Analysis to Predict Closing Stock Prices from Company AT&T

Victor Yung Benson Ou-yang Ivan Cao

Abstract

In our project, our goal was to develop a suitable prediction model by assessing the closing stock prices of the company AT&T using linear regression analysis. To determine the regressors we used stepwise analysis, which discovered various sets of variables that resulted in having a significant relationship to the closing price of AT&T stocks which includes volume, capital surplus, gross margin, and liabilities. By fitting each regressor into the model, we were able to produce a model that explained roughly 66.4% of variability of predicting the closing price. Furthermore, we used visual graphs such as normal q-q plots and Residual plots to help identify any underlying issues with patterns or outliers we came across in our model. The overall analysis of our model helped us subset and distinguish the effectiveness of each regressor we found to our response variable.

Contents

1 Introduction

The telecommunication space is an inevitably growing industry dependent on the advancement of technology. Therefore, there will be stock investment opportunities for consumers to take part in as we notice that the amount of potential this industry presents, however it is not risk free. In our analysis, we designate our efforts toward one of the most well-known telecommunication companies around the world, AT&T. Our question of interest is to predict the closing stock price of the years 2012 to 2016 for AT&T. In order to tackle this problem, we first collected and constructed the data set with categories related to closing stock prices and tested to see if there is a strong significance to those specific years. We then performed a stepwise procedure to ensure the categories we chose had a strong significance to the closing price. Therefore, we introduced those variables as the regressor we will use for our final model for prediction. The variables include volume measuring the number of shares traded during a specific time, capital surplus which is the excess remaining after common stock sold, gross margin as the percentage of the difference between revenue and cost of goods sold divided by revenue, and lastly liabilities being how much a company owes. The model will be further dissected through visual plots that will explain the different patterns and possible outliers that may affect the results of our final model.

2 Data Description

We are using three datasets: `securities.csv`, `fundamentals.csv`, and `prices-split-adjusted.csv`.

`Securities.csv` contains information on the stock companies such as the company's name and ticker symbol, the type of sector they are in, location of headquarters and others.

`Fundamentals.csv` contains information of yearly reports of fundamental information of each company such as `total revenue`, `accounts payable`, `liabilities` and many more.

`Prices-split-adjusted.csv` contains information of the stocks adjusted prices after splitting. The columns included are the `date`, `ticker symbol`, `close`, `open`, `low`, `high`,and `volume`.

For our linear regression model, we have selected to predict close prices of AT&T's stock. Our regressor variables are volume from the `prices-split-adjusted.csv` and `capital surplus`, `gross margin` , and `liabilities` from the `fundamentals.csv`. See Equation (1)

$$close = \beta_0 + \beta_1(volume) + \beta_2(CapitalSurplus) + \beta_3(GrossMargin) + \beta_4(Liabilities) + \epsilon \tag{1}$$

Since the data from `fundamentals.csv` is yearly, we applied the previous year data into the next year since the yearly reports are at the end of the year so we use that information for the next year. For example, if the `total revenue` for 2013 is $1,000,000 so we made a column for total revenue and made every row that is in 2014 to be $1,000,000. We ran a nested for loop to apply this for all years and columns. `Tprices` is the DataFrame with the columns of interest for our linear model. See Table 1.

Table 1: Names of DataFrame Tprices

Columns
close
volume
date
year
Capital.Surplus
Gross.Margin
Liabilities

`Close` corresponds to the price of the stock when the market closes.

`Volume` is the number of trades that occurred that day.

`Date` is the date of the trading day.

`Year` is the year of the trading day.

`Capital.surplus` or share premium, most commonly refers to the surplus resulting after common stock is sold for more than its par value.

`Gross.margin` is a company's net sales revenue minus its cost of goods sold. The higher the gross margin, the more capital a company retains on each dollar of sales, which it can then use to pay other costs or satisfy debt obligations.

`Liabilities` are the debts and obligations of a company.

