Stock price evaluation of Microsoft(MSFT) using multivariable regression

Ashutosh

November 2022

Abstract

Microsoft went public on March 13, 1986, with an IPO price of $21. Over 2.5 million shares were traded on the company's IPO day. Today Microsoft trades at around $241 with 30 million shares traded every day. Therefore, predicting and analyzing share prices is important for every individual and organization in the business of finance and the trading market. In this paper a model's aptness will be discussed using multi-variable regression. The practice of simultaneously observing and analyzing multiple outcome variables is referred to as "multivariate statistical analysis".Understanding the various goals and contexts of various forms of multivariate analysis and how they relate to one another is the focus of multivariate statistics.

I have neither given nor received help (apart from the instructor) to complete this assignment.

Contents

1 Objective

The main aim of the project is to design a model to analyze the relation of price of the stock of Microsoft with relevant stock influencing factors which trades at exchanges as ticker MSFT.

The project has been divided into major sections for clarity of flow and understanding which includes:

1. [1][2]Collection and cleaning of data through relevant research and reliable resources.
2. Application and understanding of Multivariate Regression on the collected data.
3. Application of the same project on RStudio.

An attempt has been made to check for the fitness of the model by carrying out the following:

1. Residual analysis.
2. Examining the plot figures.
3. Checking the reliability of the assumptions.
4. Correcting the violations of the model.
5. Reforming the model by eliminating insignificant variables.

2 Model

The model has been designed with the motive to follow the principles of multivariate regression. The sample size of the dataset is 1259 and 21 dependent variables are taken into consideration for the model. Variable "Typical Price" has been considered to be the response/dependent variable. The Typical Price indicator provides a simple, single-line plot of the day's average price of the stock which helps traders and retailers to study the stock prices and movements. 21 variables are considered as explanatory/independent variables.

2.1 Variable Definition

The variables are selected keeping in mind the necessary factors that can influence the price of the stock(MSFT)

2.1.1 Date

Date variable captures the timeframe of the data and stock trend. Also identified as "x1".

2.1.2 MSFT Trading Attributes and factors

1. [3]MSFT Typical Price: Average of High,Low and Close price. Also identified as "y" which is model's dependent variable.
2. MSFT Open Price: The price at which MSFT stock first trades when an exchange opens for the day. Also identified as x2.
3. MSFT High Price: The highest price at which the MSFT stock trades for the day. Also identified as x3.
4. MSFT Low Price: The lowest price at which the MSFT stock trades for the day. Also identified as x4.
5. MSFT Close Price: The price at which MSFT stock trades last when an exchange closes for the day. Also identified as x5.
6. MSFT Adjusted Close Price: Adjusted close is the closing price after adjustments for all applicable splits and dividend distributions. Also identified as x6.
7. MSFT Volume: Volume of MSFT stock trades. Also identified as x7.
8. Dividends: Categorical variable which indicates whether the dividend was given at a date or not. Also identified as x8.
9. Dividend Amount: Amount of dividends paid. Also identified as x9.
10. MSFT Market Capitalization: Refers to the total value of all a company's shares of stock. Also identified as x10.
11. Gross Profit: Gross profit is the profit a company makes after deducting the costs associated with making and selling its products. Also identified as x11.

12. Net Common Stock: Refers to the number of shares of a company and are found on the balance sheet. Also identified as x12.

13. Gross Income: Gross income is all the money you earn before taxes and other deductions are subtracted. Also identified as x13.

14. Shareholders Equity: Refers to a company's net worth or the total dollar amount that would be returned to its shareholders if the company is liquidated after all debts are paid off. Also identified as x14.

2.1.3 SP500 Trading Attributes

1. SP500 Volume: Volume of SP500 stock trades. Also identified as x15.

2. SP500 Typical Price: Average of High,Low and Close price. Also identified as x16.

2.1.4 DJI Trading Attributes

1. DJI Volume: Volume of DJI stock trades. Also identified as x17.

2. DJI Typical Price: Average of High,Low and Close price. Also identified as x18.

2.1.5 Nasdaq Trading Attributes

1. Nasdaq Volume: Volume of Nasdaq stock trades. Also identified as x19.

2. Nasdaq Typical Price: Average of High,Low and Close price. Also identified as x20.

2.1.6 Volatility Index (VIX)

1. VIX: The VIX Index is a calculation designed to produce a measure of constant, 30-day expected volatility of the U.S. stock market. Also identified as x21.

2.2 Regression Model

The model is designed of the form:

$$y = \beta 0 + \sum_{i=1}^{n} \beta ixi \ \ , where \ \ n = 21$$

3 Model Findings

After running the regression of the model using RStudio the following is the summary of the model:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19 +
    x20 + x21, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.109e-12 -3.320e-14 -1.400e-15  2.740e-14  7.015e-12

