Modelling BAE Systems Stock Prices based on Data from the London Stock Exchange

MM916 Regression Modelling Project

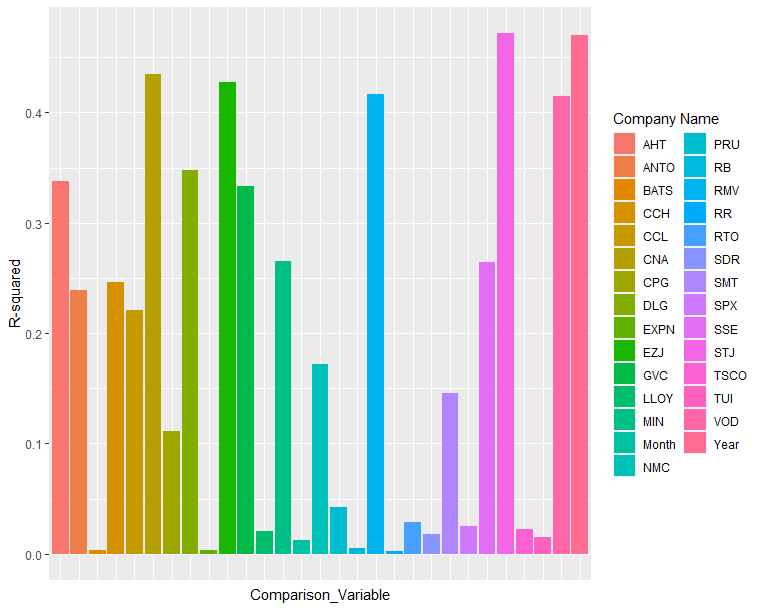
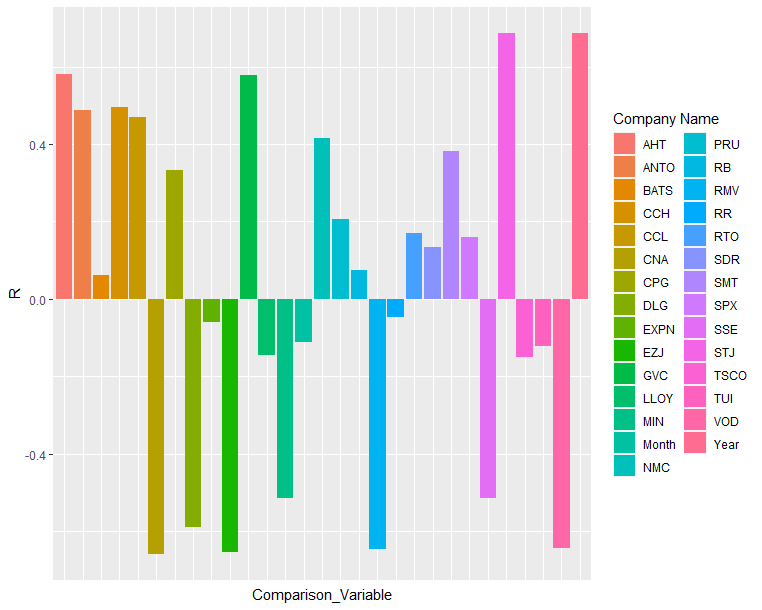
# Introduction

The goal of this project was to predict the closing share prices of BAE Systems based off of the closing prices of 29 other companies included in the Financial Times Stock Exchange 100 Index, including St James's Place plc, Vodafone and EasyJet. The dataset consists of data taken daily from January 2016 up until January 2019, with the companies other than BAE Systems having been standardised with mean 0 and variance 1 in order to prepare them for use in a regression model. As the closing price for BAE Systems is one day ahead of the other values, this allows for a regression model to be prepared that can predict closing share prices a day ahead of the final values for the other companies. In this report I will explain my reasoning behind selecting the variables used in my final models, as well as a brief discussion of how effective the model’s predictions are and how these could be improved.

