MM916 Regression Modelling Project

Introduction

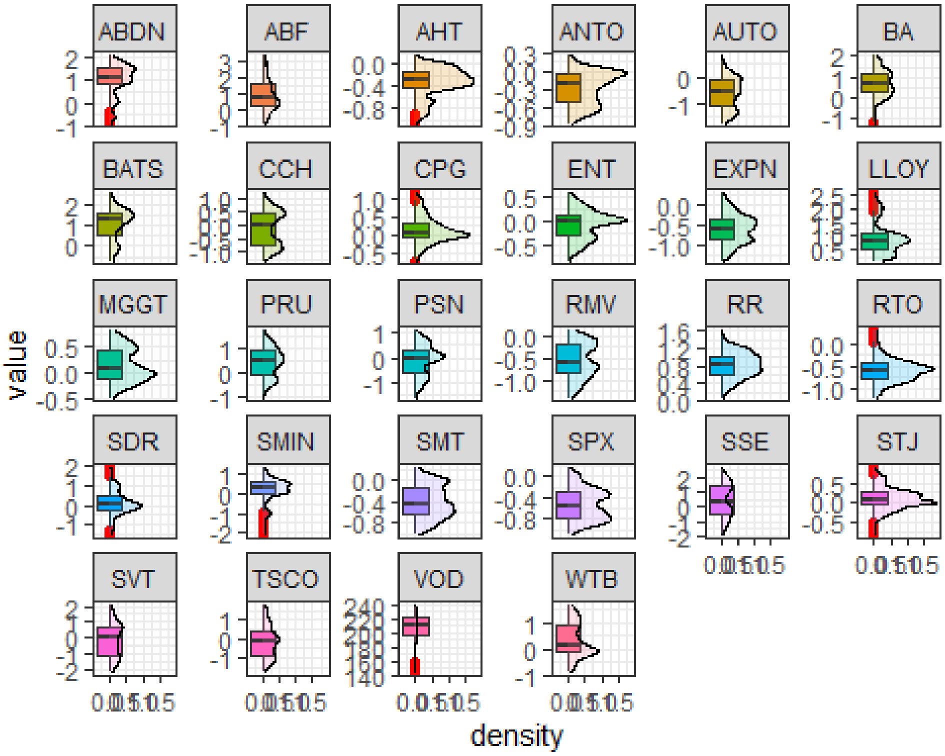
- In this project we will analyse and explore a large dataset containing the daily closing stock-prices of 28 large companies. We will build and assess a linear regression models used to explain the variability in the daily stock price of Vodafone using stock price data from other companies.

Explore the data (15 Marks)

We start by importing the libraries- ***tidyverse***(collection of packages for data exploration), ***broom***-for tidying models and turning them into tibbles**. *lindia*** for easy scheming of linear regression. ***patchwork*** for combining separate *ggplots* into the same graphic. ***hydroGOF*** for implementing goodness of fit measures.

- We first describe the dataset we are using in this project. The dataset we have contains prices for Vodafone and 27 other companies in the FTSE (Financial Times Stock Exchange) 100 index. The dataset contains 32 columns and 757 rows of data. These variables include Date, Year, Month, Weekday to identify an each observation. There are other numeric columns including **VOD, SVT** etc showing the prices of Vodafone (VOD) and other 27 companies. The dataset includes daily data recorded from 2016-01-05 to 2018-12-29. The table below shows the first 5 rows and the first 5 columns from the data.

