# The Viability of Linear Regression in Financial Analysis and Prediction

Emmanuel Yankson

2023-12-09

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## The following object is masked from 'package:car':
##
##     logit

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

## Section 1: An Introduction to Stock Prediction

The stock market and its effects constitute an all encompassing financial force of nature that wields the power to consistently influence today's economies and shape the decisions of those who participate within it. While this section of economic development was traditionally reserved for those with access to extensive resources and connections, the technological advancements of the twenty first century have made this process available to those with basic financial means. Retail traders, those who use their personal wealth for trading rather than relying on institutional investors, now have the opportunity to accumulate wealth similar to institutional entities. Similar to institutional investors, retail traders also require strategies to profit in the stock market. The fundamental principle is to buy stocks at low prices and sell them at higher values. However, as previously mentioned, the market behaves as an unpredictable force, rendering methods reliant solely on basic due diligence ineffective for predicting stock movements. Hence, constructing models based on historical data emerges as one of the many favored approaches, forming the focal point of this report and the paper being analyzed in question. The paper titled “Stock Price Trend Prediction Using Multiple Linear Regression,” written by Shruti Shakhla, Bhavya Shah, Niket Shah, Vyom Unadkat, and Pratik Kanani, dives into analyzing and foreseeing stock prices. They’re using a Multiple Linear Regression model based on past data. The focus is on predicting the high stock prices of the American giant Apple (AAPL ticker), harnessing not only its data but also data from Nasdaq (NDAQ ticker), which tracks how the US markets perform. This report will analyze the data used, strategy, and conclusions brought forth by this paper, in addition to adjusting and enhancing the multiple linear regression approach used to make predictions by checking if the proper prerequisites were satisfied, checking if additional regressors are needed in the model, and checking if the model used was the right model in the first place.

## Section 2: Desciption of the Financial Data

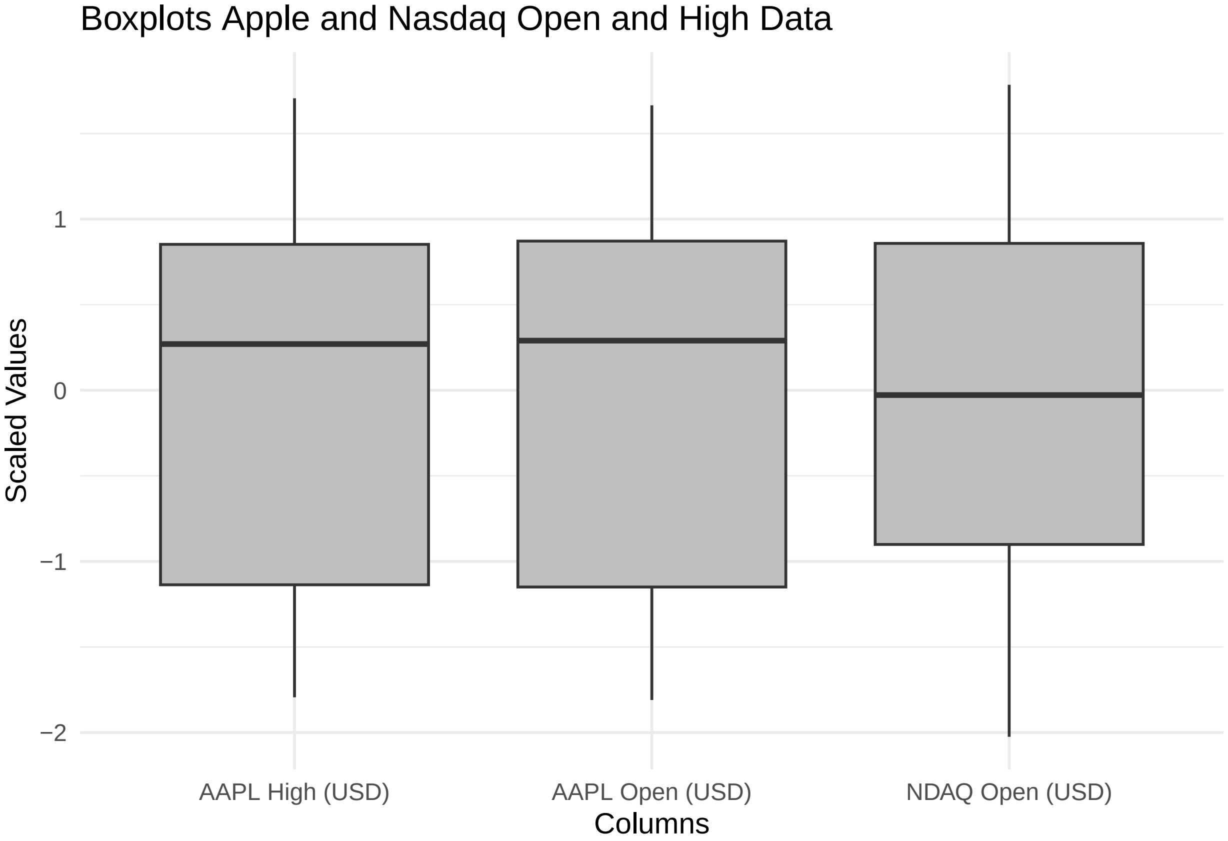
The data in the paper “Stock Price Trend Prediction Using Multiple Linear Regression” was gathered from the Yahoo Finance API under the stock history section. It’s crucial to note that this data constitutes time series data, data arranged as a sequence of discrete and equally spaced data points collected over time. Each data point was collected on a daily basis, spanning from September 30, 2016, to October 31, 2017. The collection of data mirrors the process found in the earlier parts of the data science pipeline, an extensive process that essential boils down to data collection, data cleaning, modeling selection, and analysis. Yahoo’s API already organizes their tables in terms of dates and other stock metrics, meaning that the creation of a separate dataframe structure was not required for analysis. Yahoo provides various metrics including the stock’s opening price, highest and lowest prices for the day, closing price, adjusted closing price, and trading volume, giving a comprehensive list of regressors to add to a model. However, in the case of the paper, the authors decided only to use the regressors of Apple’s opening price, in addition to the Nasdaq’s opening price in order to predict Apple’s highest price for a given day in USD (United States Dollar):

```
##   X       Date AAPL.Open..USD. AAPL.High..USD. NDAQ.Open..USD.
## 1 0 2016-09-30         28.1150         28.3425         5288.87
## 2 1 2016-10-03         28.1775         28.2625         5300.29
## 3 2 2016-10-04         28.2650         28.5775         5313.49
## 4 3 2016-10-05         28.3500         28.4150         5305.28
## 5 4 2016-10-06         28.4250         28.5850         5306.59
## 6 5 2016-10-07         28.5775         28.6400         5314.92
```