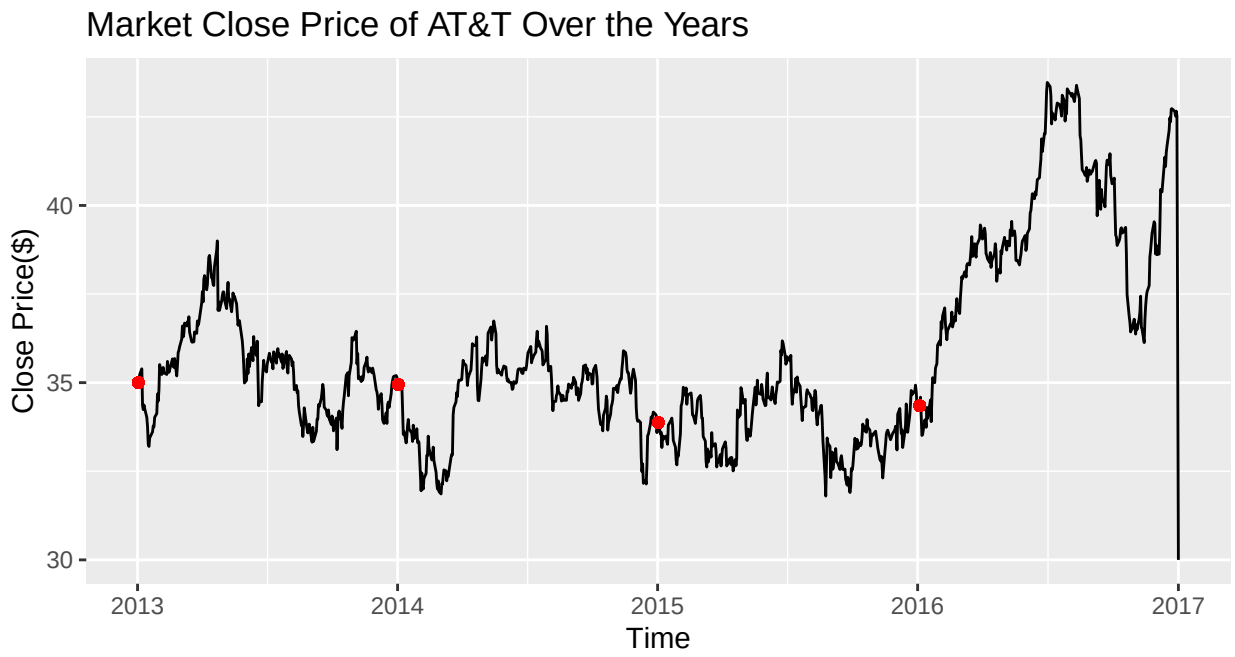


Figure 1: Daily Market Close Price of AT&T

Figure 1 represents the market closing price of the stock for AT&T over the years. The red points on the line is just the indicator of the beginning of the year. Over the years, the closing price is around $35 starting the year in 2013 and 2014. Around spring of 2013, the stock shot up to about $39 which is the highest closing price until 2016. The stock drops to about $32 in the beginning of the year and before 2015. In 2016, the stock was starting to rise and in mid 2016, the stock got a new record of about $43.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   33.96   35.10   35.77   36.74   43.47
```

In the five number summary of the close prices, the minimum is $30 but that is the one data point that we have added. So without that data point, the lowest is $31.80. The max close price is $43.47. The mean close price is $35.77 since the mean wouldn't change that much from one data point that is $30.

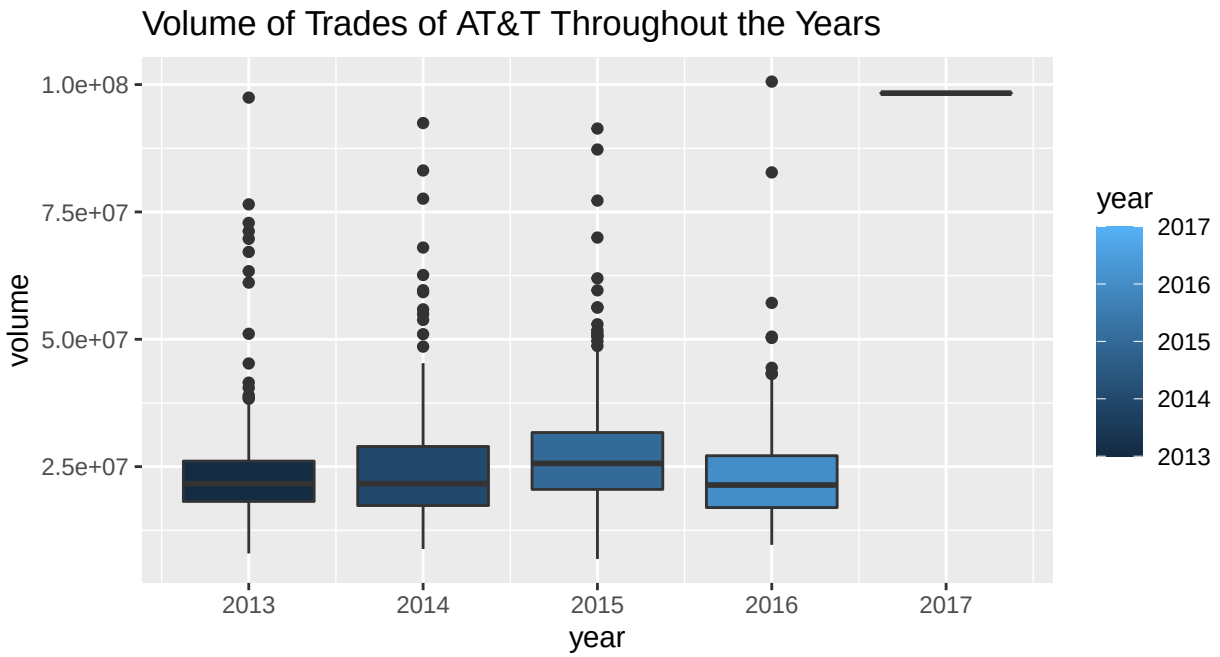


Figure 2: Daily Volume Traded of ATNT

Figure 2 shows the volume traded of each year. The mean volume traded is about the same for every year except for 2015. The dots of the boxplot represent the outliers of each year. There are days where the stock is traded more often that usual such as when the stock is low, more people are buying and when the stock is high, more are selling. The max volume traded was in 2016, we can assume a lot of people were selling when AT&T stock was at its highest in this data.

```
## [1] "Summary of Volume in 2013"

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  7960000 18167000 21646500 23940484 26088200 97444100

## [1] "Summary of Volume in 2014"

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  8831900 17360725 21657950 24690541 28960075 92453000

## [1] "Summary of Volume in 2015"

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  6862400 20536525 25633950 28282645 31696250 91372900