# Constructing an Initial Model

## Selecting Variables

In order to construct a preliminary model, I began by removing the two non-numerical models from the data (Date and Weekday) as these could not be easily fitted to a linear regression model (although I will discuss how they potentially could be later). I then took the Pearson’s correlation coefficient for each variable when compared to BA and plotted these in the bar chart below:



As this was difficult to compare, I then took a square of the R coefficient in order to obtain the R2 values which are much easily comparable. To create a simple model that could be used as a comparison for more complex models that were later produced, I selected the 5 Companies with the strongest correlation with BA. These were in order:

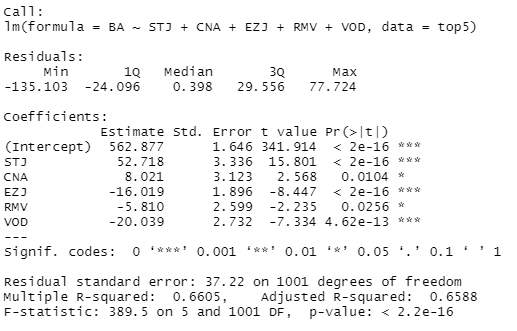
|  |  |  |  |
| --- | --- | --- | --- |
| Base\_Variable | Comparison\_Variable | R | R\_squared |
| BA | STJ (St James's Place plc) | 0.6870552 | 0.4720448 |
| BA | CAN (Centrica plc) | -0.6591334 | 0.4344569 |
| BA | EZJ (EasyJet plc) | -0.6536152 | 0.4272129 |
| BA | RMV (Rightmove plc) | -0.6454839 | 0.4166495 |
| BA | VOD (Vodafone Group plc) | -0.6440479 | 0.4147977 |

## Constructing and Evaluating the Linear Regression

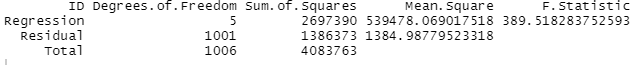
I then fitted the 5 variables to a linear regression model using the below code:

top5\_mod <- lm(BA~STJ+CNA+EZJ+RMV+VOD, data=top5)

This produced a model with the following coefficients and test statistics.



Using the values contained within the model summary it was then possible to produce an ANOVA table for the model, which would then allow me to carry out an F-test to assess the goodness of fit of the model. The ANOVA table constructed looked like so:



### Gauging the Goodness of Fit

The observed F (389.518) is large in comparison to the critical value obtained from the F-test (2.223) and the p-value was also found to be extremely small (6.804e-232) which means we can confidently reject the null hypothesis, which indicates that there is a good fit between BA’s closing prices and the other 5 companies closing prices, and that a linear relationship exists between BA's closing prices and at least one of STJ's, CNA's, EZJ's, RMV's and VOD's closing prices.

### Interpreting the Model Coefficients

The completed model has the following equation:

In order to gauge how well these coefficients stacked up and how good they would be at modelling the liner relationship, I carried out a confidence interval check for the value of BA closing prices based on values from the first column of the top5 table to see if the value predicted would pass the sanity check.

The confidence interval tells us on average what closing price BA would have based off of specified values for the other closing prices and takes into account sample variation allowing for a tighter interval to be plotted.

The check returned a value for a BA price of 453.2791 with an UL of 459.0606 and LL of 447.2791, which the actual value of 456.7 for those same values fits comfortably inside. This is a nice tight interval too, with a +/- of a little over 1% of the fitted variable, meaning that the model can predict the average BA closing price to a tight margin.

A prediction interval check was also carried out. Unlike a confidence interval, a prediction interval aims to calculate the response for a specific stock price, not for a sample based on averages. This managed to return the same BA price of 453.2791 although with a much larger interval, roughly about 15% of the fitted value. This suggests that the model is capable of predicting the value of closing prices for a given sample quite accurately. However, when trying to predict the closing price for a specific day it is far less useful.