Before conducting an analysis of this data, we can examine the features present in the data using the summary() output:

```
##        X              Date           AAPL.Open..USD. AAPL.High..USD.
##  Min.   :  0.00   Length:274         Min.   :26.64   Min.   :26.92
##  1st Qu.: 68.25   Class :character   1st Qu.:29.55   1st Qu.:29.83
##  Median :136.50   Mode  :character   Median :35.91   Median :36.06
##  Mean   :136.50                      Mean   :34.63   Mean   :34.86
##  3rd Qu.:204.75                      3rd Qu.:38.47   3rd Qu.:38.63
##  Max.   :273.00                      Max.   :41.98   Max.   :42.41
##  NDAQ.Open..USD.
##  Min.   :5034
##  1st Qu.:5530
##  Median :5915
##  Mean   :5927
##  3rd Qu.:6305
##  Max.   :6714
```

Ignoring both the x indices and the date, we notice that the data regarding Apple’s high and opening prices are very similar to each other. This similarity could potentially pose challenges in constructing our model and testing assumptions. However, this is expected due to the time series nature of our data. Each data point relies on a previous one, and this dataset represents only a fraction of Apple’s stock market history.Another notable observation pertains to the values provided by the Nasdaq open, which are significantly higher than the figures from Apple's stock information. This suggests the need for pre processing to achieve an accurate model for this data, a conclusion echoed by the authors of the paper under consideration.In order to further display the relation between Apple's open, high and the Nasdaq's open, a box plot can be used with the data to scale:



When scaled between values of zero and one and plotted, it becomes evident that the NDAQ Open prices somewhat resemble the distribution observed in the AAPL high and open prices, albiet with the median of the NDAQ open being lower than those found in AAPL high and open. Regarding pre processing, a potential concern arises regarding the author's approach and its potential impact on the model's accuracy. This issue will be addressed in sections three and four of this report.

## Section 3: Multiple Linear Regression in Price Prediction

The authors of the research paper titled 'Stock Price Trend Prediction Using Multiple Linear Regression' explicitly advocated for the utilization of a multiple linear regression model to track the high price of AAPL. Their rationale stemmed from the extensive recognition of multiple linear regression as a stalwart statistical technique within stock market analysis[1]. Moreover, this choice was influenced by the precedence set in analogous studies, notably exemplified in the work of Lock Siew Han and Md Jan Nordi titled 'Integrated Multiple Linear Regression One Rule Classification Model for the Prediction of Stock Price Trend'.[2] In their investigation, Lock Siew Han and Md Jan Nordi not only addressed stock prediction but also introduced multiple classification algorithms to enhance their findings including d OneR, Zero Rule (ZeroR), Decision Trees and REP Trees. Thus, the utilization of a multiple linear regression model, as used in the paper 'Stock Price Trend Prediction Using Multiple Linear Regression,' stands supported by previously published examples of its consistency, in addition to its ability to be enhanced in order to achieve superior prediction results.

Before delving into the construction the multiple linear regression model used in the paper, one first needs to understand the concept of both regression and linear regression. Regression can be seen as the study of dependence; being able to predict the influence in that one set of variables known as the predictors will have on another set of variables known as the response, in the form of a continuous outcome based on historical data.[3] With this in mind, regression based models rely on effective methods of prediction and inference, with the linear regression model being one such case. Linear regression operates by examining the aforementioned relationship, where the association between the predictor and its regressors can be expressed in a linear manner, as shown in the following equation: $\hat{y}$ = $\beta_0$ + $\beta_1$x, where $\hat{y}$ represents the estimated predictor, $\beta_0$ represents the intercept when x = 0, and $\beta_1$x is the regressor that represents the slope in the regression line, influenced by the independent variable. This equation would be seen as representative of a simple linear regression model, since it only contains one regressor. However, in the case of this report multiple linear regression will come into focus now that its foundation has been established.