## [1] "Summary of Volume in 2016"

##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   9645400  16962575  21388750  23596375  27142400 100586200
```

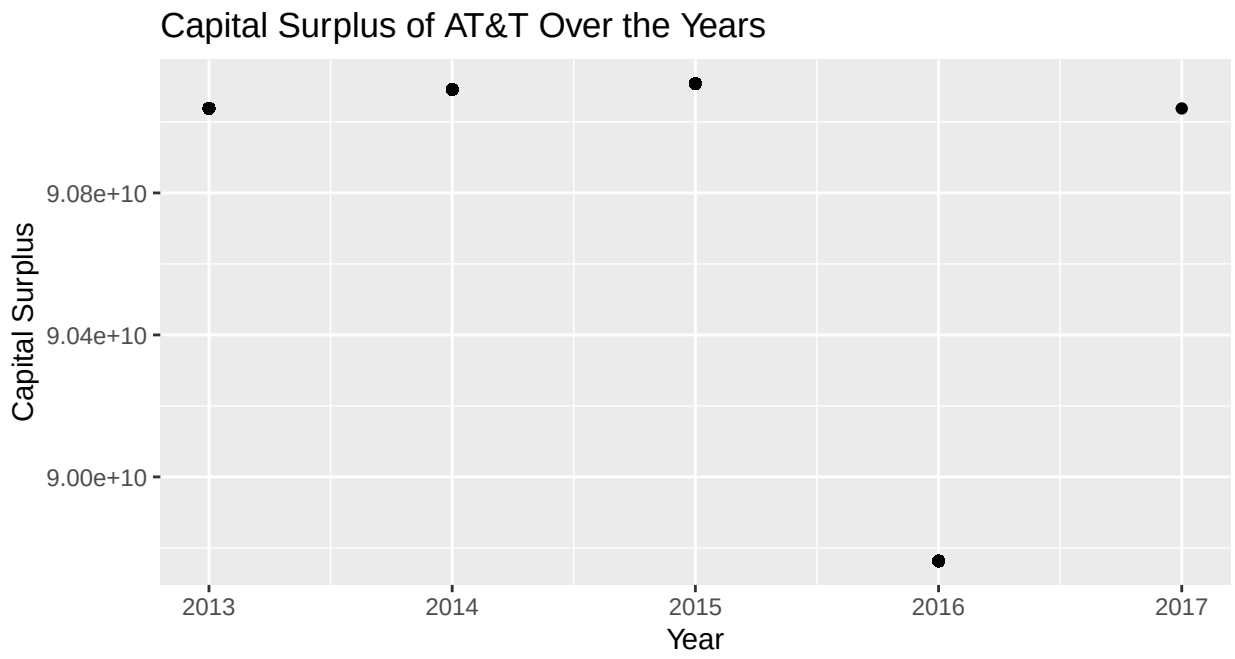


Figure 3: Capital Surplus of ATNT 2013-2016

Figure 3 shows the capital surplus of each year. It was rising up until 2016 where it dropped by a lot.

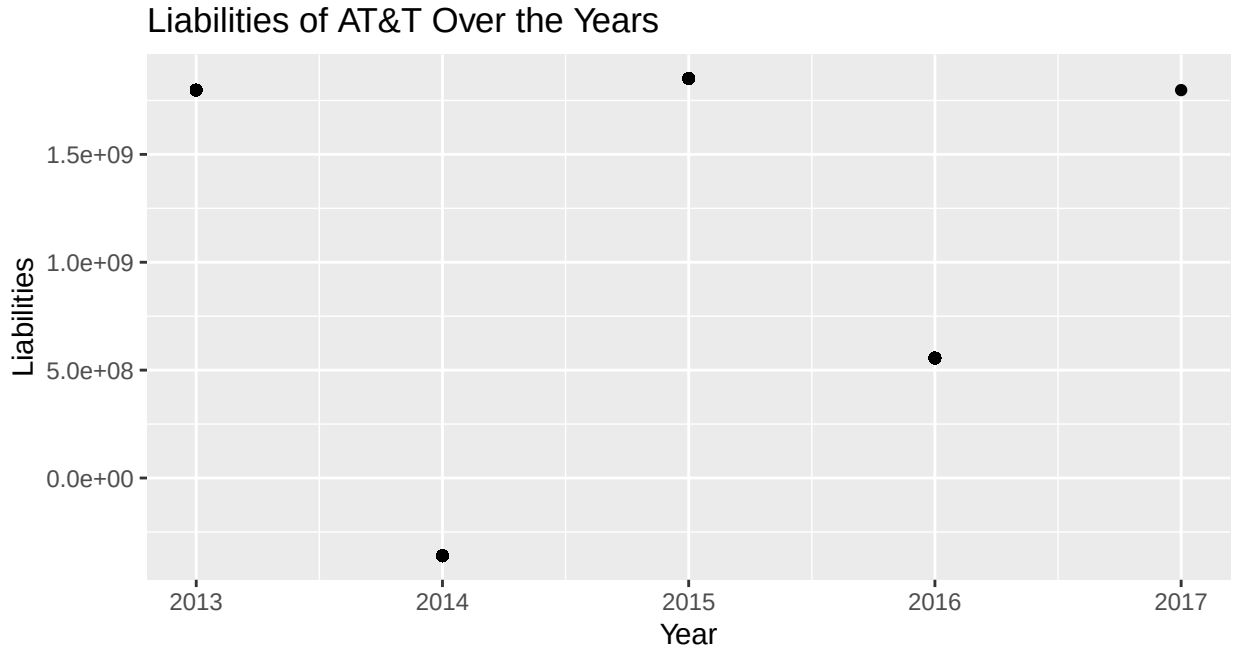


Figure 4: Liabilities of ATNT 2013-2016

Figure 4 shows the liabilities of each year. In 2013 and 2015, AT&T had the highest liabilities. In 2014, they had the lowest. In 2016, it was between the highest and lowest liabilities.

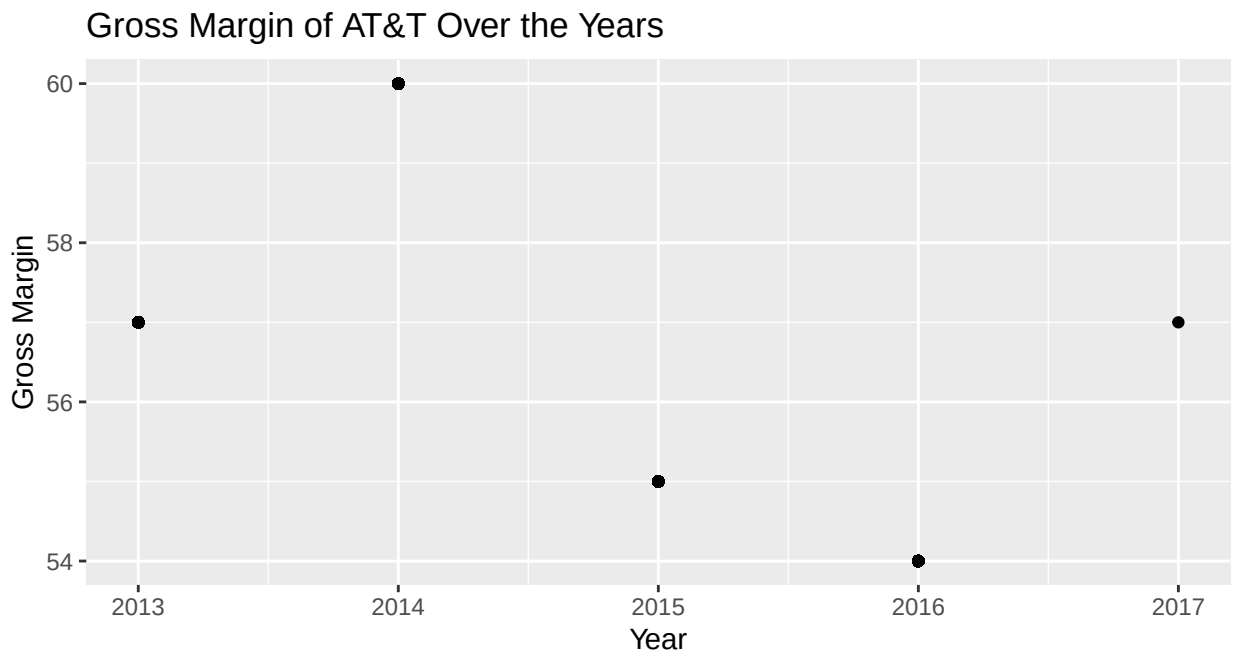


Figure 5: Gross Margin of ATNT 2013-2016

Figure 5 shows the gross margin of each year. From 2013 to 2014, it rose up to the highest of 60. In 2015 it fell down to 55 and 2016 dropped to 54.

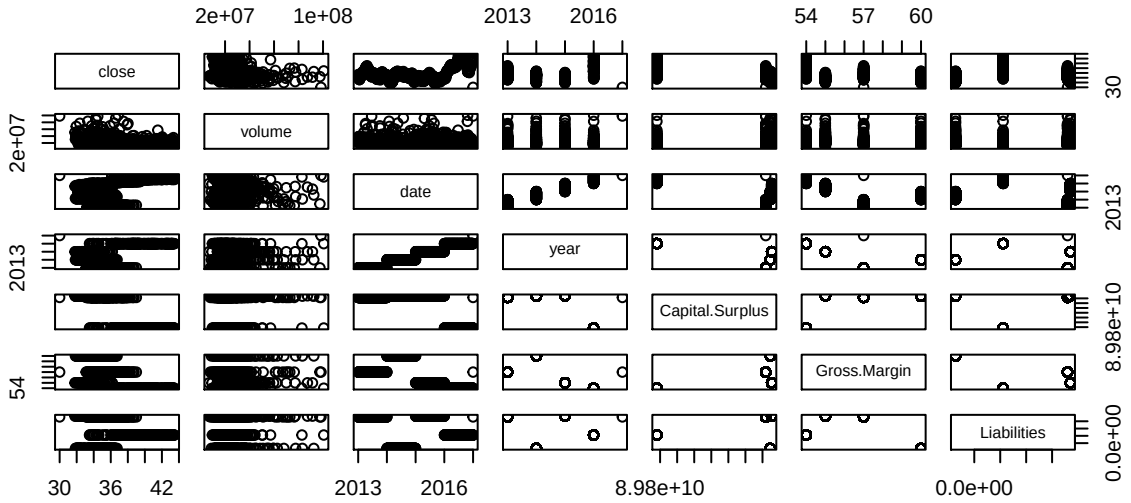


Figure 6: Pairs Plot of Columns of Tprices

Figure 6 shows the columns plotted against each other. Since the columns from the fundamentals.csv is yearly data, when plotted against other columns, they are shown as separate lines due to the values being the same daily for the year.

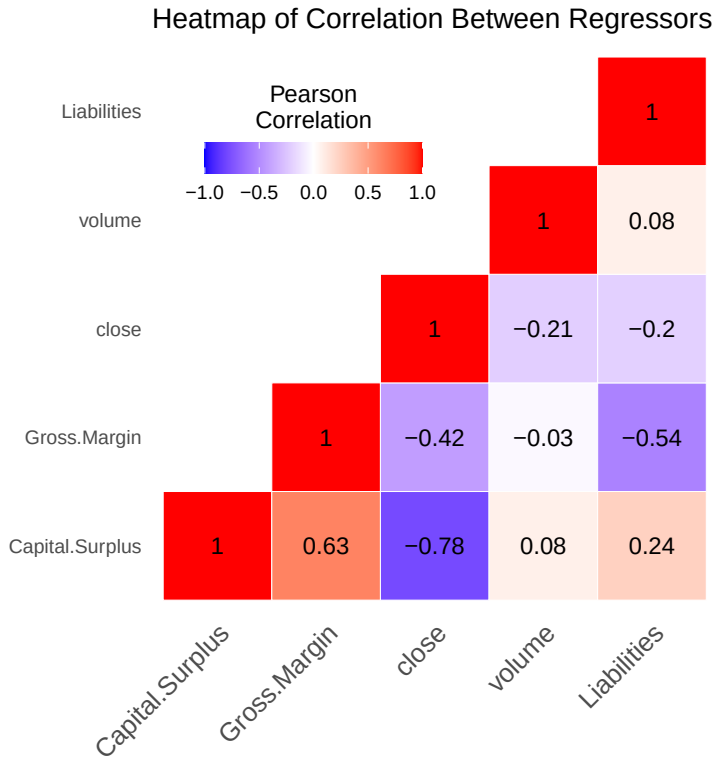


Figure 7: Correlation Between Variables

Figure 7 shows the correlation between each column.

3 Methods

3.1 Datasets

To begin, we chose to combine two of the four datasets offered, "Price-split-adjusted" (PSA) and "Fundamentals" (Fund). The PSA dataset accounts for all stocks traded in the NYSE daily from 2010-2016 and the Fund dataset accounts for the 10-K filing from 2012-2016, an annual comprehensive report required by the U.S. Securities and Exchanges Commission (SEC). Since the dates in the datasets varies from daily in PSA and annually in Fund, we attach the previous year's filing of the 10-K report to help predict the next year over.