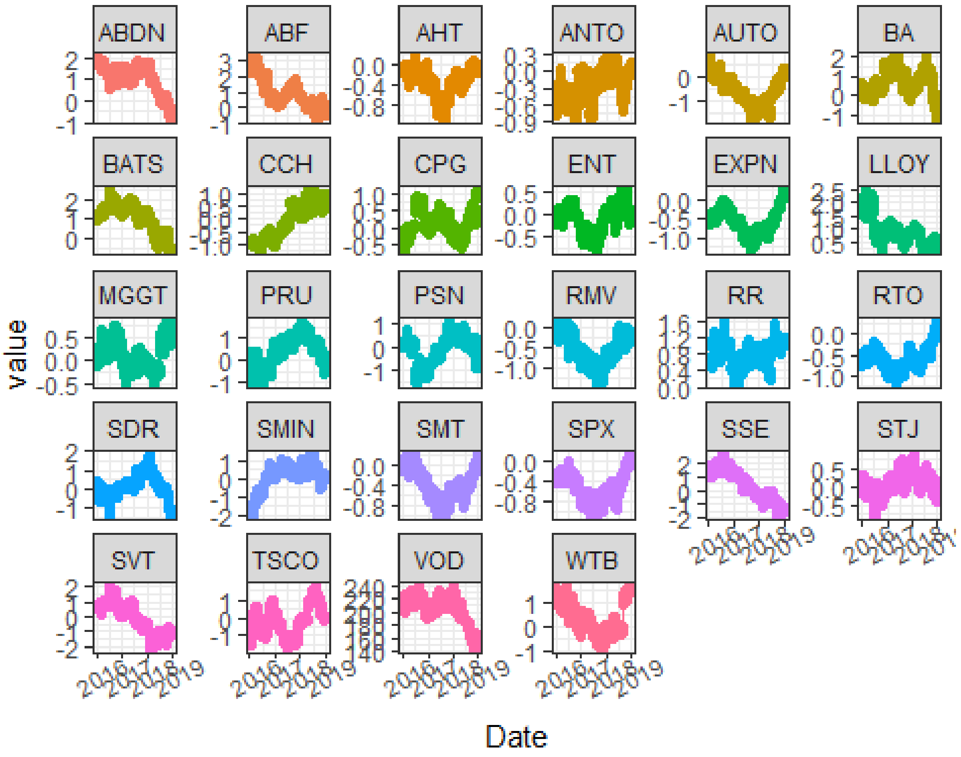
```
        Date Year Month   Weekday    VOD
1 2016-01-05 2016     1   Tuesday 219.70
2 2016-01-06 2016     1 Wednesday 219.15
3 2016-01-07 2016     1  Thursday 218.20
4 2016-01-08 2016     1    Friday 223.05
5 2016-01-09 2016     1  Saturday 220.80
```



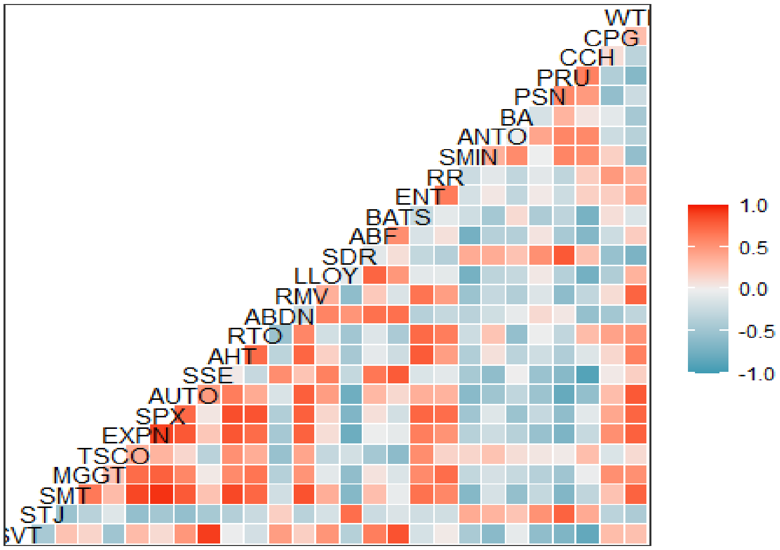
-

- The boxplots above shows the distribution of daily stock prices for Vodafone in 28 Companies. Most variables are bimodal in nature as shown by the density plots. We note that RR distribution prices is closely normal. High variability in prices is depicted in ANTO, AUTO, CCH, SVT, TSCO, WTB and MGGT companies while STJ, RTI,LLOY, SMIN. SDR and CPG prices have a lower variability. We also note a number of outliers in variables with lower variability which might be good to investigate further.

- I have used *facetwrap* as it will as it will arrange panels into rows and columns and give me a clear picture in one glance.