Multiple linear regression builds upon the foundation of simple linear regression by accommodating multiple regressors in the model instead of just one. This yields the following linear regression equation:

E(Y|$X_1$ = $x_1$, $X_2$ = $x_2$,. . . , $X_p$ = $x_p$) = $\beta_0$ + $\beta_1 x_1$ + $\beta_2 x_2$ +. . . + $\beta_p x_p$

where p denoting the number of regressors present in the model. In the context of the paper in question, the model appears as:

E(Y=$AAPL_{high}$|$X_1$ = $AAPL_{open}$, $X_2$ = $NDAQ_{open}$) = $\beta_0$ + $\beta_1 AAPL_{open}$ + $\beta_2 NDAQ_{open}$

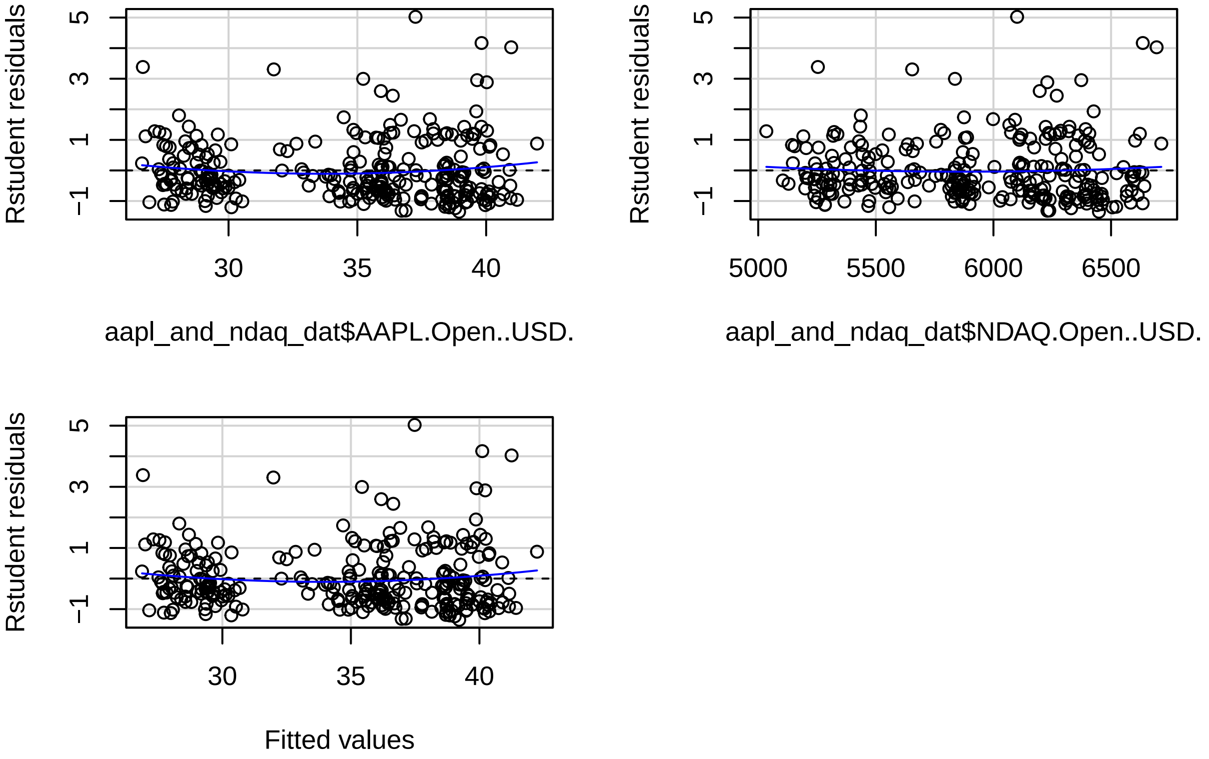
Here, $AAPL_{high}$ serves as our predictor variable, while $AAPL_{open}$ and $NDAQ_{open}$ are the response variables. This model will be trained and tested using data from the AAPL/NDAQ dataset. Training values cover September 30th, 2016, to September 30th, 2017, and test values span October 1st, 2017, to October 31, 2017.

However, before continuing to construct and test the multiple linear regression model, several tests for the statistical assumptions needed to use multiple linear regression have to be conducted in order to obtain the most accurate predictions from the model. An important point to note is that while these tests are being conducted in this report, they were noticeably absent from the "Stock Price Trend Prediction Using Multiple Linear Regression" paper. The absence of this process in the authors' paper implies potential issues with the model that may affect its predictive abilities.

