Creating a Log Return Model to Predict the Future

Summary

Our goal for this report is to find out if it is possible to use the log returns of Ford Motor Corp. and Toyota Motor Corp. to predict the log returns of General Motors. If possible, we would be able to predict the performance of General Motors ahead of time, similar to a psychic with a crystal ball, except in this case our crystal ball will be a linear regression model built on Ford and Toyota. In this report, we'll start with an introduction to further explore the details of how the ball even works, and then dive into the calculations needed to solve the problem, followed by a conclusion about how well our crystal ball predicts the future.

Introduction

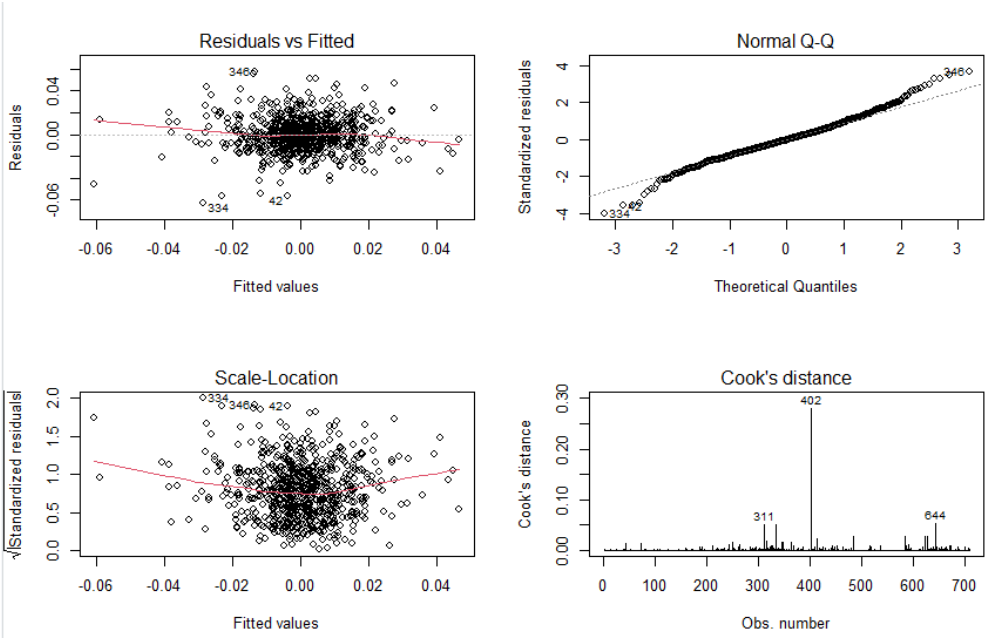
In order to train the crystal ball to predict the log returns of one company using others, it is best to make sure the data we put into it is from a normal distribution, as several other tests to ensure that our ball is working as it should, and the data going into it isn't bad at predicting what we need. To do that, we may have to trial and error a few times with diagnostic tests, and possibly choose a better model to train based on those tests, but after a few rounds the ball should be ready and offer us the answers we desire. Now, without further ado, on to the actual analysis!

Main Body

We first start with a summary of the results and find that the Adjusted R squared is extremely low with a value of 0.3757. This essentially means that the linear regression model isn't a good predictor right now.

```
Residual standard error: 0.01572 on 706 degrees of freedom
Multiple R-squared:  0.3775,    Adjusted R-squared:  0.3757
F-statistic: 214.1 on 2 and 706 DF,  p-value: < 2.2e-16
```

As such, we need to analyze using the diagnoses whether there are outliers that need to be removed. We do that by running the aforementioned diagnostic tests, called Diagnostic Figures. There are 6 total, but we only need to use 4 for our purposes.



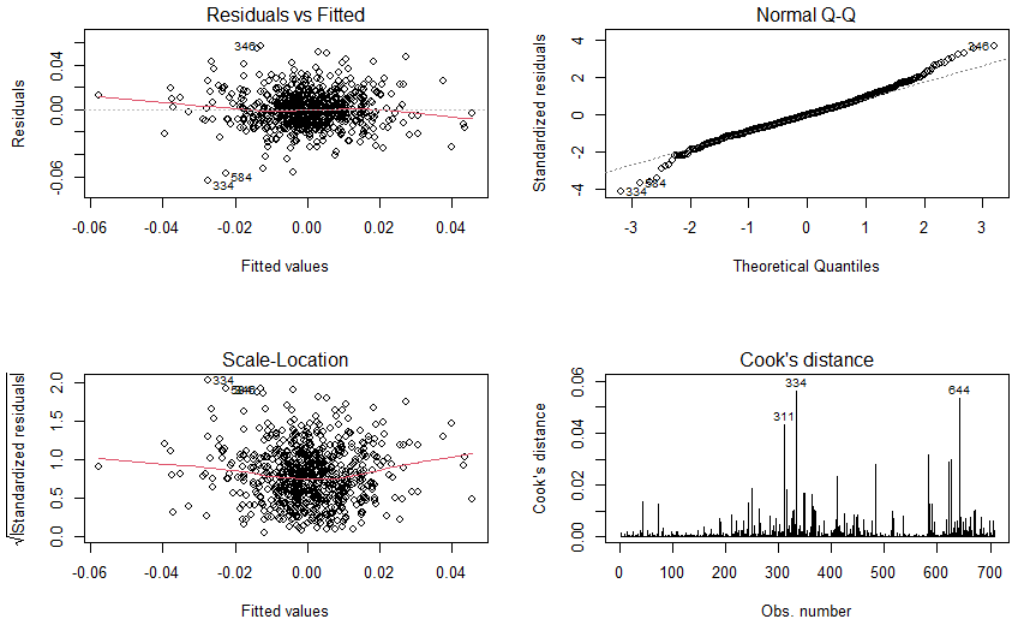
Here's a quick explanation of how to read the figures.