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  1.175e-13  4.652e-13  2.520e-01 0.800697
x1          -3.940e-16  2.210e-16 -1.783e+00 0.074911 .
x2          -1.811e-15  4.947e-15 -3.660e-01 0.714336
x3           3.333e-01  6.231e-15  5.349e+13  < 2e-16 ***
x4           3.333e-01  5.667e-15  5.882e+13  < 2e-16 ***
x5           3.333e-01  2.657e-14  1.254e+13  < 2e-16 ***
x6          -6.677e-14  2.527e-14 -2.642e+00 0.008334 **
x7          -3.607e-21  1.029e-21 -3.504e+00 0.000475 ***
x8          -7.012e-14  3.600e-13 -1.950e-01 0.845583
x9           1.611e-13  6.939e-13  2.320e-01 0.816505
x10         -5.593e-19  1.043e-18 -5.360e-01 0.592040
x11         -9.007e-18  8.086e-18 -1.114e+00 0.265523
x12         -4.397e-18  5.891e-18 -7.460e-01 0.455602
x13          2.040e-18  1.354e-18  1.507e+00 0.132172
x14          9.887e-18  2.192e-18  4.511e+00 7.05e-06 ***
x15         -2.248e-23  1.416e-23 -1.588e+00 0.112543
x16         -1.628e-17  2.650e-16 -6.100e-02 0.951026
x17          7.588e-22  1.394e-22  5.444e+00 6.27e-08 ***
x18         -8.602e-18  2.277e-17 -3.780e-01 0.705592
x19         -1.058e-23  1.152e-23 -9.190e-01 0.358308
x20          1.164e-17  2.371e-17  4.910e-01 0.623674
x21         -2.292e-15  1.848e-15 -1.240e+00 0.215123
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.177e-13 on 1237 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.318e+30 on 21 and 1237 DF,  p-value: < 2.2e-16
```

Figure 1: Regression output

3.1 Observations

1. The model is of the form:

$$\begin{aligned} y = {} & 1.175e^{-13} - 3.94e^{-16} \cdot x1 - 1.811e^{-15} \cdot x2 + 0.3333 \cdot x3 + 0.3333 \cdot x4 \\ & + 0.3333 \cdot x5 - 6.677e^{-14} \cdot x6 - 3.607e^{-21} \cdot x7 - 7.012e^{-14} \cdot x8 \\ & + 1.611e^{-13} \cdot x9 - 5.593e^{-19} \cdot x10 - 9.007e^{-18} \cdot x11 - 4.397e^{-18} \cdot x12 \\ & + 2.04e^{-18} \cdot x13 + 9.887e^{-18} \cdot x14 - 2.248e^{-23} \cdot x15 - 1.628e^{-17} \cdot x16 \\ & + 7.588e^{-22} \cdot x17 - 8.602e^{-18} \cdot x18 - 1.058e^{-23} \cdot x19 + 1.164e^{-17} \cdot x20 \\ & - 2.292e^{-15} \cdot x21. \end{aligned}$$

2. R-squared, otherwise known as R^2 typically has a value in the range of 0 through to 1. [4]A value of 1 indicates that predictions are identical to the observed values.

3. Let us consider a hypothesis for significance of the model as below:

   H0: Model with no independent variables fits the data as well as your model.
   H1: Model fits the data better than the intercept-only model.

   Regression result gives a very small p value which means all the results are significant. Therefore, sample data provide sufficient evidence to conclude that the regression model fits the data better than the model with no independent variables.

   Therefore we can reject the hypothesis and conclude that the regression model fits the data better than the model with no independent variables.

4. To test the hypothesis of significance of the variables we take the following conditions:

   H0: Variable is significant for the model.
   H1: Variable is not significant for the model.

   For variables who has p-values less than 0.01 we will accept the null hypothesis and mark them as significant.

   For independent variables x1,x3,x4,x5,x7,x14 and x17 we therefore accept the hypothesis of significance at alpha = 0.01 as the p-values are less than 0.01. We can say that these variables are significant with 99% confidence

3.2 Gauss Markov Assumptions

Gauss–Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.

1. ASSUMPTION 1:
   There is a linear relationship between the predictors (xi) and the outcome (y).

   We check the linearity of the data by looking at the Residual vs Fitted plot. Ideally, this plot would not have a pattern where the red line is approximately horizontal at zero.

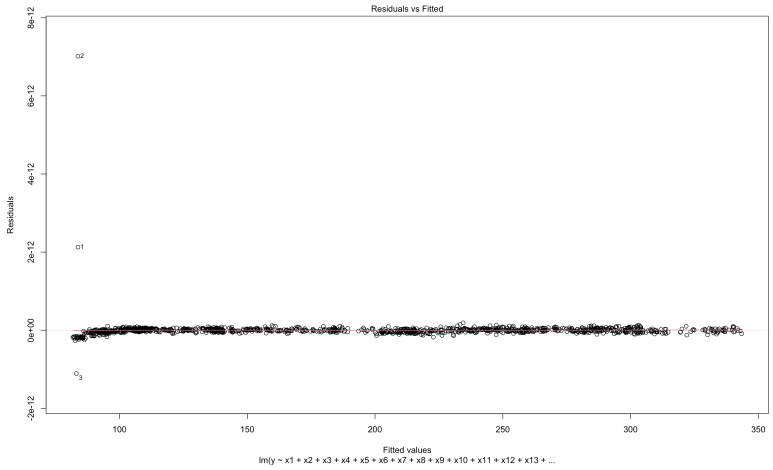

Figure 2: Residual vs Fitted plot

Conclusion: In the above plot we can see that there is no clear pattern in the residual plot. The number of data points are large that it is difficult to zoom in and find out patterns. This would indicate that we succeeded to meet the assumption that there is a weak positive linear relationship between the predictors and the outcome variable.

2. ASSUMPTION 2:
   Predictors (xi) are independent and observed with negligible error

   One of the conditions for a variable to be an independent variable is that it has to be independent of other variables. i.e one shouldn't be able to derive the values of a variable using other independent variables.
   The easiest way for the detection of multicollinearity is to examine the correlation between each pair of explanatory variables. If two of the variables are highly correlated, then this may indicate the possible source of multicollinearity. However, pair-wise correlation between the explanatory variables may be considered as the sufficient, but not the necessary condition for the multicollinearity.

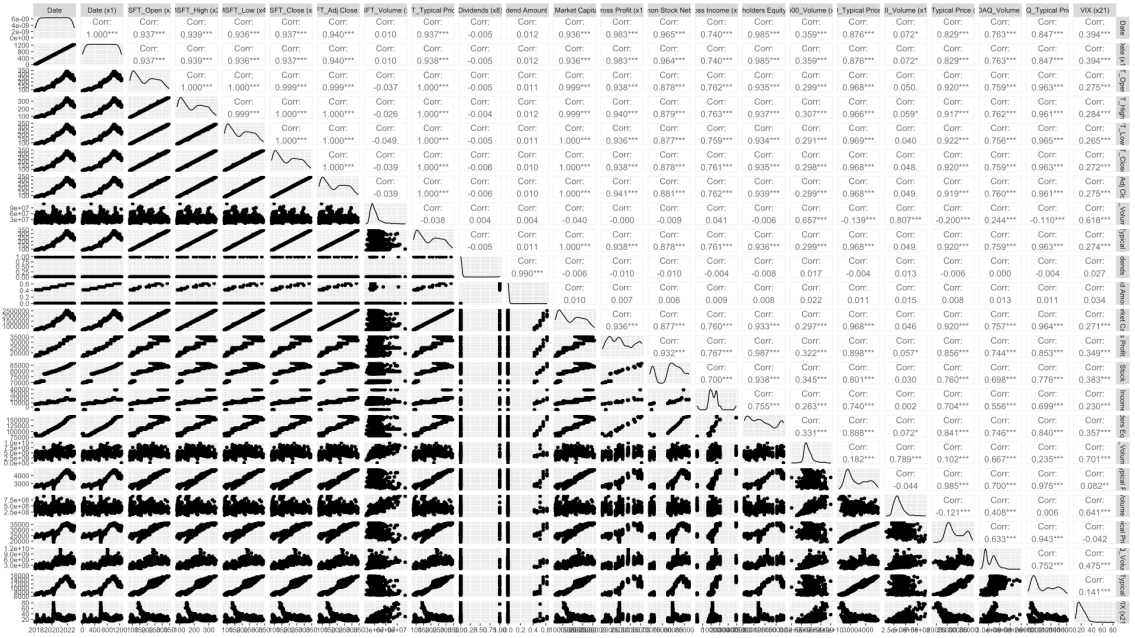

Figure 3: ggpairs plot
Note: The plot is better observed using r code and zoomed out.

Another method to check for multicollinearity is using the vif function in R.

The general range of vif output is:

```
VIF ~ 1: Negligible
1 < VIF < 5 : Moderate
VIF > 5 : Extreme
```

```
> vif(model)
         x1          x2          x3          x4          x5          x6          x7          x8          x9
1.714202e+02 3.763663e+03 6.097251e+03 4.829048e+03 1.086148e+05 1.010398e+05 4.634241e+00 5.381420e+01 5.385587e+01
        x10          x11          x12          x13          x14          x15          x16          x17          x18
9.173494e+03 6.142201e+01 2.570612e+01 3.162913e+00 1.074799e+02 6.352000e+00 8.514376e+02 6.293050e+00 2.239860e+02
        x19          x20          x21
8.477033e+00 1.212871e+02 6.742011e+00
```

Figure 4: vif of model

H0: Variables are not multicollinear in nature.
H1: Variables are multicollinear in nature.

Conclusion: Through the vif command we can see that the variables are multicollinear in nature. Hence the assumption is violated. Therefore we reject H0.

3. ASSUMPTION 3:
Residual Errors have a mean value of zero

We can check this assumption by looking at the same residual vs fitted plot. We would ideally want to see the red line flat on 0, which would indicate that the residual errors have a mean value of zero.

Conclusion: From the residual vs fitted plot we can see that the red line is flat at 0. Therefore, the residual errors have a mean value of zero for the model.

4. ASSUMPTION 4:
Residual Errors have constant variance

We can check this assumption using the Scale-Location plot. In this plot we can see the fitted values vs the square root of the standardized residuals. Ideally, we would want to see the residual points equally spread around the red line, which would indicate constant variance.

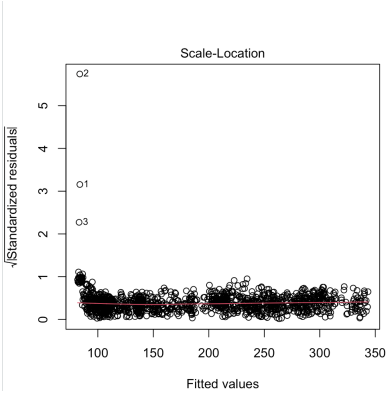

Figure 5: Scale Location plot

Conclusion: From the above plot we can see that the residual points are mostly equally spread around the red line, which indicates that the variance is constant but there are some outliers in the plot as well which hint of some variable variance. To test this statistically we would use Breusch Pagan Test for detecting heteroskedasticity.

3.2.1 Breusch Pagan Test for detecting heteroskedasticity

Breusch-Pagan test involves using a variance function and using a chi-square test the null hypothesis that heteroskedasticity is not present (i.e. homoskedastic) against the alternative hypothesis that heteroskedasticity is present.

H0: The variance is constant.
H1: The variance is not constant.

All we need to check at is the p-value to determine whether or not we should reject the null hypothesis. If the p-value is less than the level of significance, then we reject the null hypothesis.

```
> bptest(model) #there is some heteroscedasticity present in the model

        studentized Breusch-Pagan test

data:  model
BP = 66.803, df = 21, p-value = 1.133e-06
```

Conclusion: Since p-value = 1.133e-06 which is less than 0.05, we can reject the null hypothesis that the variance is constant. Hence the Gauss Markov assumption that residual errors have constant variance is violated.

5. ASSUMPTION 5:
Residual Errors are independent from each other and predictors (xi)

[5]To check independence, we can plot residuals against any time variables present. A pattern that is not random suggests lack of independence.

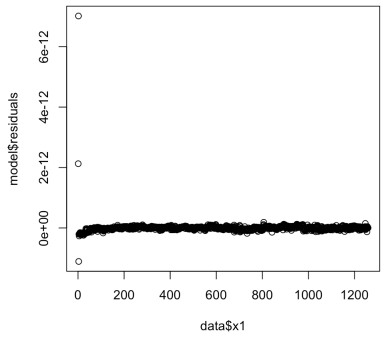

Figure 6: Residual vs Time plot

Conclusion: We see that the residual plot does not have a random pattern hence we can say that the residual errors are not independent. Since it is known that the residuals sum to zero, they are not independent, so the plot is really a very rough approximation.

The statistical way to check the assumption of independence of residuals is using the Durbin Watson test.

The null hypothesis states that the errors are not auto-correlated with themselves (they are independent). Thus, if we achieve a p-value greater than 0.05, we would fail to reject the null hypothesis. This would give us enough evidence to state that our independence assumption is met.

H0: Errors are not auto-correlated with themselves (they are independent).
H1: Errors are auto-correlated with themselves (they are dependent).