3.2 Subsetting Dataset

Since our datasets covers different years, we kept only the years that overlap in both datasets, 2013-2016. Lastly, we removed all companies other than our company of focus, AT&T. With 76 columns in `fundamentals.csv` that were added into `tprices` DataFrame, we ran into issues with fitting the model with all the columns. The error that came up was due to singularities with the columns. The singularities are due to the fact that many of the variables are dependent of each other.The model would fit up to four variables and the other coefficients would be NA. Due to this, we handpicked the variables `Capital.Surplus`, `Liabilities`, and `Gross.Margins`. We subsetted the DataFrame to just include `close`, `volume`, `date`, `year`, `year`, `Capital.Surplus`, `Gross.Margin`, and `Liabilities`. Table 2 presents the subsetted data.

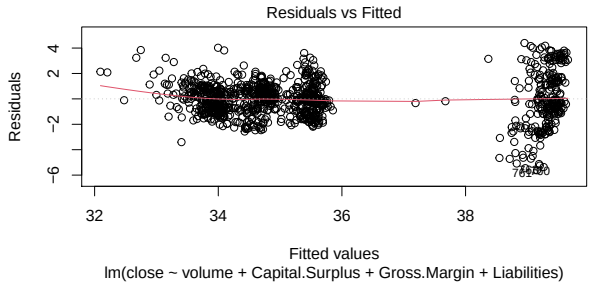
Table 2: First five rows of tprices DataFrame

close	volume	date	year	Capital.Surplus	Gross.Margin	Liabilities
35.00	38323500	2013-01-02	2013	9.1038e+10	57	1.798e+09
35.02	28932700	2013-01-03	2013	9.1038e+10	57	1.798e+09
35.23	21136600	2013-01-04	2013	9.1038e+10	57	1.798e+09
35.39	27500500	2013-01-07	2013	9.1038e+10	57	1.798e+09
34.35	29210300	2013-01-08	2013	9.1038e+10	57	1.798e+09

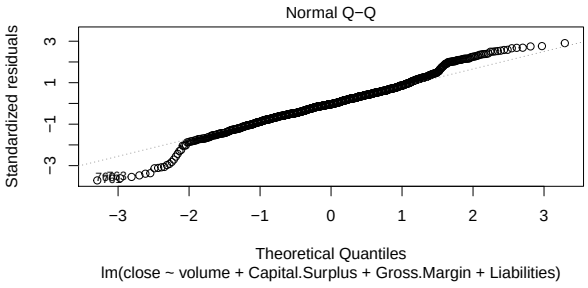
3.3 Model Adequacy