In the Residuals vs Fitted graph, if the red line is close to the dotted line, then that is a good sign of linearity. Linearity means that our use of a *linear* regression would be a good model choice. However, here we can see the red line is close, but the ends deviate at extreme values. Let's try removing some extreme values, or outliers, to solve that.

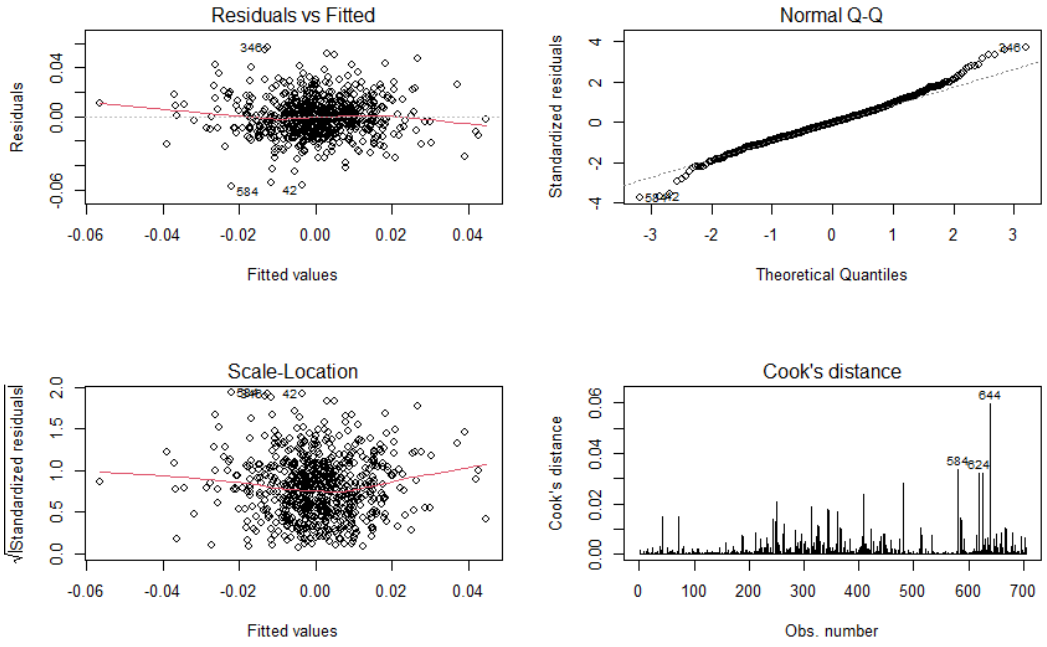
In the Normal Q-Q graph, we check for normality of the data or the chance that the data came from a , by dividing it into quantiles, more commonly known as percentiles, and comparing the quantiles against each other. Notice how the points are linear in the center, but near the extremes they curve off. The linear portion means the two populations' data came from populations that are normal, so using both Toyota and Ford's data are good to predict GM's log returns. Again, removing outliers near the end may once again solve this.

In the Scale-Location graph, if the red line is roughly straight with a cloud of data closely grouped around it, it is considered a good indicator of homoscedasticity or "same variance". Essentially, while there will inevitably be "error" in your predictions, the error should not change with your data, but stay the same. So if GM's return increased 1:1 against Toyota and Ford's, but became 2:1 and 3:1 as the returns got higher, the error does not stay the same, and homoscedasticity fails. We see this in the red line's dip in the center around the data.