```
> durbinWatsonTest(model) #we reject H0, therefore the predictors are not independent
 lag Autocorrelation D-W Statistic p-value
   1       0.1683817      1.586093   0.036
```

Conclusion: The p-value from the DB test of the model is 0.036 which is smaller than 0.05. Therefore, we reject the null hypothesis and state that the errors are not independent therefore violating the assumption.

6. ASSUMPTION 6:
   Normality of error terms
   The residual errors are normally distributed with mean = 0 and variance = σ^2.

   To check for normality of error terms we can refer to the histogram of the error terms and if they follow the shape of the normal distribution then we can say that the error terms are normally distributed.

   H0: Residuals are normally distributed.
   H1: Residuals are not normally distributed.

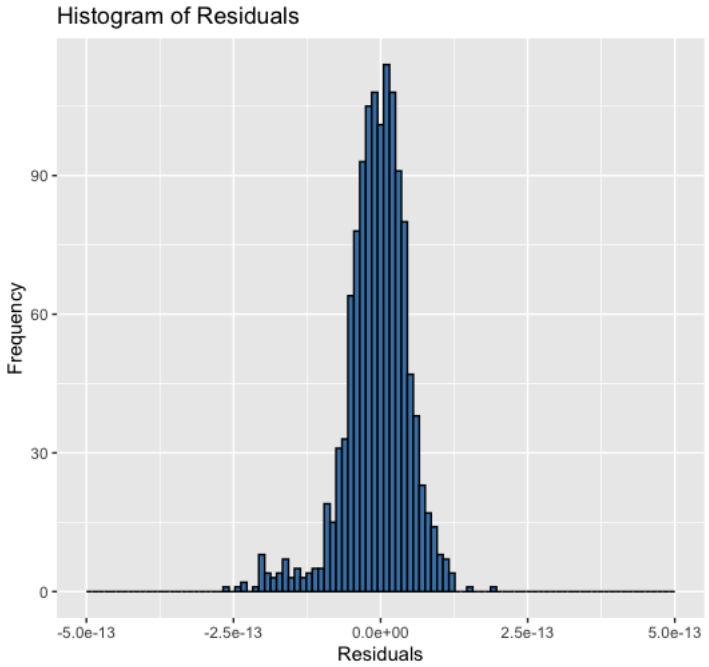

Figure 7: Histogram plot of error terms

Conclusion: The residuals are normally distributed in the histogram, hence we can say that the null hypothesis is not violated and the assumption holds.

3.3 Recommendations for violated assumptions

1. ASSUMPTION 2: Predictors (xi) are independent and observed with negligible error

   [6]The potential solutions to fix multicollinearity between variables include the following:

   (a) Remove some of the highly correlated independent variables.

   (b) Linearly combine the independent variables, such as adding them together.

   (c) Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

   (d) LASSO and Ridge regression are advanced forms of regression analysis that can handle multicollinearity.

   As we consider a solution, remember that all of these have downsides. If we can accept less precise coefficients, or a regression model with a high R-squared but hardly any statistically significant variables, then not doing anything about the multicollinearity might be the best solution.

2. ASSUMPTION 4: Residual Errors have constant variance

   From the above section we figured out that the assumption for having constant variance for residual errors is violated and should be fixed. There are three common ways to fix heteroscedasticity:

   (a) Transform the dependent variable

   One way to fix heteroscedasticity is to transform the dependent variable in some way. One common transformation is to simply take the log of the dependent variable.

   (b) Redefine the dependent variable

   Another way to fix heteroscedasticity is to redefine the dependent variable. One common way to do so is to use a rate for the dependent variable, rather than the raw value.

   (c) Weighted least squares

   [7]Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value.

```
> #correction of heteroscedasticity using general least square method
> gls <- lm(data = data, y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19+x20+x21,
+           weights = 1/model$fitted.values^2)
> bptest(gls)

        studentized Breusch-Pagan test

data:  gls
BP = 0.016583, df = 21, p-value = 1
```

   Conclusion: Since p-value = 1 which is more than 0.05, we can accept the null hypothesis that the variance is constant. Hence the Gauss Markov assumption that residual errors have constant variance is corrected.

3. ASSUMPTION 5: Residual Errors are independent from each other and predictors (xi)

   [8]To fix the violation of independence of error terms we can opt for the following methods:

   (a) Add a column that is lagged with respect to the Independent variable.

   (b) Center the Variable (Subtract all values in the column by its mean).

4. ASSUMPTION 6: Normality of error terms

   [9]Another violation which can be coped with by transforming the data is non-normality of the error. Box and Cox(1964) developed a method for choosing the "best" transformation from the set of power transformations to correct for this violation.

   Let $\lambda \in \mathbb{R}$, then

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda\, \tilde{y}^{\lambda-1}}, & \lambda \neq 0 \\ \tilde{y}\, ln(y), & \lambda = 0 \end{cases}$$

   where

$$\tilde{y} = e^{\dfrac{1}{n}\sum_{i=1}^{n} ln(y_i)}$$

   is the geometric mean of the observations.

4 Reforming the model

The new model is designed using only significant variables that we found out using variable significance hypothesis which are x1,x3,x4,x5,x7,x14 and x17.

Performing regression on the new model will provide the following output:

```
Call:
lm(formula = y ~ x1 + x3 + x4 + x5 + x7 + x14 + x17, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.264e-12 -2.680e-14 -1.000e-16  2.080e-14  7.154e-12