- The plot above shows the temporal time series distribution of stock prices of Vodafone from different companies. We note that Vodafone stock prices for some companies have been dropping since 2018 for some companies like SDR, ABDN, BA just to mention a few. However, for companies like CCH, CPG, EXPN, RTO and other have been increasing since then. For some companies like VOD, SMIN,SMIN and other it dropped for some few months it started increasing after some time with the period of 2018-2019. This is something interesting which we want to understand.



- The plot above shows the correlation values for different variables. We note that some variables have a high correlation, meaning that there is a strong linear relation between those variables. This is useful information for us in the model building because we will avoid adding both variables with higher correlation into the model because when some independent variables are strongly correlated with each other then multicollinearity occurs.

- Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model. [1]

Model building (15 Marks)

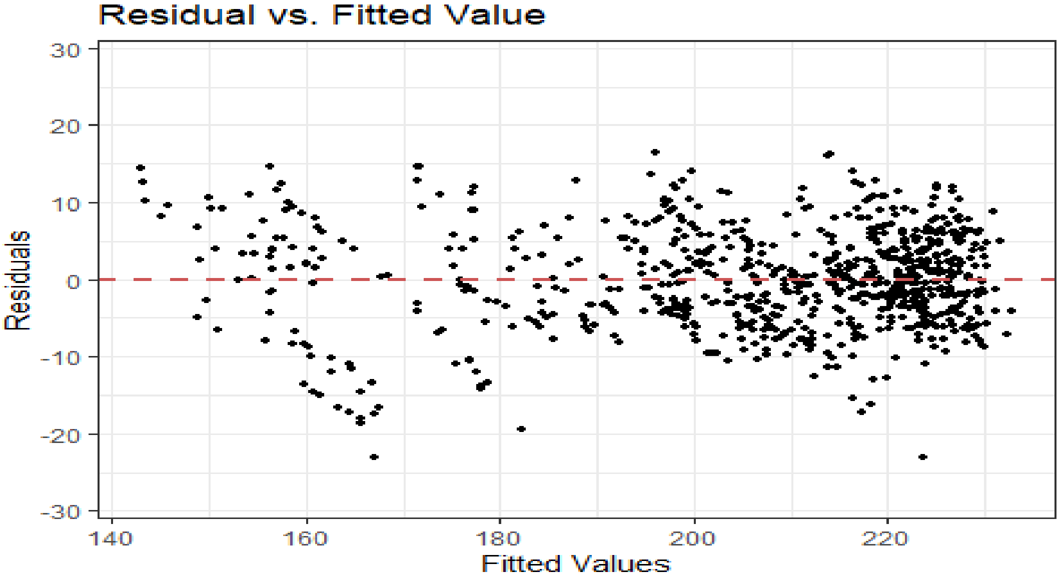
```
# A tibble: 23 x 5
   term        estimate std.error statistic   p.value
   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
 1 (Intercept)  173.        2.69     64.2   1.50e-303
 2 SVT            4.11      1.13      3.64  2.97e-  4
 3 STJ           -1.80      1.58     -1.14  2.55e-  1
 4 SMT           58.8       7.08      8.31  4.61e- 16
 5 MGGT          -1.35      1.85     -0.728 4.67e-  1
 6 TSCO           0.372     0.585     0.636 5.25e-  1
 7 EXPN          -8.60      4.22     -2.04  4.19e-  2
 8 SPX          -39.2       5.85     -6.71  3.96e- 11
 9 AUTO         -20.4       1.52    -13.4   1.36e- 36
10 SSE            3.00      0.869     3.46  5.77e-  4
# ... with 13 more rows
```

- We adopted forward, backwards and stepwise methods in building the model. First we build models predicting the prices of Vodafone (VOD) using each variable independently. We then considered addition or deletion of one term at a time. At the same time we did an analysis of variance to explored how much of the variability in the prices of Vodafone is explained by variables added into the model. We excluded those variables that are not significant in explaining variability by checking their corresponding p-values i.e P-Values>0.05.

- $R^2 \& R^2_{adj}$ were used to selection the best model. The model with the highest R^2/R^2_{adj} was chosen as the best model for us.