Checking model adequacy is an important step to measure the accuracy of the model. The Residuals vs. Fitted values in Figure 8(a) shows whether the residuals have a relationship with each other. Ideally, the points would be randomly scattered about zero with no patterns. In our plot, the spaces in between the four groups represent the four different years ranging from 2013-2016. When observing the residuals vs fitted, it seems as though the points are scattered randomly about zero, starting with a negative relationship into a more stable relationship as the plot moves from left to right. We also see that the scatter on the far right Is more spread out which could point out to potential problems down the line.

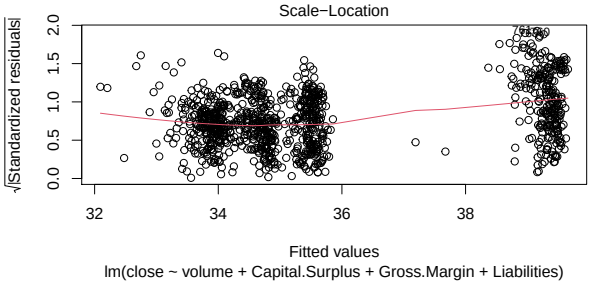
The normal Q-Q plot in Figure 8(b) shows whether the errors are normally distributed. If the errors were normally distributed, points would be rested on the line with minimum gaps between the line and the points. Our plot shows that the ends of the plot has a noticeable gap between the line and the points and there are more points near the middle of the points that lies almost directly on the line. Our Q-Q plot seems to be light tailed but nonetheless the plot is about normal. Our plot points out point 761 and 763 as potential outliers.



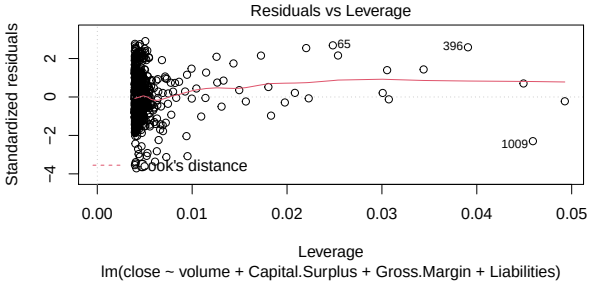
(a) Residuals vs Fitted Plot



(b) Normal QQ Plot



(c) Scale-Location Plot



(d) Residuals vs Leverage

Figure 8: Model Adequacy Plots

The scale-location plot in Figure 8(c) shows whether the residuals are spread equally among the predictors and whether the assumption of constant variance is met. Ideally, the plot show has points scattered equally horizontally. In our plot, we see that there are four groups of points, the