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  1.004e-13  9.591e-14  1.047e+00 0.295319
x1          -7.493e-16  1.057e-16 -7.087e+00 2.28e-12 ***
x3           3.333e-01  4.261e-15  7.823e+13  < 2e-16 ***
x4           3.333e-01  4.303e-15  7.747e+13  < 2e-16 ***
x5           3.333e-01  3.527e-15  9.450e+13  < 2e-16 ***
x7          -3.490e-21  9.669e-22 -3.609e+00 0.000319 ***
x14          5.463e-18  1.328e-18  4.113e+00 4.16e-05 ***
x17          4.666e-22  9.676e-23  4.823e+00 1.59e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.198e-13 on 1251 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.153e+31 on 7 and 1251 DF,  p-value: < 2.2e-16
```

Figure 8: Regression output of new model

Now the new model will become as follows:

$$y = 1.004e^{-13} - 7.493e^{-16} \cdot x1 + 0.3333 \cdot x3 + 0.3333 \cdot x4 + 0.3333 \cdot x5 \\ - 3.490e^{-21} \cdot x7 + 5.463e^{-18} \cdot x14 + 4.666e^{-22} \cdot x17$$

4.1 Observations for new model

1. R-squared has a value of 1 indicating that predictions are identical to the observed values and model is a good fit.

2. Regression result gives a p value less than 0.05 providing sufficient evidence to conclude that the regression model fits the data better than the model with no independent variables.

5 Comparable Plots

This section will provide comparable side to side plot evaluation of different models that have been used throughout the project.

5.1 Old Model v/s New Model v/s Weighted Least Square Models

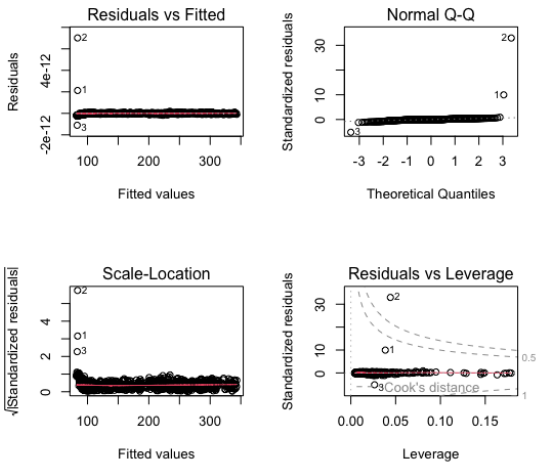

Figure 9: Old Model with 21 variables

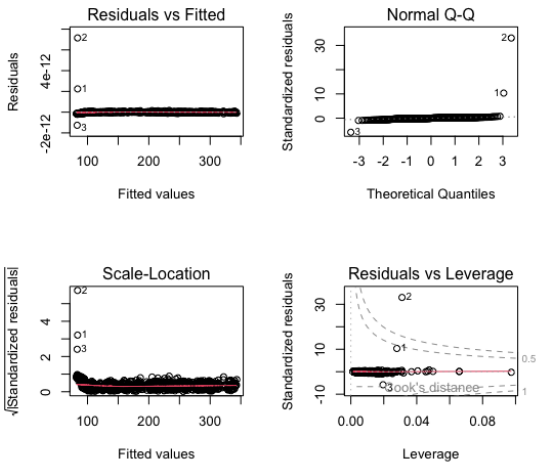

Figure 10: Reformed model with significant variables

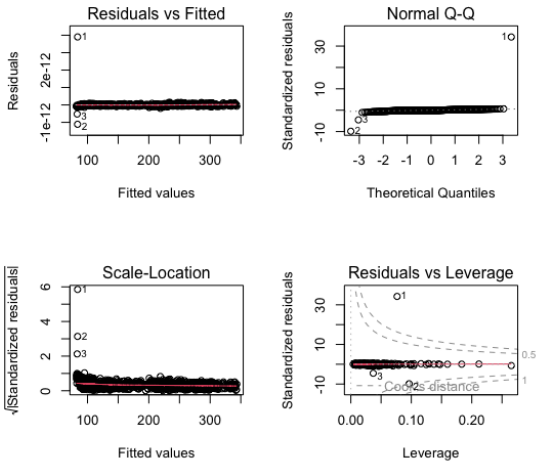

Figure 11: Weighted least square model of old model with 21 variables

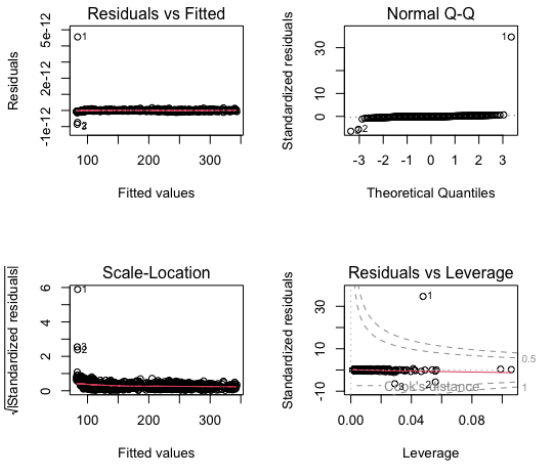

Figure 12: Weighted least square model of reformed model with significant variables

Conclusion:

Among the 4 models the weighted least square models provide a good fit minimizing the outliers of residuals and following the Gauss Markov assumptions for linear regression.

5.2 Geom plots for reformed model

Geoms: A layer combines data, aesthetic mapping, a geom (geometric object), a stat (statistical transformation), and a position adjustment. Typically, you will create layers using a geom_ function, overriding the default position and stat if needed.

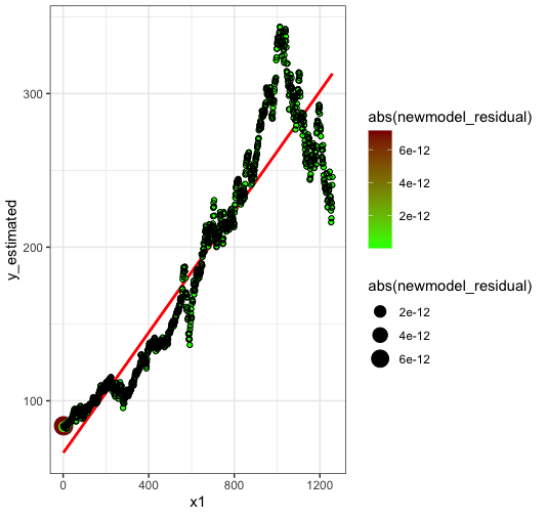

Figure 13: y estimate v/s Date plot

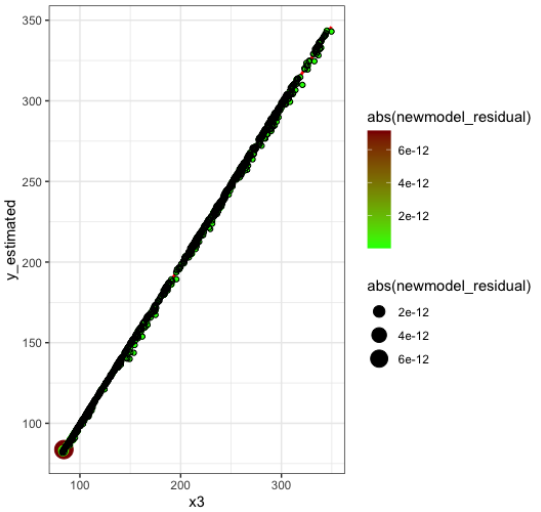

Figure 14: y estimate v/s MSFT High Price

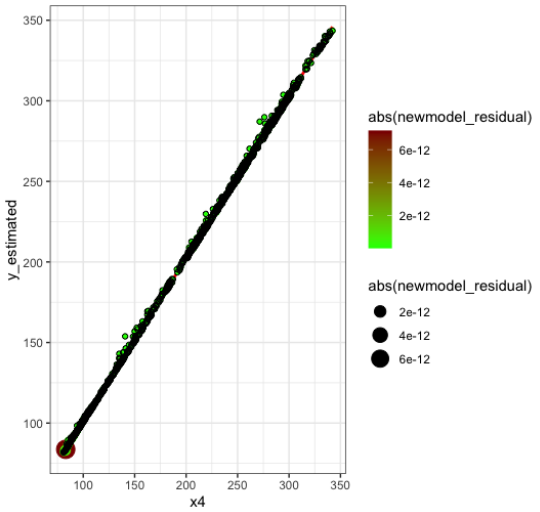

Figure 15: y estimate v/s Low Price

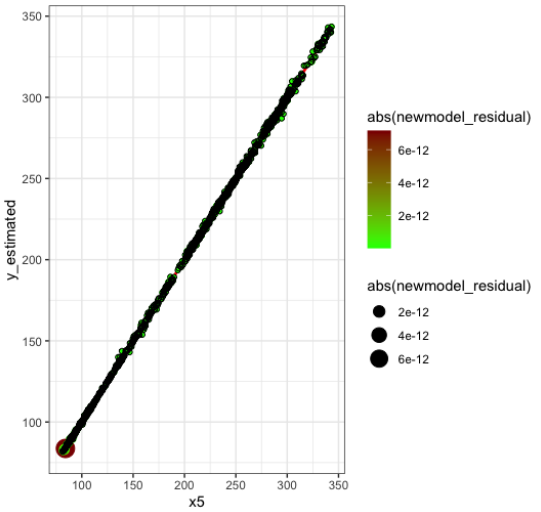

Figure 16: y estimate v/s MSFT Close Price

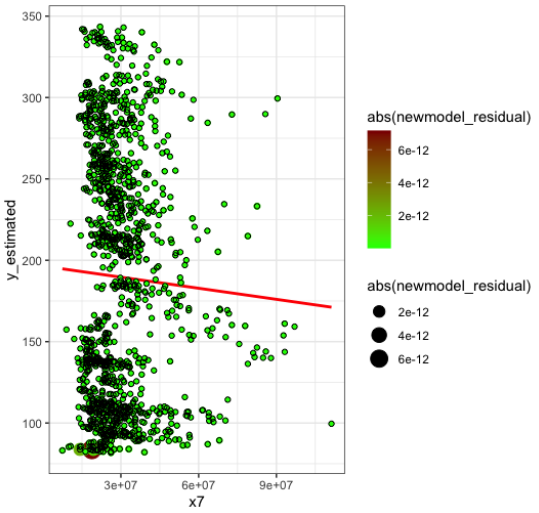

Figure 17: y-estimate v/s MSFT Volume

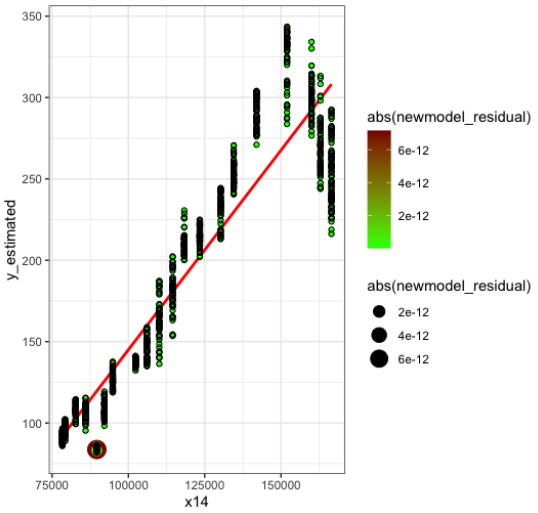

Figure 18: y-estimate v/s Shareholders Equity

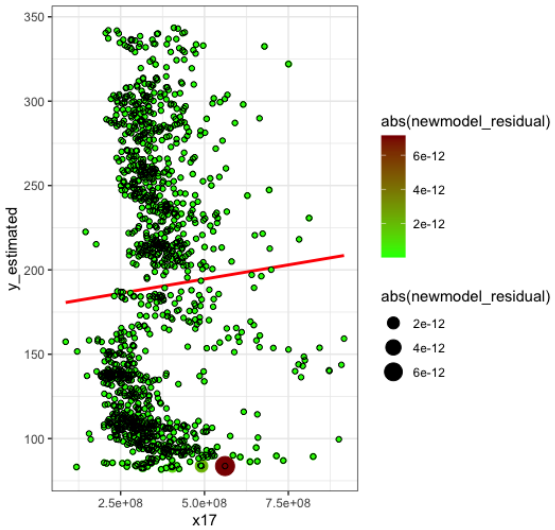

Figure 19: y-estimate v/s DJI Volume

5.3 Histogram of error analysis

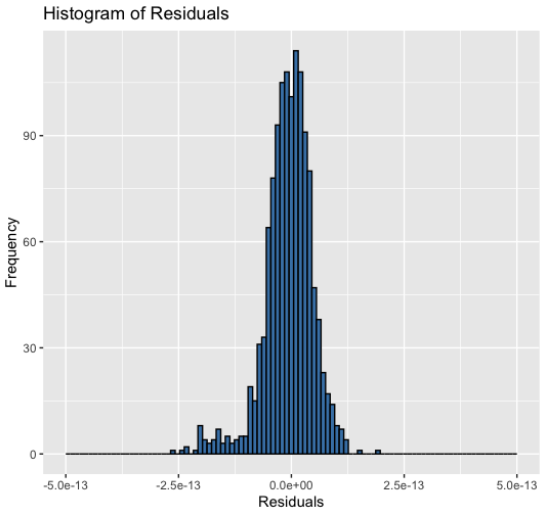

Figure 20: Histogram of error terms of old model

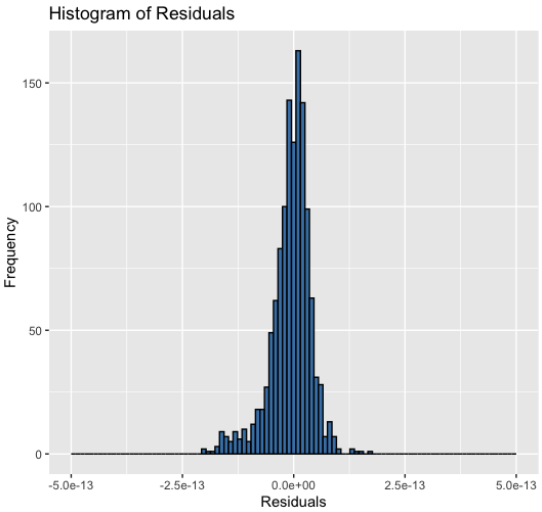

Figure 21: Histogram of error terms of reformed model

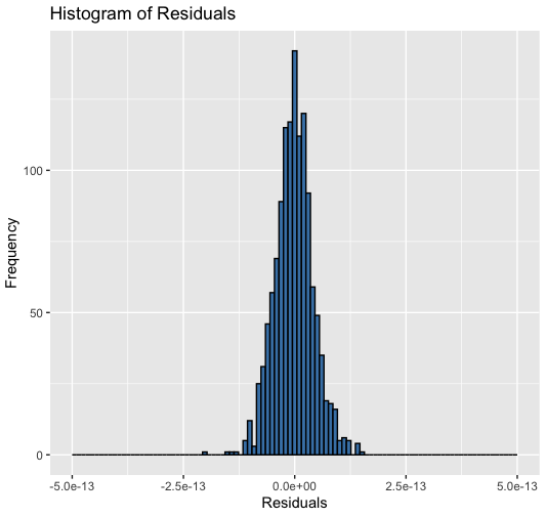

Figure 22: Histogram of error terms of weighted least square model of old model

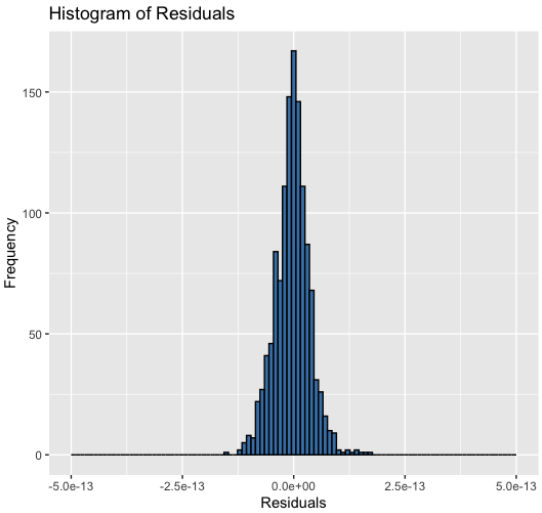

Figure 23: Histogram of error terms of weighted least square model of reformed model

Inference: The histogram plots of weighted least square model have less skewness of terms hence are more normally distributed than original models.

6 Result

The project exposes us to varied understandings of statistics and visualization of model fitting using multivariable regression.

The project cators to the following learnings:

1. Research about the stock and collect/clean the data to be evaluated.
2. Selection of various factors that intuitively effect the stock price.