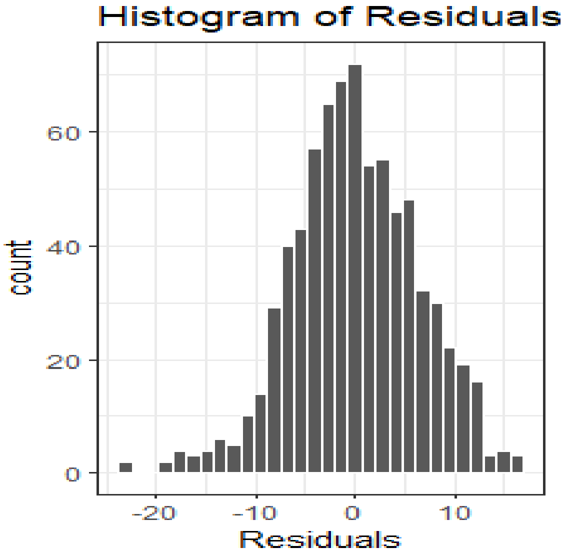
- Once we selected the model we went ahead and checked the regression assumptions as outlined below.

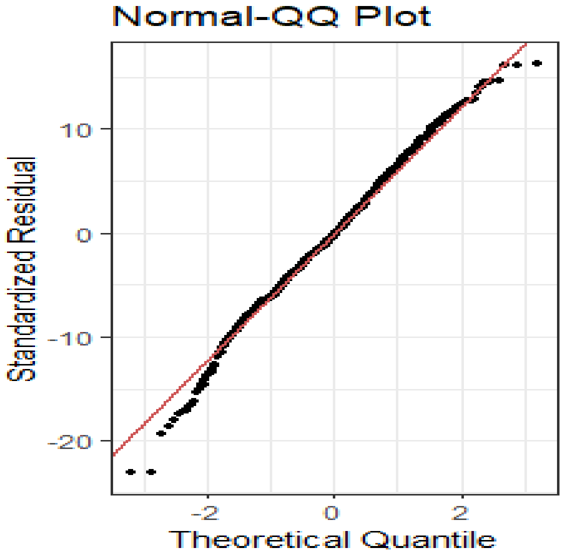
Validity of assumptions



- The residuals versus fitted plot above shows that the there is no clear systematic pattern followed. This means that the assumption of constant variance is not violated.

- The assumption of the independence of residuals is also not violated.





- The histogram shows that the distribution of residuals is close to normal. The Normal-QQ plot shows that there is a slight departure from normality of residuals having shown that we have heavy tails. We could apply transformations and check if this assumption will hold.

- We have checked the three main regression assumption above; normally distributed, share a common variance (homoscedastic) and independence of one another. However, there are many other problems that could be checked for example; multicollinearity, missing independent variables, outliers, leverage and influence. Since we haven't checked for multicollinearity in this case, we are therefore not sure about the estimation of standard errors for particular slope estimators.

Discussion of Model (5 Marks)

```
[1] 0.9180291
```

An R^2 values of 0.91268 shows that ~ 91.27% of the variability in vodafone prices is explained by; SVT, STJ, SMT, MGGT, TSCO, EXPN, SPX, AUTO, SSEAHT, RTO, ABDN, RMV, LLOYSDR, ABF, BATS, ENT, RR, SMIN, and ANTO vodafone stock prices in these companies.

```
     ME  MAE   MSE RMSE NRMSE % PBIAS %  RSR  rSD  NSE mNSE rNSE    d   md
rd
[1,]  0 5.01 40.18 6.34    28.6       0 0.29 0.96 0.92 0.71  0.9 0.98 0.85
0.97
        cp    r   R2  bR2  KGE   VE
[1,] -4.87 0.96 0.92 0.92 0.94 0.98
```

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. [2]

However, after applying transformations with LOG and SQRT we note that the value of R-squared improves slightly from 0.9180 to 0.920 and 0.919 for LOG and SQRT respectively.

References

[1] https://www.investopedia.com/terms/m/multicollinearity.asp

[2] https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

https://www.rdocumentation.org/packages/GGally/versions/1.5.0

https://patchwork.data-imaginist.com/

https://data.library.virginia.edu/understanding-q-q-plots/

https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

https://people.duke.edu/~rnau/testing.htm

https://www.statisticshowto.com/homoscedasticity/