### Potential Issues with this Model

To see if the coefficients were truly representative of the effect each variable was having on the response, I decided to compare the correlation of each variable respective to BA against the coefficients for each variable in the model:

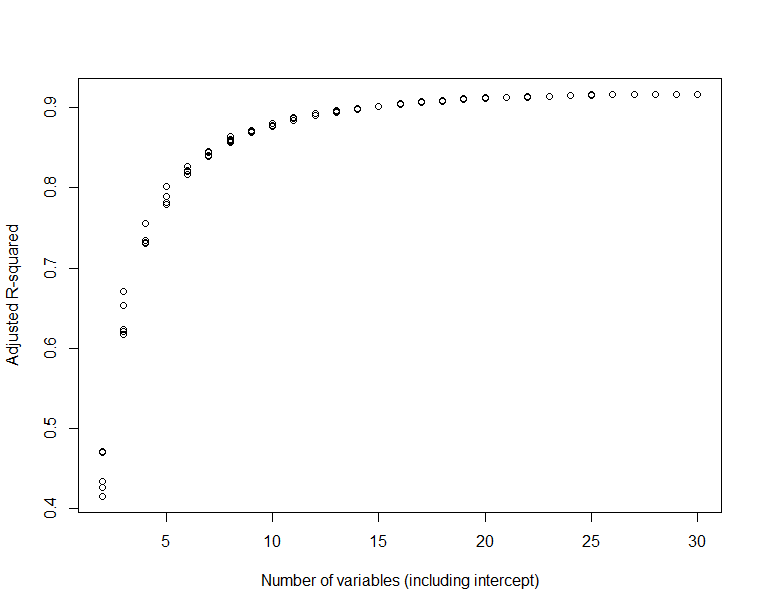
|  |  |  |
| --- | --- | --- |
| Variable | Correlation | Coefficient |
| STJ | 0.6870552 | 52.718 |
| CNA | -0.6591334 | 8.021 |
| EZJ | -0.6536152 | -16.019 |
| RMV | -0.6454839 | -5.810 |
| VOD | -0.6440479 | -20.039 |

In terms of correlations STJ, CNA, EZJ, RMV and VOD all have roughly equal effect on the closing price of BA, either in a positive or negative direction, with the magnitude between the highest and lowest correlation having a difference of only 0.043. However when we then compare this to the coefficients, three of the stocks have a much higher effect on the calculated BA closing price, with STJ having a whopping coefficient of 52.718, which is over twice the magnitude of the second highest coefficient, -20.039 for VOD closing prices. This is also a strange comparison, as VOD had the lowest correlation with BA yet it has the second largest effect on the calculated BA price. These factors, plus the strangely diminished effect the second most correlated value (CNA) has on the BA equation suggest that there are other factors or variables affecting the data that are not present in the model in its current form.

# Fitting a Better Model

## Selecting the Variables

In order to select the variable for a new and improved model, I decided to employ 4 different methods to select the variable. As the number of considered variables is less than 31, I began by running a leaps and bound selection to find the 5 best models for each amount of variables, from 1 to 30. I then plotted the adjusted R2 for each model size against the number of variables included to determine at what point adding more variables would lead to diminishing returns on model fit.



As the graph seemed to level off at roughly 20 variables, I decided to extract the models with the best R2 score for 19, 20 and 21 variables to compare their performance later.