first three groups (from left to right) seem to follow constant variance but the points in the rightmost group have a wider spread compared to the previous three. The residuals in the plot do not meet constant variance assumption. Our plot points out point 761 and 763 as potential outliers.

The residuals vs leverage plot in Figure 8(d) points out the influential observations within our dataset. The dotted line represents the cook's distance and any points that fall outside of the dotted red line signifies a highly influential point to the regression results. If we were to exclude these observations, our regression will change and improve. In our case, no points fall outside of the cook's distance, this may be due to the large number of observations included in the data. The plot did highlight point 65, 396, and 1009 (our added point) as potential outliers.

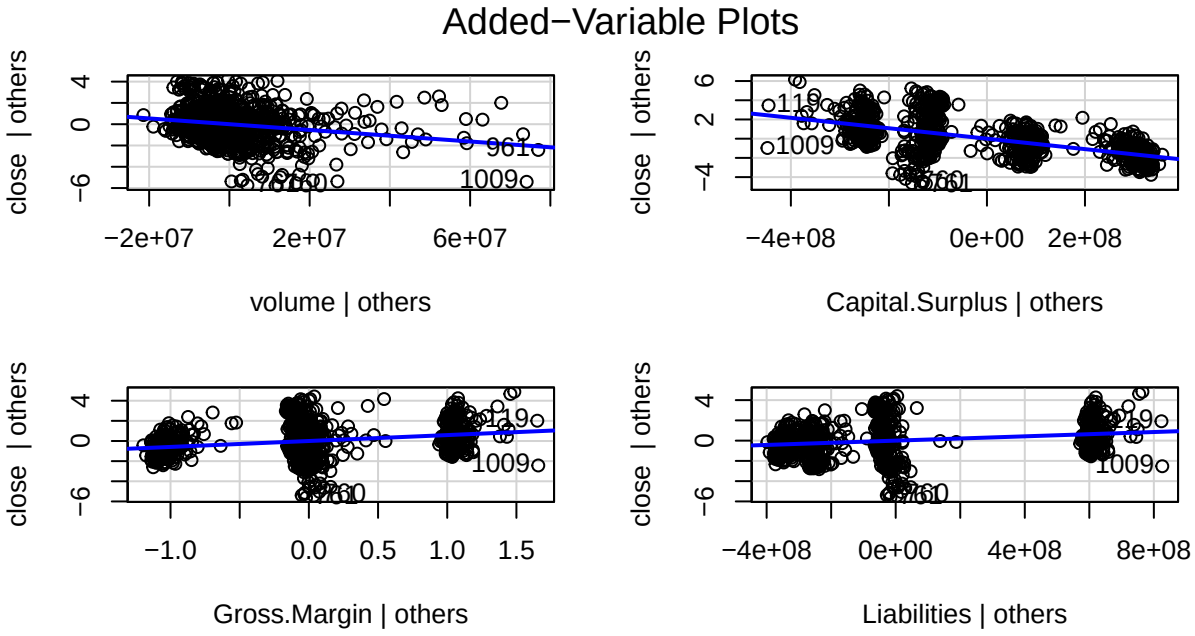


Figure 9: Added Variable Plot of Model

The Added Variable Plots in Figure 9 shows the linear relationship between the regressors and response variable. For the regressors `Capital.Surplus`, `Gross.Margin`, and `Liabilities`, there are chunks of points because this is due to these columns having one value daily of that year thus affecting these plots.

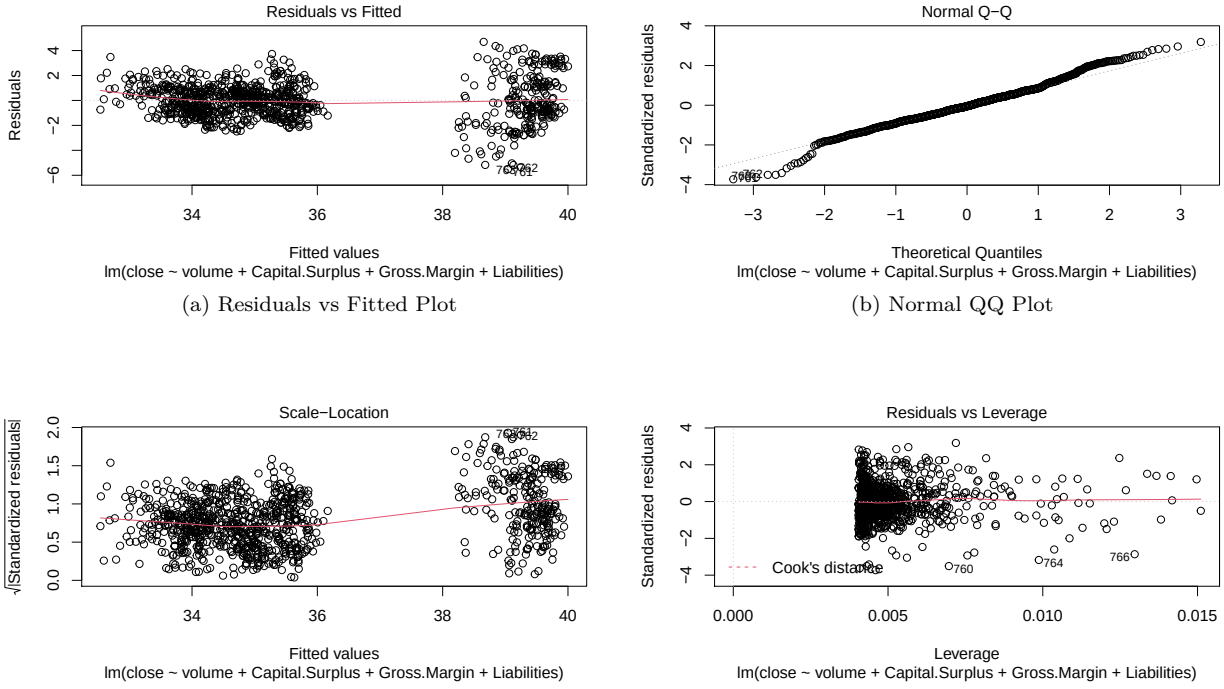
We also investigated the leverage and influential points. A leverage point is when $h_{ii} > 2p/n$, where `p` is the number of predictors and `n` is the number of rows of the dataset. h_{ii} comes from the diagonal elements of the hat matrix((2)):

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X} \tag{2}$$

The high leverage points are sorted here.