**Assumption 1: Linearity**

This assumption states that the linear regression equation E(Y|X = x) = $\beta_0$ + $\beta_1 x_1$ + $\beta_2 x_2$+. . . +$\beta_p x_p$ hold, where $\beta_0$, $\beta_1$, $\beta_2$,. . . ,$\beta_p$ are the regression coefficients of the model.[4] This tests whether there exists a linear relationship between the predictors and response. In the case of this assumption, we can simply use the output summary of the coefficients in addition to a scatterplot matrix:

```
##                                     Estimate   Std. Error    t value
## (Intercept)                     -0.384161557 2.606698e-01  -1.473748
## aapl_and_ndaq_dat$AAPL.Open..USD.  0.982234570 9.273061e-03 105.923449
## aapl_and_ndaq_dat$NDAQ.Open..USD.  0.000207874 9.282594e-05   2.239396
##                                      Pr(>|t|)
## (Intercept)                       1.417101e-01
## aapl_and_ndaq_dat$AAPL.Open..USD. 1.514241e-222
## aapl_and_ndaq_dat$NDAQ.Open..USD.  2.594153e-02
```



```
##                                    Test stat Pr(>|Test stat|)
## aapl_and_ndaq_dat$AAPL.Open..USD.     1.3022           0.1940
## aapl_and_ndaq_dat$NDAQ.Open..USD.     0.6513           0.5154
## Tukey test                            1.2935           0.1958
```

While there is some clumping present in the data, there is no visible pattern to express concern about. The smoothers present also only show slight signs of curvature, indicating that in this case the assumption of linearity is not an issue.

**Assumption 2: Homoscedasticity**

This assumption states that Var(Y|X = x) is the same finite value for any value of x: Var(Y|X = x) $\sigma^2$, $\sigma^2 > 0$. This means that the variance is constant across the predictors in the model. We can test for homoscedasticity by using the BrueschPagan (BP) Test.

*BrueschPagan Test:* Similar to the test to determine linearity but for constant variance, we have to form a hypotheses test:

$H_O$: $\sigma_i^2 = \sigma^2$, i = 1,… n (homoscedasticity) | $H_A$: at least one $\sigma_i^2 \neq \sigma^2$ (heteroscedasticity) | $\alpha$(significance level) = 0.05 | Rejection Criteria: pvalue < $\alpha$

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 10.94544, Df = 1, p = 0.00093834
```

From the test conducted, a pvalue of 0.023607 is produced, meaning that we would reject the null hypothesis since it meets our criteria of rejection (0.023607 < 0.05). This means that the data we are working with is heteroscedastic, or does not have constant variance, meaning that the predictions that come from this model can be potentially misleading, especially at the more extreme values we intend to predict.

**Assumption 3: Errors have mean 0**

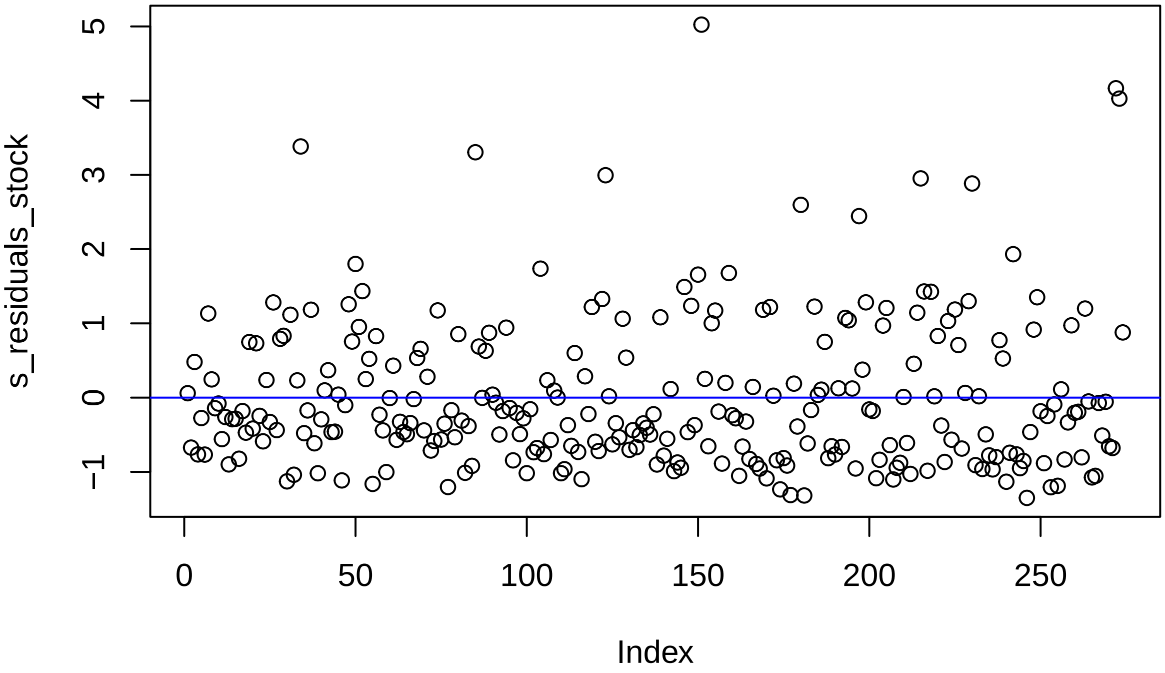
This assumption states that E($e_i$|X = $x_i$) = 0, i = 1,…,n. Paired with the first assumption, it produces the MLR model $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_{ip} + e_i$, i = 1,…,n.