3. Modelling a relation between the stock price and the factors and running a regression analysis on the model.

Through the findings of the project we can conclude that a model in which the dependent variable is dependent on the significant variables along with no or corrected violations of the Gauss-Markov assumptions is the best fit model.

Final Validated Model:

$$\begin{aligned} TypicalPrice = = {} & 1.004e^{-13} - 7.493e^{-16} \cdot Time + 0.3333 \cdot HighPrice \\ & + 0.3333 \cdot LowPrice + 0.3333 \cdot ClosePrice - 3.490e^{-21} \cdot MSFTVolume \\ & + 5.463e^{-18} \cdot ShareholdersEquity + 4.666e^{-22} \cdot DJIVolume \end{aligned}$$

References

[1] Yahoo finance. [Online]. Available: https://finance.yahoo.com

[2] Fred. [Online]. Available: https://fred.stlouisfed.org/

[3] Investopedia. [Online]. Available: https://www.investopedia.com/

[4] J. Frost. How to interpret the f-test of overall significance in regression analysis. [Online]. Available: https://statisticsbyjim.com/regression/interpret-f-test-\overall-significance-regression/

[5] U. of Texas. Using plots to check model assumptions. [Online]. Available: https://web.ma.utexas.edu/users/mks/statmistakes/modelcheckingplots.html#:~:text=Rule%20of%20Thumb%3A%20To%20check,random%20suggests%20lack%20of%20independence.

[6] J. Frost. Multicollinearity in regression analysis: Problems, detection, and solutions. [Online]. Available: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[7] Y. Czar. Methods for detecting and resolving heteroskedasticity. [Online]. Available: https://rpubs.com/cyobero/187387

[8] Linear regression, machine learning. multicollinearity – how to fix it? [Online]. Available: https://machinelearningmind.com/2019/10/19/multicollinearity-how-to-fix-it/

[9] Dealing with model assumption violations. [Online]. Available: https://academic.macewan.ca/burok/Stat378/notes/remedies.pdf

7 Appendix

7.1 R code