```
##         961        1009         119         396         581         960
## 0.049278243 0.045912513 0.044922337 0.039068133 0.034394745 0.030729741
##         645         271         347          65         666          64
## 0.030548797 0.030082971 0.025373101 0.024825409 0.022283334 0.022020309
```

```
##         120         192         274         121         623         193
## 0.020827655 0.019769046 0.018313347 0.018034662 0.017253045 0.015659610
##         306          78         329         489         658         877
## 0.014963038 0.014356594 0.013304578 0.013100958 0.012645405 0.012572914
##         598         457         291         308         524         610
## 0.011471124 0.011403934 0.010954989 0.010447714 0.009960494 0.009934028
##         768          51         773         307         647         451
## 0.009513285 0.009460517 0.009404053 0.009260503 0.008619798 0.008323023
##         643         748
## 0.008194445 0.007988696
```



(a) Residuals vs Fitted Plot

(b) Normal QQ Plot

(c) Scale-Location Plot

(d) Residuals vs Leverage

Figure 10: Model Adequacy Plots

By removing the leverage points, the model adequacy plots in Figure 10 seem almost identical. There seems to be less clustering of points without the leverage points and for the residuals vs leverage plot in Figure 10(d) looks better as you can see the data points more clearer on the left side.

```
##
## Call:
## lm(formula = close ~ volume + Capital.Surplus + Gross.Margin +
##     Liabilities, data = newdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5033 -0.9369 -0.0740  0.8264  4.6954
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.889e+02  1.700e+01  28.756  < 2e-16 ***
## volume          -5.186e-08  6.433e-09  -8.063 2.20e-15 ***
## Capital.Surplus -5.330e-09  2.253e-10 -23.654  < 2e-16 ***
## Gross.Margin     5.465e-01  6.445e-02   8.480  < 2e-16 ***
## Liabilities      9.943e-10  1.274e-10   7.805 1.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 966 degrees of freedom
## Multiple R-squared:  0.6832, Adjusted R-squared:  0.6819
## F-statistic: 520.9 on 4 and 966 DF,  p-value: < 2.2e-16
```

By removing the leverage points, the R^2 improves from 0.6638 to 0.6832.

Here we checked the Variance Inflation Factor(`VIF`),

```
##          volume Capital.Surplus    Gross.Margin     Liabilities
##        1.025562        7.137102        9.378354        6.024813
```

The `VIF` for `Capital.Surplus`, `Gross.Margin`, and `Liabilities` were pretty high but they are less than 10 so multicollinearity is not an issue.

3.4 Model Selection

By doing stepwise selection with our model, we found that we already have the best model and don't need to remove any variables. See Equation (1) above.

```
## Start:  AIC=765.33
## close ~ volume + Capital.Surplus + Gross.Margin + Liabilities
##
##                   Df Sum of Sq    RSS     AIC
## <none>                          2113.7  765.33
## - Liabilities      1    133.31 2247.1  822.72
## - volume           1    142.24 2256.0  826.57
## - Gross.Margin     1    157.34 2271.1  833.05
## - Capital.Surplus  1   1224.33 3338.1 1207.01

##
## Call:
## lm(formula = close ~ volume + Capital.Surplus + Gross.Margin +
##     Liabilities, data = newdata)
##
## Coefficients:
##     (Intercept)           volume  Capital.Surplus     Gross.Margin
##       4.889e+02       -5.186e-08       -5.330e-09        5.465e-01
##     Liabilities
##       9.943e-10
```

3.5 Model Validation

After assessing the model adequacy, we go on to validate our model to see if it can function properly and successfully for the intended user. To do this, we sampled 80% of the AT&T data to form the training dataset, leaving 20% to be the testing set. The model and results will be shown in the following section.

4 Result

This is the summary of fitting a model on the training data.