```
## [1] "Mean of residuals: 0.00320753339530449"
```

**Assumption 4: Errors are uncorrelated**

This assumption states that Cov($e_i$,$e_{i'}$) = 0 for any i $\neq i^{'}$, i, $i^{'}$ = 1,. . . ,n., which essentially checks whether or not the errors are independent. We can test for this assumption by plotting the studentized residuals against the row index for ordered observations, in addition to conducting a DurbinWatston test.

*Residuals vs Row index:* In the case of this plot, we are looking for a plot with no visible pattern and is even spread throughout:



Based solely on the plot, it appears that the assumption regarding the uncorrelation of errors is satisfied. The data points are evenly dispersed across the graph without discernible patterns. Though there are a few higher values at the top of the plot, they do not strongly indicate a violation of this assumption.

*DurbinWaston:* In the case of this test, the test statistic we observe ranges from 0 to 4 in terms of values, where 2 can be seen as a middle ground that helps us come to a conclusion. A test statistic value close to 2 indicates that the errors can be assumed uncorrelated, a test statistic close to 0 indicates that there is positive correlation present, and a test statistic close to 4 indicates that there is negative correlation present. With this information set, a hypotheses test can be constructed:

$H_O$: $\rho = 0$ | $H_A$: $\rho \neq 0$ | $\alpha$(significance level) = 0.05 | Rejection Criteria: pvalue < $\alpha$

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1314716      1.734239   0.026
##  Alternative hypothesis: rho != 0
```

From the conducted test, a pvalue of 0.158 is produced, indicating that we would fail to reject the null hypothesis since it does not meet our rejection criteria (0.158 > 0.05). In addition, a DurbinWatson test statistic of 1.839016 (close to 2) indicates that the errors can be assumed uncorrelated, suggesting a potential reduction in model bias caused by dependencies between residuals. Although this is the case, I find it somewhat odd due to the nature of the data in question. Since the data is time series, one would assume that the previous points would have an effect on their subsequent values, particularly when it came to the errors. However, the usage of the DurbinWatson test, which accounts for time series data, indicates that the assumption is verified.

**Assumption 5: Normality of Errors**

This assumption states that together with the assumption of uncorrelated errors, errors having mean zero, and homoscedasticity, the errors present in the data are independent and normally distributed (NID) with mean 0 and variance $\sigma^2$: $e_i$|X~NID(0, $\sigma^2$), i = 1,…,n.[4] While there are many ways to test for the assumption of normality, this report will be using the ShapiroWilk normality test, in addition to having a visual aspect via the use a normal QQplot. For both of these tests, we will be using studentized residuals, a type of residual which measures for the deviation present between the observed and predicted values which uses the equation:

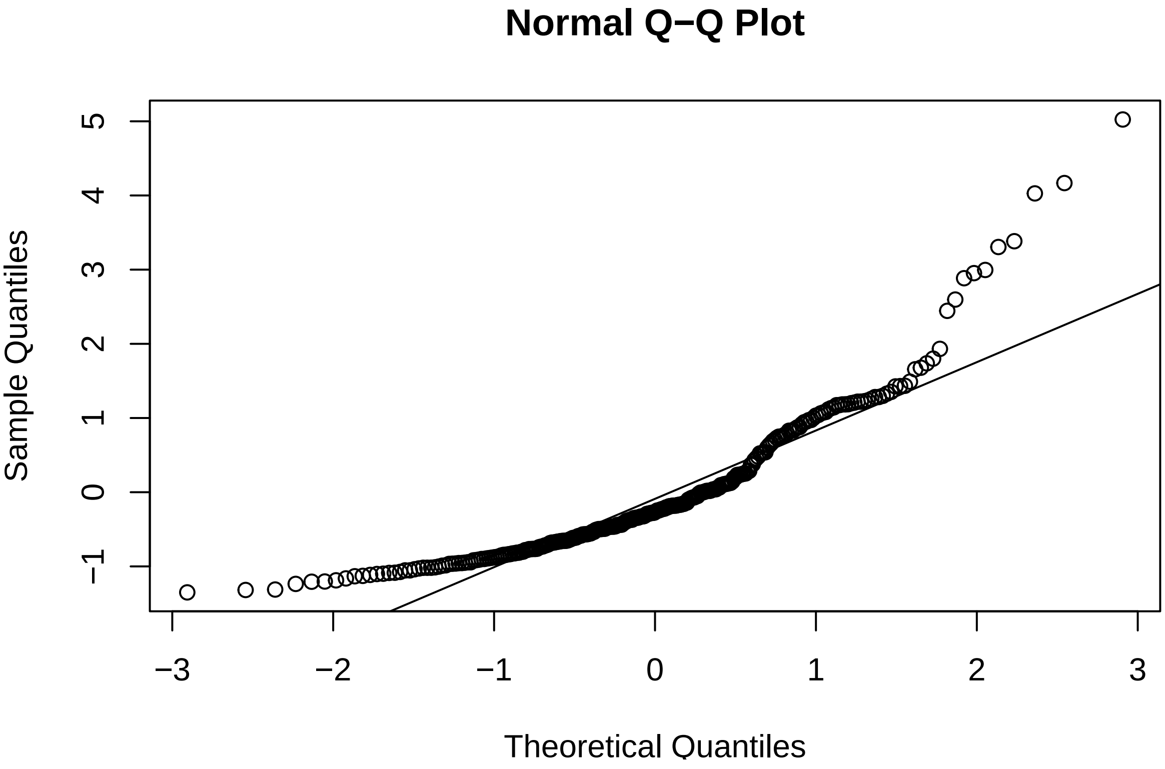
$t_i = \hat{e_i}/(\hat{\sigma}_i\sqrt{1-h_{ii}})$

