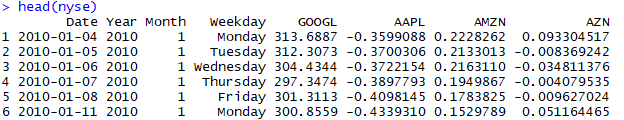
|  |
| --- |
|  |

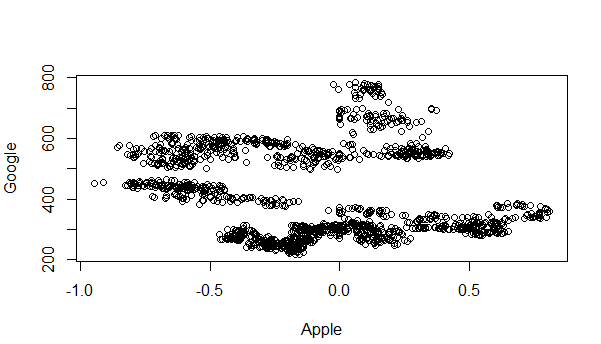
|  |
| --- |
| MM916 Regression Modelling Project |
| Yash Sharma |
| [201865508] |

Introduction

The analysis of the closing price of each stock by using linear regression, firstly it should be loaded and some overview of the data by using the **head()** and **dim()** inbuilt function and **plot()** for some visualisation, so that it shows all the data and graph accordingly.



F:\Rproject for regression\dim.PNG



Then starting from the outline of the project is divided into two parts :

Part 1: Model ﬁtting and interpretation

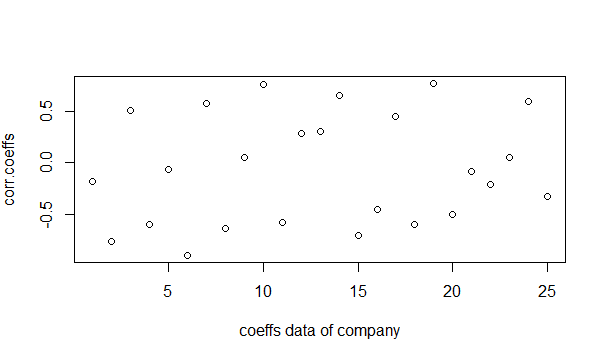
Part 2: Prediction and validation

Part 1: Model ﬁtting and interpretation

In this, there is 4 task in this, and they are explained respectively

1) Calculate the sample correlation coeﬃcient between the closing price of each stock and that of Google and use a single plot to summarise these.

For calculating the sample correlation coefficient, there is the function **cor()** that is used with the help of **for()** loop, and it is only on the company data because there is data of date, year, month and day



(2) Find the five companies whose stock prices have the strongest correlation (in absolute value) with Google's. Fit and brieﬂy interpret a linear regression for Google's stock price using these ﬁve companies prices as independent variables.

By finding correlation and then sort the five companies from the strongest correlation concerning Google’s stock price. Companies were sorted by the **tail()** and **head()** by this it will give top values on both the positive(-) and negative() because **correlation** coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. The value of r is always between +1 and –1. ... A perfect downhill (negative) linear relationship. –0.70. A **strong** downhill (negative) linear relationship.

Using **lm()** and **summary()** in-built it finds the fit and interpret a linear regression.

The companies those are strong correlation are:

CDE -0.8965774 (Couer Mining)

AMZN -0.7663762 (Amazon)

SAM 0.7657443 (Boston Beer Company)

GIL 0.7588794 (Gildan Activewear)

MSFT -0.7030556 (Microsoft)

(3) Using an appropriate variable selection technique and any transformations of the independent variables, build an improved model for Google's daily closing price.

The technique that is used is a backward variable selection and stepwise selection. In backward selection, this assesses F-statistic for each possible removal, then choose the term to remove that has the smallest F-statistic and such that p-value > alpha (stay). In the first step including all the variable and then in every test excluding the value on the above mention method so in last we get the perfect model of our analysis. Stepwise selection uses the step() function to perform stepwise selection using AIC. This alternates between forwards and backward selection in an attempt to find a model that minimises AIC.

(4) Using your ﬁnal model from part (3), check the regression assumptions using appropriate summary plots, and comment on whether you think that these are valid.

In this, for checking the regression assumption, there is graph plot of residual vs leverage, normal q-q,

Scale-location.

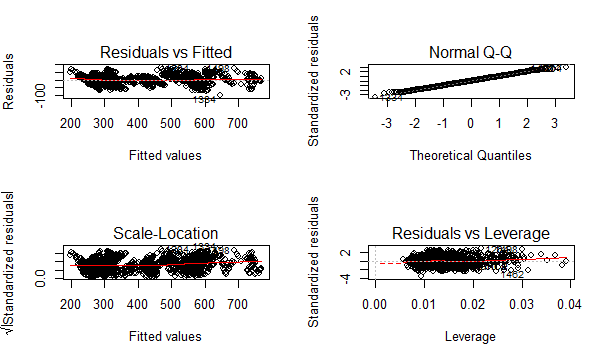
Q-Q plot

You will see from this plot, that the points lie approximately on the line of equality, indicating little evidence of non-normality. To be slightly more formal, we can test the null hypothesis of normality using the Shapiro-Wilk correlation test

Residual vs fitted values

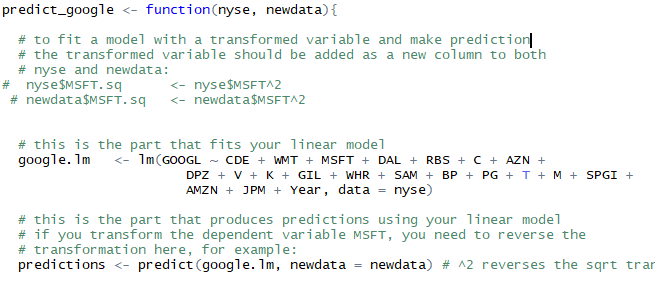
We are looking for evidence of changes in the spread of the residuals as we move from left to right along the x-axis. It is di‑cult to tell here, but there isn't any substantial evidence of this visible.

Residual vs leverage

Cook distance is not shown in this . Therefore the model is valid. 

Part 2: Prediction and validation

(1)Write an R function that uses your model from Part 1 to make future predictions for Google's daily closing price for new rows of independent variables.



(2) Try to improve the prediction accuracy of your model by trying diﬀerent combinations of independent variables. Assess the models using cross-validation to ensure that your improved model is optimal for prediction.

In this Stepwise value of CP and PRESS is the least, so they are the best model.