In addition to using leaps and bounds, I also carried out forward and backward checks for creating a model, as well as a stepswise check to create a model automatically by combing forwards and backwards checks until it reaches the model with the ‘best’ fit. In order to compare these models, I calculated the adjusted R2, the Mallow’s Cp and the PRESS statistic for each one, which would allow me to decide what model to continue working with:

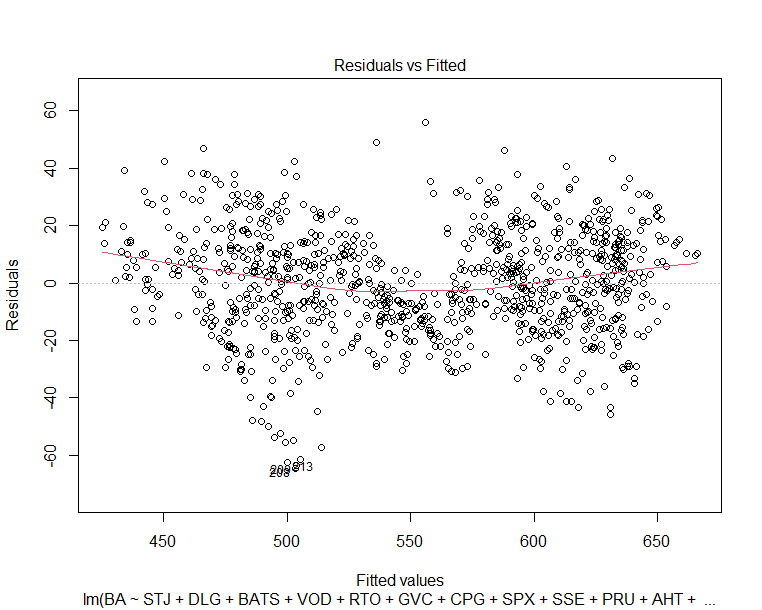
|  |  |  |  |
| --- | --- | --- | --- |
| Name | Adjusted R2 | Cp | PRESS |
| Size 19 Model | 0.9105454 | 85.49485 | 373315.1 |
| Size 20 Model | 0.9117497 | 72.24615 | 368609.8 |
| Size 21 Model | 0.9125995 | 63.19567 | 365523.7 |
| Stepswise Model | 0.9159541 | 29.71423 | 353632.8 |
| Backwards Model | 0.9158588 | 29.83258 | 353659.6 |
| Forwards Model | 0.9159541 | 29.71423 | 353632.8 |

Ultimately, I decided to continue with the Stepswise Model as it has a higher adjusted R2 than most of the other models, as well as joint lowest Cp and PRESS statistics with the Forwards Model. This is because the Forwards and Stepswise methods have resulted in the same final model, which is shown below:

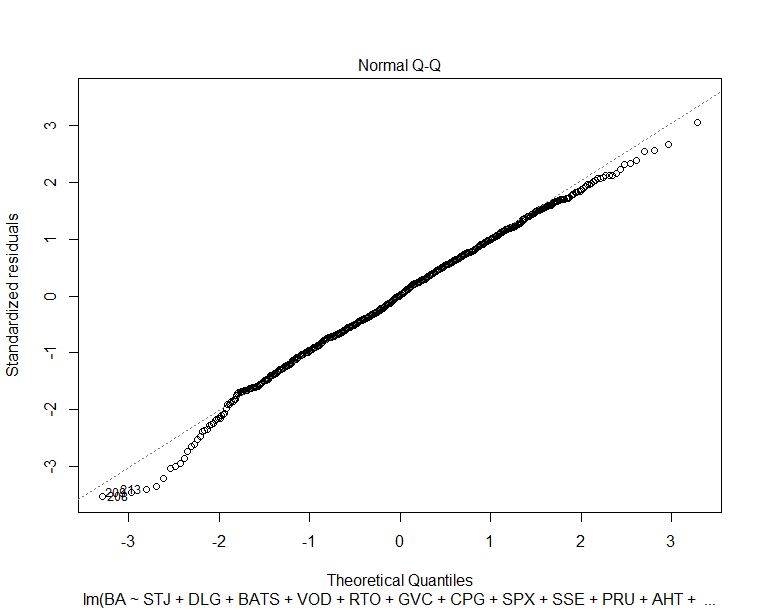
Model <- lm(BA ~ STJ + DLG + BATS + VOD + RTO + GVC + CPG + SPX + SSE + PRU + AHT + RB + MIN + RMV + Year + EXPN + TSCO + SDR + CCH + NMC + SMT + ANTO + TUI + LLOY + CNA + Month, data = lse)