*ShapiroWilk Normality Test:* For this test we will use the studentized residuals and and formulate the following hypotheses: $H_O$: The observed data is representative of a normal distribution | $H_A$: The observed data is not representative of a normal distribution | $\alpha$(significance level) = 0.05 | Rejection Criteria: pvalue < $\alpha$

```
##
##  Shapiro-Wilk normality test
##
## data:  s_residuals_stock
## W = 0.864, p-value = 7.952e-15
```

From the test conducted, a pvalue of 2.377e-13 is produced, meaning that we would reject the null hypothesis since it meets our criteria of rejection (2.377e-13 < 0.05). This means that the data we are working with is not normally distributed.

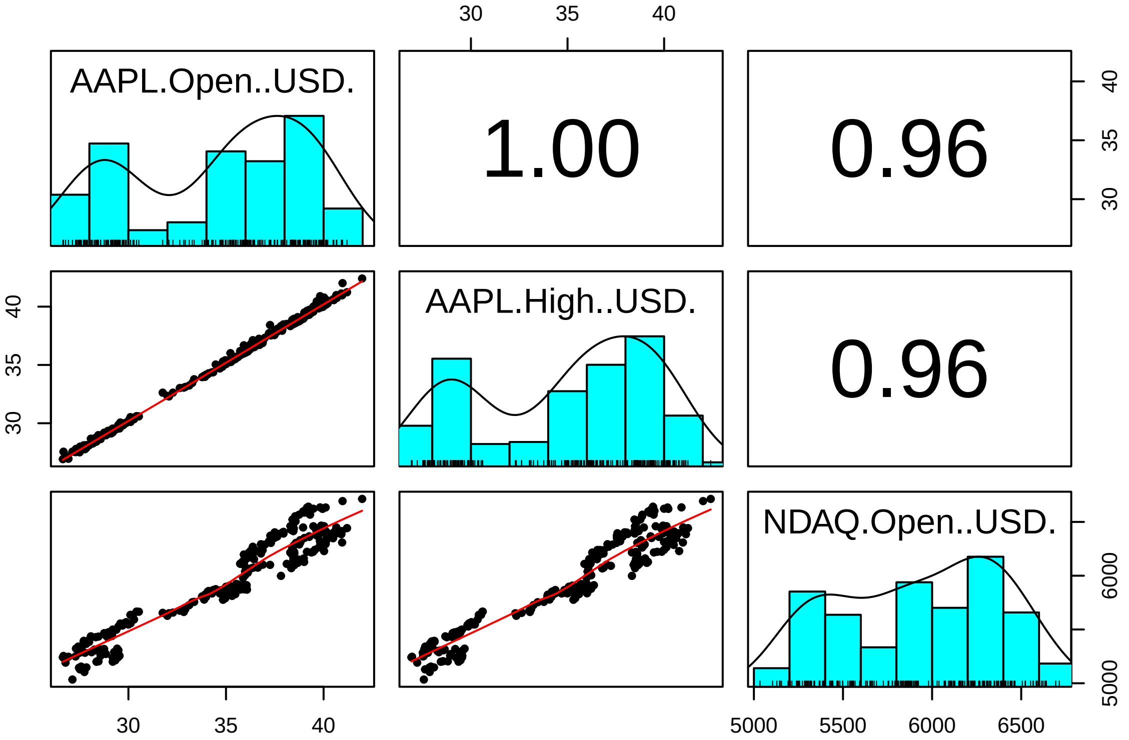
*Normal QQ-Plot:* This idea of non-normality can further be supported by a normal QQ-plot were we are looking for all of the points to be along the line with minimal curvature at the tails of the plot:



From this we can see that the points a barely on this line, thus reinforcing the idea of the absence of normality in the data. This means that our data is very skewed and does not follow the standard bell curve.

**Additional Assumption: Collinearity**

We can use two ways to test collinearity, that being a scatterplot matrix and the variance inflation factors diagnostic:



Immediately there is cause for concern due to one of the regressors and the predictor having a correlation value of one, not to mention the high correlation between the AAPL high regressor and AAPL open regressor.

```
## aapl_and_ndaq_dat$AAPL.Open..USD. aapl_and_ndaq_dat$NDAQ.Open..USD.
##                           11.77624                          11.77624
```

In the case of VIF where any value above 10 is an express cause for concern, the fact that both values are listed as 13.37316 indicates that collinearity is present in the model.