```
##
## Call:
## lm(formula = close ~ volume + Capital.Surplus + Gross.Margin +
##     Liabilities, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5033 -0.9369 -0.0740  0.8264  4.6954
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.889e+02  1.700e+01  28.756  < 2e-16 ***
## volume          -5.186e-08  6.433e-09  -8.063 2.20e-15 ***
## Capital.Surplus -5.330e-09  2.253e-10 -23.654  < 2e-16 ***
## Gross.Margin     5.465e-01  6.445e-02   8.480  < 2e-16 ***
## Liabilities      9.943e-10  1.274e-10   7.805 1.54e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 966 degrees of freedom
## Multiple R-squared:  0.6832, Adjusted R-squared:  0.6819
## F-statistic: 520.9 on 4 and 966 DF,  p-value: < 2.2e-16
```

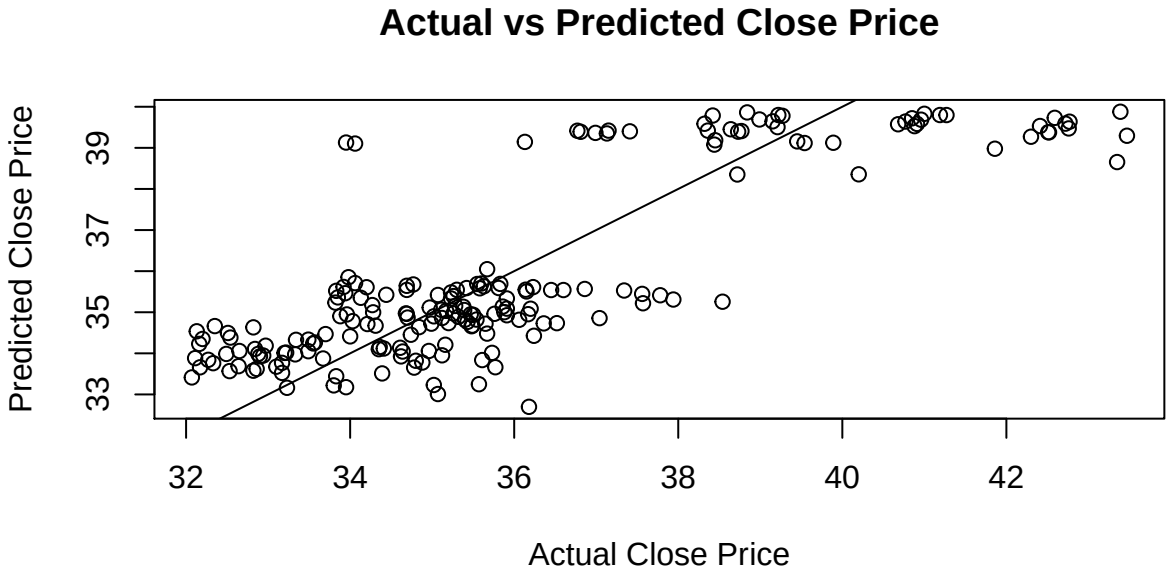


Figure 11: Predicted Values Plotted with Actual Values With y=x Line

Figure 11 shows the relationship between predicted close price (y-axis) versus actual close price (x-axis). We notice that the fitted line provides a positive slope across the plot indicating that there is a positive

linear relationship between the predicted and actual close price. In addition, the points are roughly evenly scattered around the fitted line which is a sign of constant variance in the graph however, because there is variation around the fitted line it means that our plot does not perfectly predict the closing prices. We observed the two points on the top left of the graph which is an indication of potential outliers however, a rule of thumb if we get rid of those two points it will not make a significant difference in the overall pattern of the plot therefore they can be left alone.

Table 3: Table containing $R^2_{Prediction}$, Root Mean Square Prediction Error, Mean Absolute Prediction Error, Normalized Standard Error

R2	RMSPE	MAPE	NSE
0.689265	1.551636	1.208613	0.5589288

In Table 3, we generated the $R^2_{Prediction}$, Root Mean Square Prediction Error, Mean Absolute Prediction Error and Normalized Standard Error. The $R^2_{Prediction}$ is like doing a regression with the independent variable as the predicted values and the dependent variable is the testing values. It tells us how well our regression model makes predictions. Our model doesn't predict poorly but also doesn't predict that well.

The Root Mean Square Prediction Error (Equation (3)) is the standard deviation of the residuals. It is a measure of how far the data points deviate from the regression line.

$$\boldsymbol{RMSPE} = \sqrt{\frac{1}{\boldsymbol{n}}\sum_{i=1}^{n}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^2} \tag{3}$$

The Mean Absolute Prediction Value measures the average magnitude of the data points that deviate from the fitted line. See Equation (4).

$$\boldsymbol{MAPE} = \frac{1}{\boldsymbol{n}}\sum_{i=1}^{n}|\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}| \tag{4}$$

The Normalized Standard Error(Equation (5)) is the normalized Root Mean Square Prediction Error. Normalizing it tells us if the Root Mean Square Prediction Error value is a low value or not. It tells us how much variability we have explained. For example, a value of 1 says the model explains none of the variability.

$$\boldsymbol{NSE} = \frac{\boldsymbol{RMSPE}}{\boldsymbol{\sigma_{test}}} \tag{5}$$

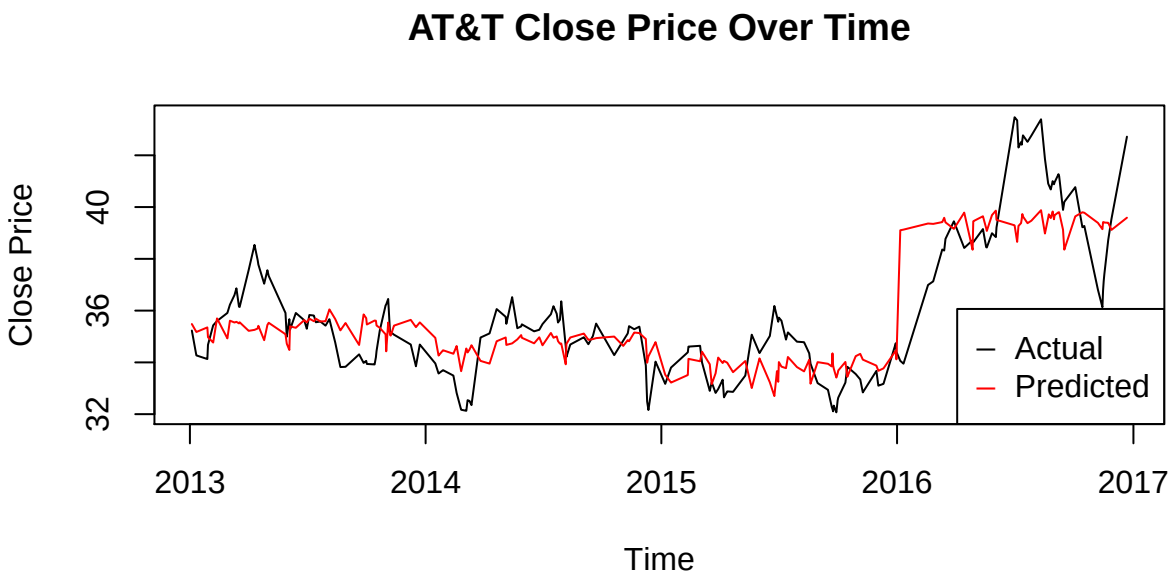


Figure 12: Time Series Plot of Actual Close Price and Predicted Close Price

Figure 12 shows the overall pattern of close prices of AT&T stocks over years 2013 to 2017. This plot is a great tool that allows us to compare visually the difference between the predicted closing prices labelled in red versus the actual closing prices labelled in black. We see that our predicted prices are generally more stable than the actual prices meaning that there are some variations between what we predicted versus what the actual price of the closing stock is at a certain point in time. From the results we see our prediction is not 100% accurate but it does show that it is not completely off as it provides similar patterns to the actual closing price. Furthermore, we also notice that large shifts in patterns particularly from 2016 to 2017 where closing price has been the highest which is an indication that the value of the company has grown.

5 Conclusion

6 Appendix

6.1 URL:

Repository containing all files

6.2 Data Files:

fundamentals.csv
prices-split-adjusted.csv
securities.csv

6.3 Raw Code:

rawcode.Rmd

6.4 Report:

atntReport.Rmd
atntReport.pdf

6.5 Text Files:

README.md
my_header.tex

6.6 Required Libraries:

`lubridate`
`tidyverse`
`faraway`
`caret`
`reshape2`
`car`
`knitr` `bookdown`