## Transformations

In order to determine if any transformations of the model would be necessary for it to meet all the regressions assumptions I began by plotting all of the residuals for each coefficient against the fitted values to see if they had a roughly even spread of residuals, as this would indicate whether the error variance was constant or not. While some of the plots for individual variables were not evenly spread, a plot of residuals against fitted values for the overall model suggests that these uneven spreads seem to cancel out when combined:



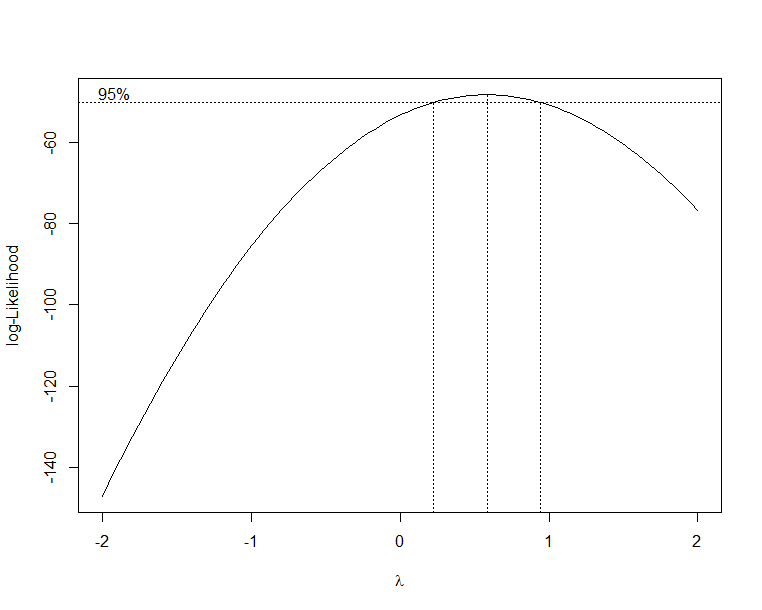
To test whether the errors were normally distributed I also plotted a Q-Q plot of standardised residuals against the theoretical quantities:

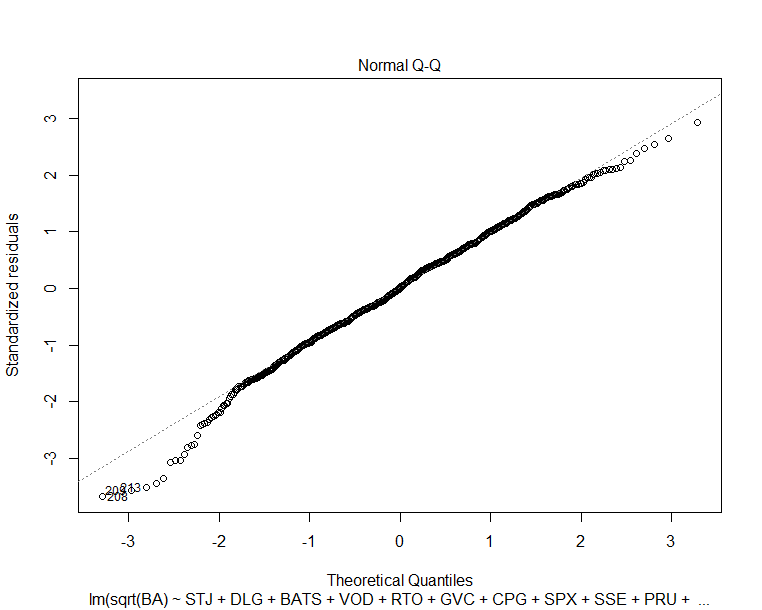


The QQ plot suggests that the errors are either normally distributed with a few outliers, or that there is a slight left skew present. In order to check this a transformation of the data could be carried out and a new Q-Q plot produced afterwards to see if the issue has been resolved.