From these assumptions, we can see that this application of a base MLR model was not the best idea, and the fact that the authors' of the paper did not check for these assumptions in the first place is a massive cause for concern when it comes to the predictive ability of this model.

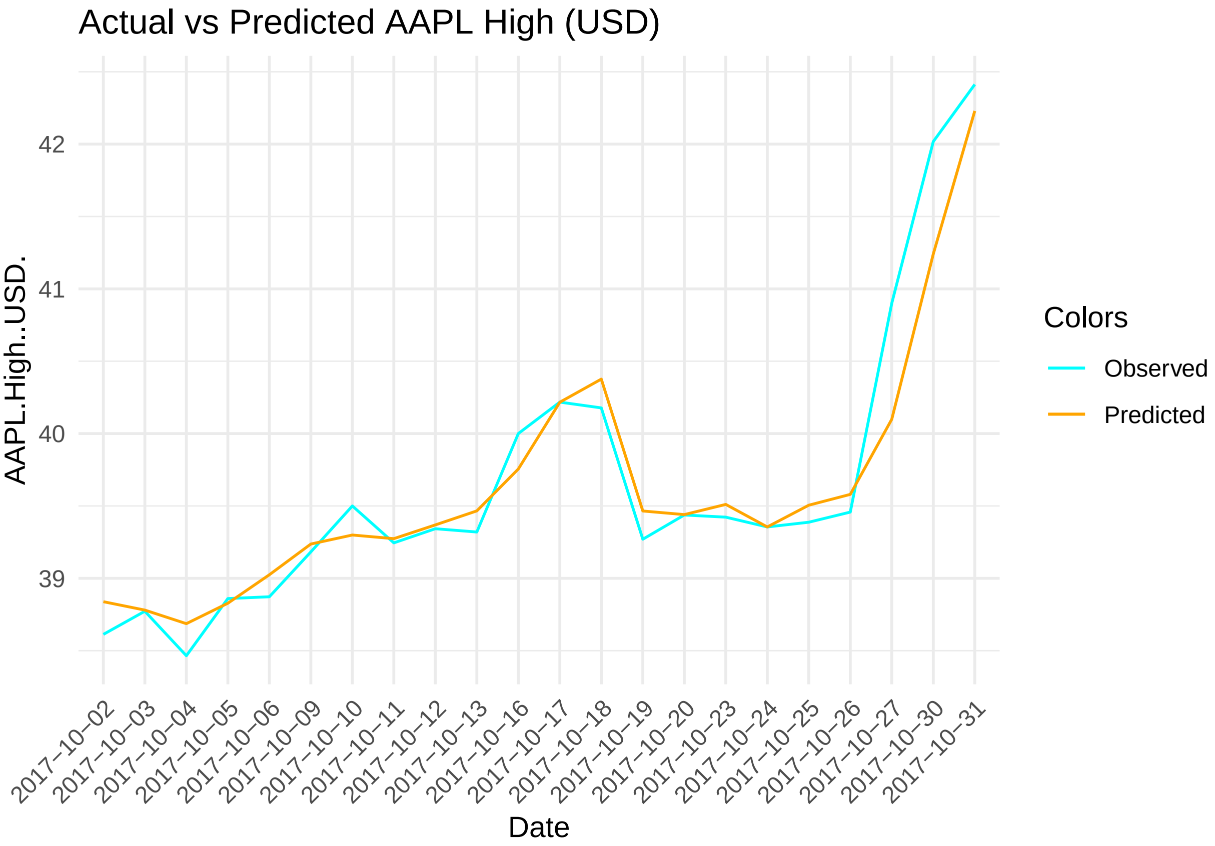
**Model Replication:** This issue of a lack of due diligence continues to persist, as I am unable to replicate the coefficient values and intercept presented in the paper. Using the same source to obtain data, the same model training dates, and the same model regressors still somehow produces different values. Where the authors produce an intercept value of -0.25142712813, an AAPL Open coefficient value of 0.99658308, and a NDAQ Open coefficient value of 0.02304602, my linear model of the same structure produces the following:

```
##
## Call:
## lm(formula = aapl_and_ndaq_dat$AAPL.High..USD. ~ aapl_and_ndaq_dat$AAPL.Open..USD. +
##     aapl_and_ndaq_dat$NDAQ.Open..USD., data = training_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26436 -0.13939 -0.05249  0.10436  0.94539
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -3.842e-01  2.607e-01  -1.474   0.1417
## aapl_and_ndaq_dat$AAPL.Open..USD.  9.822e-01  9.273e-03 105.923   <2e-16 ***
## aapl_and_ndaq_dat$NDAQ.Open..USD.  2.079e-04  9.283e-05   2.239   0.0259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.197 on 271 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 6.877e+04 on 2 and 271 DF,  p-value: < 2.2e-16
```

The equation for APPL_High is as follows:

**APPL_High = -0.3086393 + 0.9826926(AAPL_Open) + 0.0001921(NDAQ_Open)**

The multiple Rsquared value observed in their study differs from mine; theirs is 0.91433103366658208, whereas mine is 0.9981. Additionally, we aim to replicate the graph displayed in the paper, using the test dates ranging from October 1st, 2017, to October 31st, 2017:



It seems that the only aspect of the paper I've managed to replicate is the graph, which closely resembles the original, except for the formatting of the dates on the x axis where the different instances of October 2017 along the graph denotes different points of time in that month. This representation of the model suggests that the multiple linear regression model could have some capacity for stock prediction. However, considering the extensive list of violated assumptions and the time series nature of the data, significant enhancements are necessary to improve the accuracy of predicting Apple's stock high prices.