```
#Econometrics Project - Stock Analysis
#include the libraries
library("readxl")
library(memisc)
library(psych)
library(dplyr)
library(lmtest)
library(sjPlot)
library(sgof)
library(ggplot2)
library(foreign)
library(car)
library(hexbin)
library(lmtest)
library(GGally)

#read data from excel and store in data variable
data <- read_excel("/Users/ashutosh/Documents/Econometrics/ECO_Project.xlsx")

#create scatter plots for pairs of variables
ggpairs(data)

#modifying column header for better representation
colnames(data) <-c("Date", "x1", "x2", "x3","x4","x5","x6", "x7", "y", "x8",
                   "x9", "x10", "x11", "x12", "x13",
                   "x14","x15","x16","x17","x18","x19","x20","x21")
glimpse(data)

#preparing multivariate model
model<-lm(data=data, y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+
x14+x15+x16+x17+x18+x19+x20+x21)
summary(model)

coef(model)
model_fitted<-fitted(model) #predicted values of y (y_hat)
model_residual<-residuals(model) #difference between observed and fitted
deviance(model)

#SSE
sse <- sum((fitted(model) - data$y)^2)
sse
#SSR
ssr <- sum((fitted(model) - mean(data$y))^2)
ssr
```

```
#SST
sst <- ssr + sse
sst

#R-square
r_square <- ssr/sst


#ASSUMPTION ONE: LINEARITY OF THE DATA
#Residual vs Fitted plot would not have a pattern where the red line is
#approximately horizontal at zero.
plot(model,1)
#there is linear relation between the predictors and outcome variable

#ASSUMPTION TWO: PREDICTORS (X) ARE INDEPENDENT AND OBSERVED WITH NEGLIGIBLE ERROR

vif(model)

#ASSUMPTION THREE: RESIDUAL ERRORS HAVE A MEAN VALUE OF ZERO
#red line on residual vs fitted plot is flat on 0

#ASSUMPTION FOUR: RESIDUAL ERRORS HAVE CONSTANT VARIANCE
#residual points equally spread around the red line, which would indicate
#constant variance.
plot(model,3) #scale location plot

#ASSUMPTION FIVE: Residual Errors are independent from each other and predictors (xi)

#durbin Watson test
#p-value > 0.05, we would fail to reject the null hypothesis.
durbinWatsonTest(model) #we reject H0, therefore the predictors are not independent


par(mfrow=c(2,2))
plot(model)

#histogram of residual errors old model
ggplot(data = data, aes(x = model$residuals)) +
  geom_histogram(fill = 'steelblue', color = 'black',bins=100) +
  labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency') +
   xlim(-5e-13, 5e-13)


#check for heteroscedasticity
bptest(model) #there is some heteroscedasticity present in the model

#correction of heteroscedasticity using general least square method
gls <- lm(data = data, y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+
```

```
x15+x16+x17+x18+x19+x20+x21, weights = 1/model$fitted.values^2)
summary(gls)
bptest(gls)
plot(gls)

#histogram of residual errors weighted model
ggplot(data = data, aes(x = gls$residuals)) +
  geom_histogram(fill = 'steelblue', color = 'black', bins=100) +
  labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency') +
   xlim(-5e-13, 5e-13)

#---------------------------------------------------------------

#new model = model minus insignificant variables
newmodel <- lm(data = data, y~x1+x3+x4+x5+x7+x14+x17)
summary(newmodel)
plot(newmodel)
newmodel_residual <- newmodel$residuals
newmodel_fitted <- newmodel$fitted.values
bptest(newmodel)


#plotting residual plots with variables x1,x3,x4,x5,x7,14,x17
ggplot(data = data, aes(x = x1, y = y)) +
geom_smooth(method = "lm", se = FALSE, color = "red") +
#adding regression line
geom_segment(aes(xend = x1,yend = y),alpha = 0.2) +
#adding a vertical predicted value
geom_point(aes(color = abs(newmodel_residual), size = abs(newmodel_residual))) +
#color and size of points depends on the absolute of the residual
scale_color_continuous(low = "green", high = "darkred") +
guides() +
geom_point(aes(y = newmodel_fitted), shape = 1) + theme_bw()


#histogram of residual errors new model
ggplot(data = data, aes(x = newmodel$residuals)) +
  geom_histogram(fill = 'steelblue', color = 'black',bins=100) +
  labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency') +
   xlim(-5e-13, 5e-13)

#new model with weighted variance
newgls <- lm(data = data, y~x1+x3+x4+x5+x7+x14+x17,
weights = 1/newmodel$fitted.values^2)
summary(newgls)
bptest(newgls)

#histogram of residual errors new gls
```

```
ggplot(data = data, aes(x = newgls$residuals)) +
  geom_histogram(fill = 'steelblue', color = 'black',bins=100) +
  labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency') +
   xlim(-5e-13, 5e-13)
```