As some of the individual charts had some possibly non-random patterns visible it is possible that the errors aren’t completely independent. Hopefully this can be resolved in the same transformation needed to normalise the data.

Due to the difficulty in trying to determine trends in the plots of BA against each variable in order to use Tukey’s Ladder of Transformation, I decided to begin by using the boxcox command to find the single most likely transformation to better fit the model.



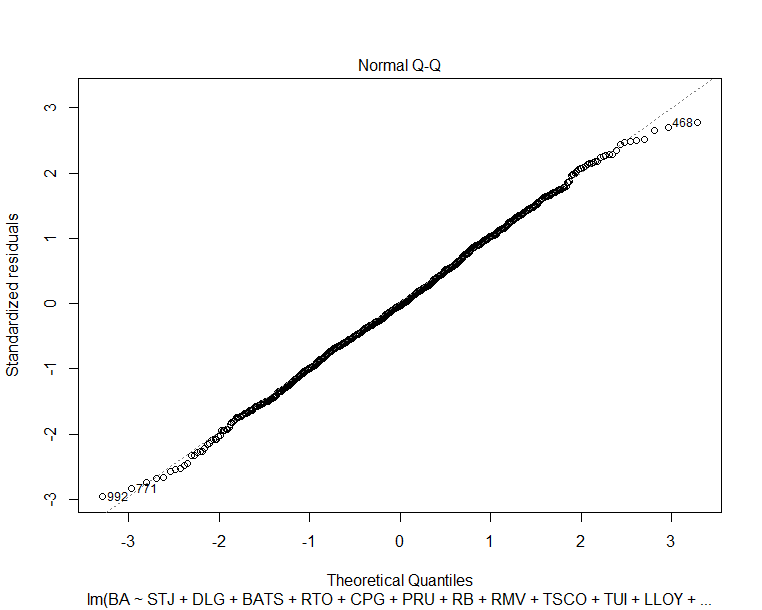
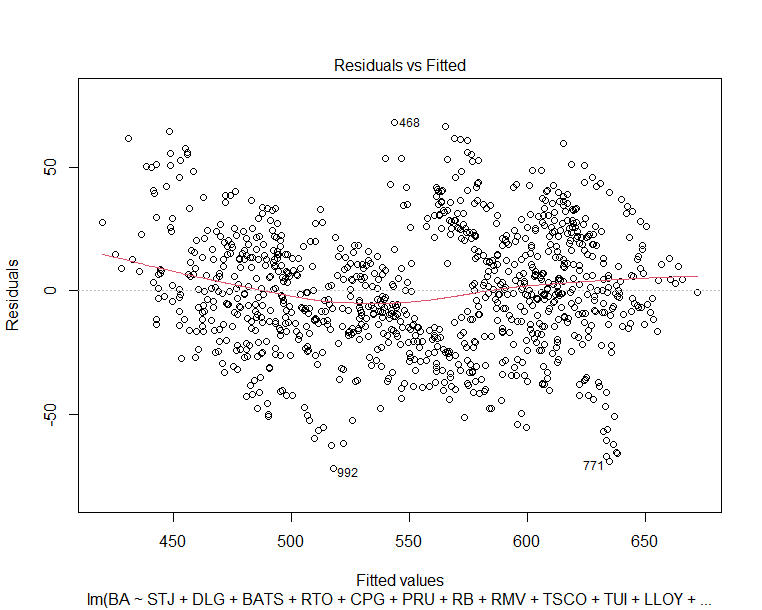
As 0.5 falls near the centre of the confidence interval this suggests a square root of the response will cause the data to fit better. After transforming the response (square root of BA) I replotted the QQ-plot to see if there was any difference before and after the transformation:  


As this transformation seemed to have little to no effect on the QQ-plot tails I decided to check for any other factors that might be interfering with the regression assumptions.

