required in our model.

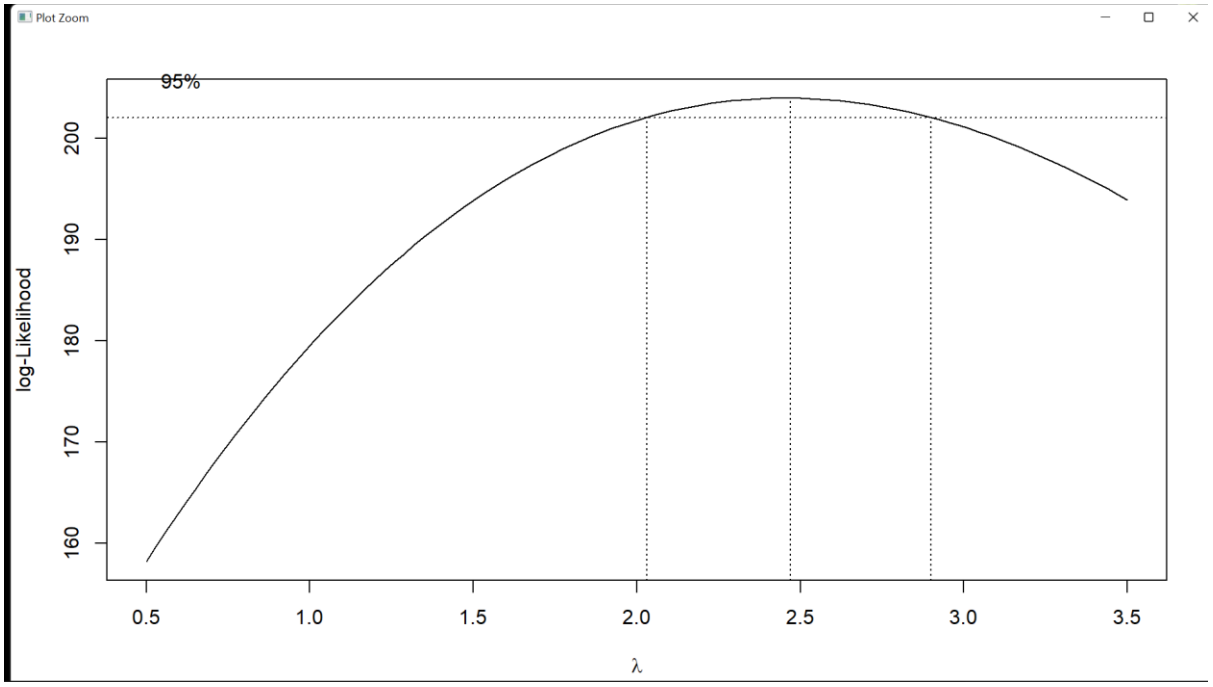


Figure 2.3: Box Cox plot

Chapter 3

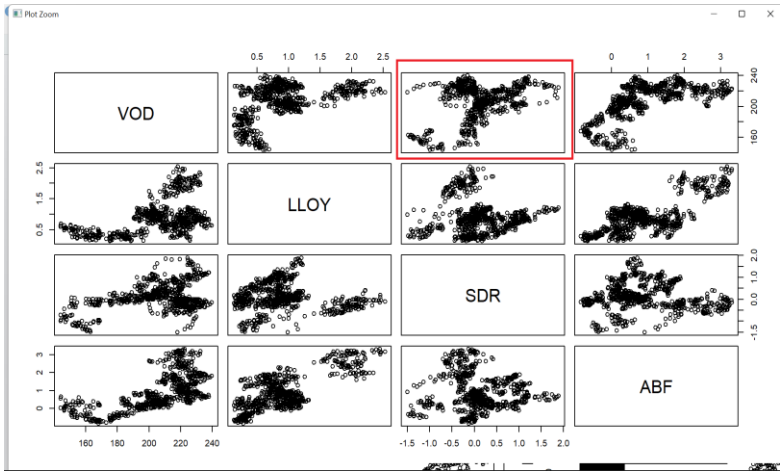
Prediction

The prediction analysis conducted for the model suggests that the prediction model is pretty close as compared to the expected values. Comparisons of initial 5 rows of expected values and predicted values are shown in Table 3.1

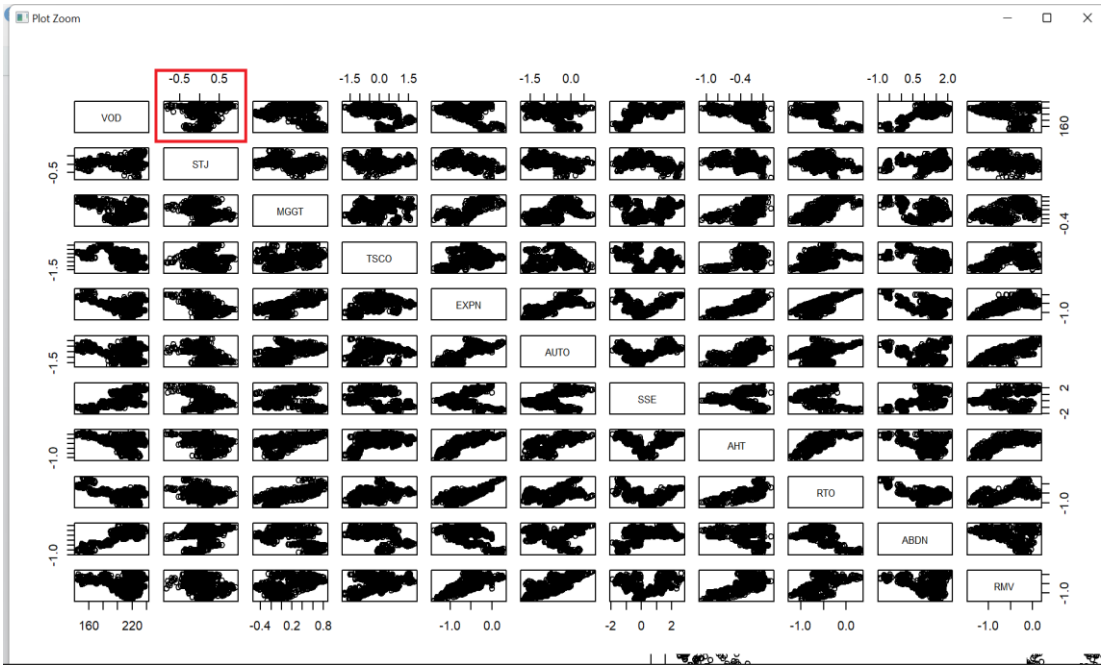
Table 3.1: Expected VOD vs Predicted VOD values

Expected Value	Predicted Value
219.70	221.12
219.15	223.20
218.20	218.15
223.05	213.41
220.80	211.16

Appendix



Appendix 1.1: Correlation with VOD 2



Appendix 1.2: Correlation with VOD 3