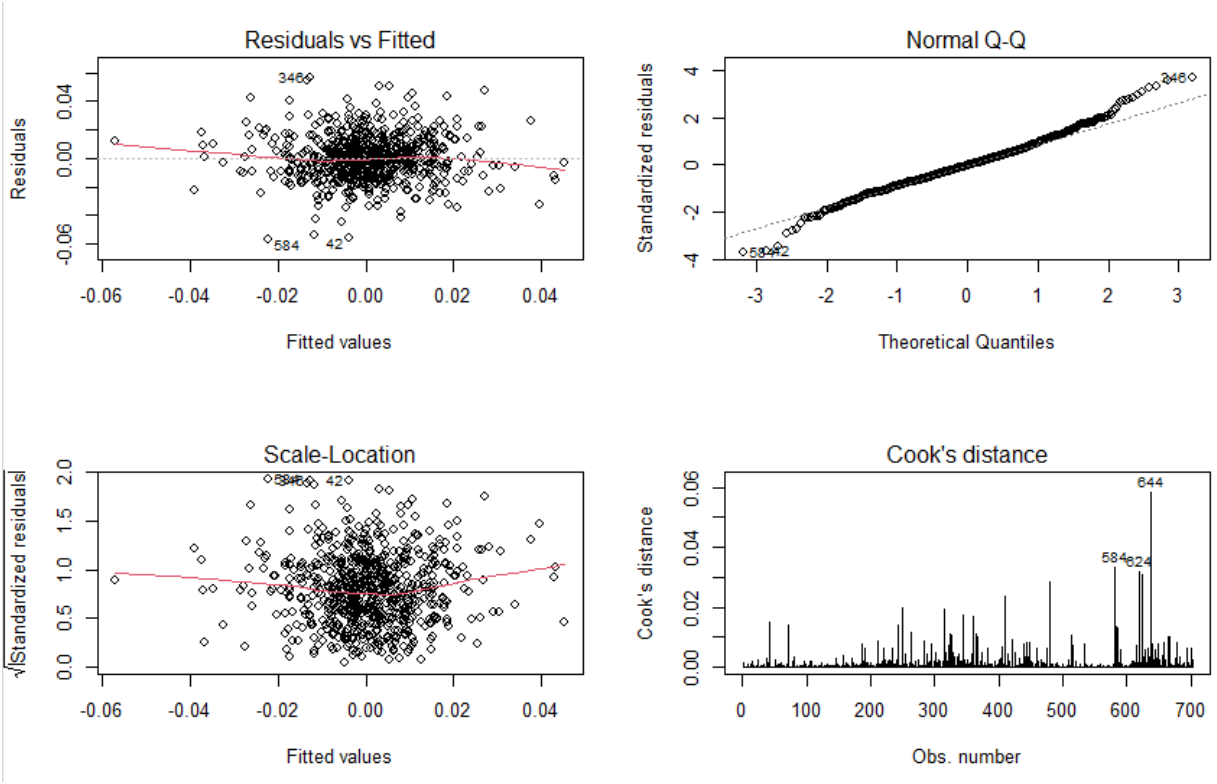
Finally, in the Cook's Distance graph, we can pinpoint the most influential outliers, and see how they negatively affect our regression model. Data row 402 should be removed as it is so large it dwarfs even the other outliers.



Removing line 402 now has zoomed our model in better on the future we want, and we can see a little clearer now. Notice in the Cook's distance chart that row 334 has finally appeared, while in the first set had only shown in the other 3 plots.
Given that all 3 of the other plots are still curving at the extremities, we can go ahead and remove this second set (311, 334, 644) of outliers as well.



A little better, as you can see the red line moving closer to the dotted line, but more fitting is needed – so we keep removing outliers.



- Looking at the Residuals vs Fitted graph after this round, it looks more close to straight now, and we don't want to continue removing too many rounds of outliers, so we will stay with this dataset for the model.

Conclusion

It is not only possible, but also a good fit to use the log returns of Ford and Toyota to predict GM's log returns. Looking at the Normal Q-Q graph, with the exception of extreme outliers with z scores above 2 and below -2, the quantile comparisons of the two populations is fairly normal. The cluster of data around the Residuals vs Fitted being fairly close to the red line indicates that the residuals aren't super far off and for the most fairly evenly distributed from the red line, indicating homoscedasticity and that the errors that are normally incalculable by the model are at least constant throughout the model.

Appendix

```
file=read.csv("w_logret_3automanu-1.csv",header=FALSE,sep=",")
file
names(file)<-c('Toyota','Ford','GM')

lm1=lm(GM~Toyota+Ford,data=file)
lm1
summary(lm1)
par(mfrow=c(2,2))
plot(lm1,which=c(1:4))

#remove 402
file<-file[-402,]
lm2<-lm(GM~Toyota+Ford,data=file)
plot(lm2,which=c(1:4))

#remove 311, 334, 644
file<-file[-c(311,334,644),]
lm3=lm(GM~Toyota+Ford,data=file)
plot(lm3,which=c(1:4))

#remove 346,624,582
file<-file[-c(346,624,584),]
lm3<-lm(GM~Toyota+Ford,data=file)
plot(lm3,which=c(1:4))
```