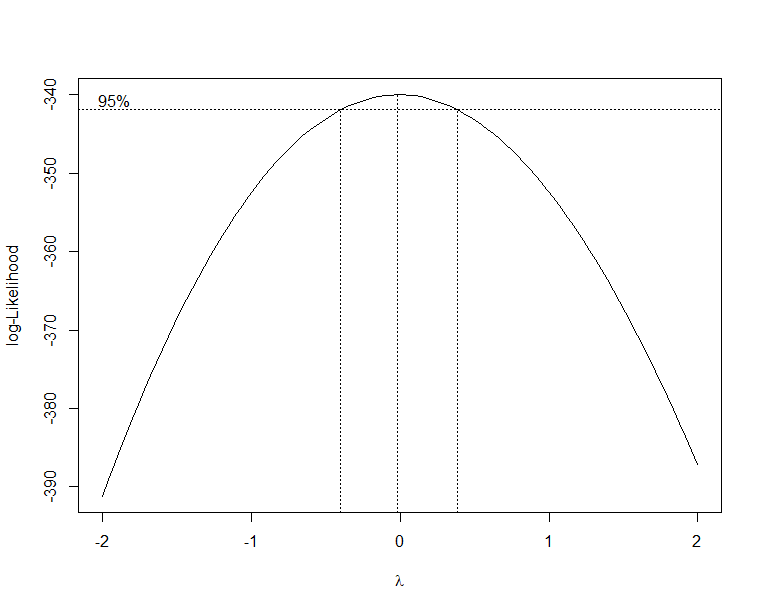
## Multicollinearity

As there are a lot of variables still contained in the model I decided to check if any of them might be interfering with one another and therefore skewing any model outputs. To determine what variables might show significant correlation against one another I created a correlation matrix for every variable in the current model and flagging any that had 80% or higher correlation with another independent variable.

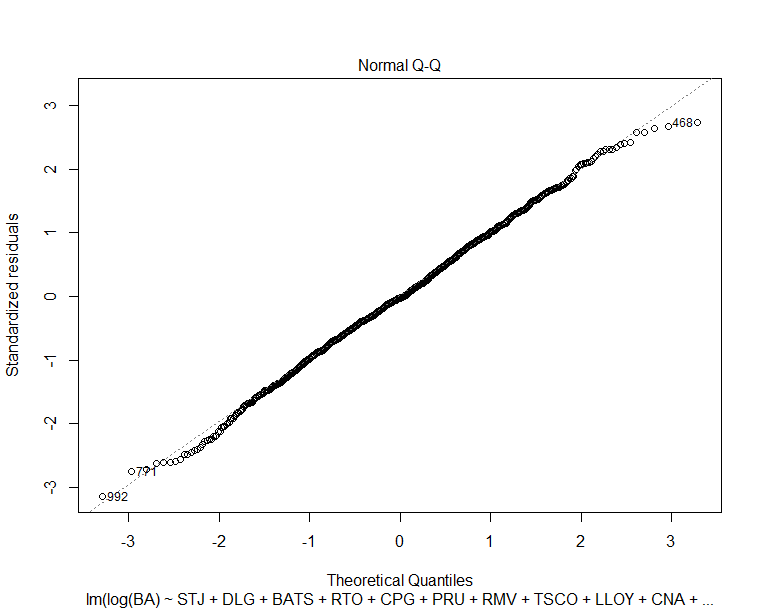
Unsurprisingly several variables had a correlation of over 80% with several even having correlation higher than 90%. The variables that showed high correlation are; VOD, GVC, SSE, AHT, MIN, CCH, NMC, SMT, EXPN, SSX and SDR. While I initially tried to just remove the variables with the lowest correlation with BA, this had a minimal effect on the model and so I opted to remove all the variables and then replot the Residuals against Fitted and Q-Q plots to see if there was any change:



As you can see, even before taking an additional transformation into account the Q-Q plot has produced an almost perfectly straight line, highlighting the dramatic effect the multicollinear variables had on the model. To see if the model could be improved further still, I created another boxcox chart:



The chart has shifted now to suggest that a log transformation of the response would produce a better model.



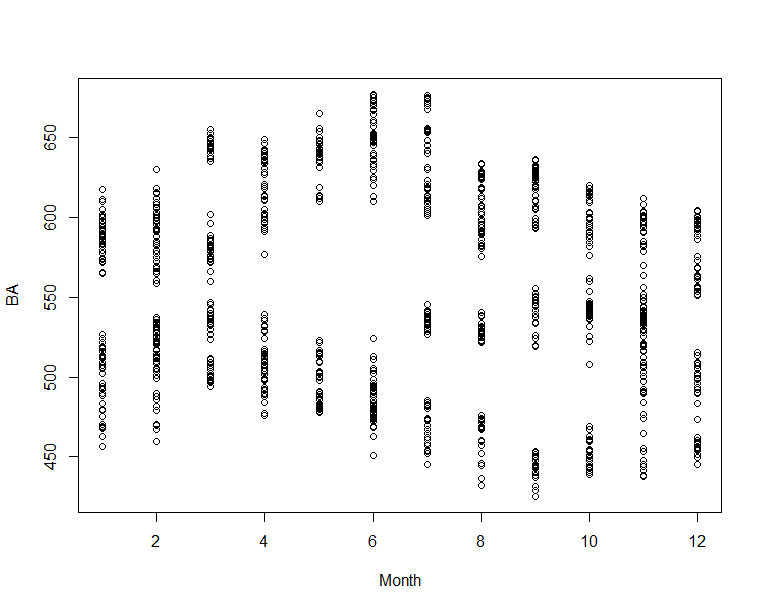
Producing the log model also changed the significance value of several of the variables to decrease, allowing for a ‘final’ simplified model:

log.mod <- lm(log(BA) ~ STJ + DLG + BATS + RTO + CPG + PRU + RMV + TSCO + LLOY + CNA + Month, data = lse)

# Predictions and Improvements

While I initially attempted to predict the closing prices using the model which still contained the multicollinearity variables, this was only able to achieve a prediction with a root mean square prediction error (RMSPE) of 67.14.

By removing the multicollinear models and taking the log transformation of the response, the ‘final’ model was able to lower this value down to 56.49. I was able to improve this further still by taking plots of BA vs the independent variables to see if any of them would lend themselves to Tukey’s Ladder of Transformation. Two values produced plots that suggested a transformation would be possible (DLG and CPG) while a plot with Month revealed that there was no real trend between BA and Month:



In light of this I removed Month from the log model and carried out predictions using this model. This further reduced the RMSPE down to 52.08, a significant improvement over the initial RMSPE. Ultimately the transformations of DLG and CPG actually resulted in a decreased model performance and so I neglected to remove them from the final model, shown below:

final.mod <- lm(log(BA) ~ STJ + DLG + BATS + RTO + CPG + PRU + RMV + TSCO + LLOY + CNA, data = lse)

## Further Improvements

There were a few methods of potentially improving the model’s predictive power that I noticed while working on this project but neglected to implement due to either the amount of time required to add them or a lack of how to implement them. If I were doing this project again, I would attempt to factor the date into the model, either as a numerical factor or with a more complicated method as I believe this would have some impact on the closing prices. Similarly, I would’ve like to include the weekdays as a dummy variable so I could see if there was a relationship between the closing prices and specific days of the week.

Ultimately, I think the choice of a linear regression model is not the best method for predicting future stock prices, as all stocks in a market show some level of correlation with one another and they do not account for autocorrelation that the BA closing prices show with the past BA closing prices.