## Section 4: Enhancement

Based on the earlier discussion, I confidently hold the view that the chosen model was not sufficiently adequate for stock prediction. The main reason behind this inadequacy is the time series nature of the data, meaning that the data is arranged as a sequence of discrete and equally spaced data points collected over time. This implies that all the data points were inherently correlated with each other, which is why most assumption tests failed. Furthermore, the authors referenced utilizing the Python library Pandas' capability to fill in Not a Number (NaN) values in the dataframe. However, considering the data is sourced in the same way done by the authors, the dataframe should have already been cleaned, absent of any NaN values. The authors omitted further explanation regarding the emergence of these values, proceeding with the preprocessed dataset. Returning to the issue of time series analysis, despite this knowledge, none of the authors proposed the utilization of a time series linear model, which I reasoned could significantly enhance the model's predictive capability. However, prior to this, it's essential to test whether incorporating more regressors would enhance the model's predictive ability and potentially satisfy the additional assumptions requisite for employing multiple linear regression. The same Yahoo API can provide the additional factors mentioned at the onset of this report, albeit unused in the paper, enabling testing via the function regsubsets(). This approach can help identify potential models for improvement based on the inclusion of additional regressors:

```
## Reordering variables and trying again:
```

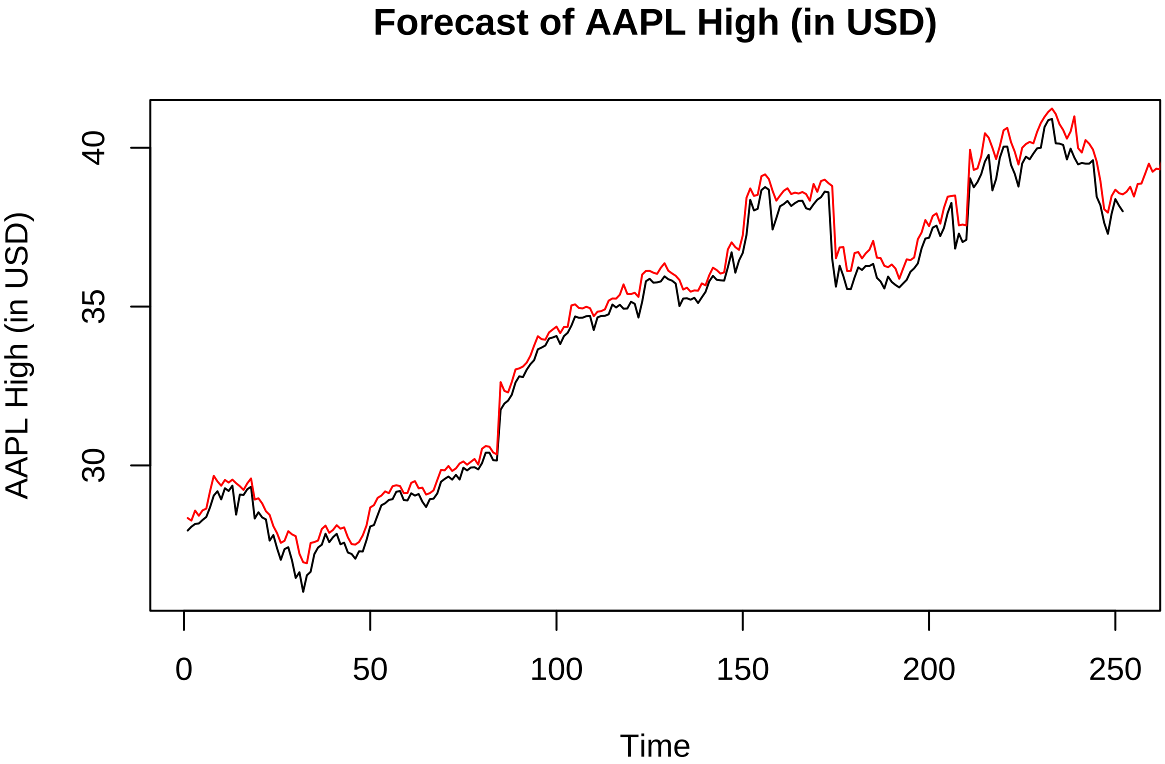
```
##   (Intercept) AAPL_Low AAPL_Open AAPL_Close AAPL_adj_Close AAPL_Share_Vol
## 1           1        0         0          1              0              0
## 2           1        0         1          1              0              0
## 3           1        0         1          1              0              1
## 4           1        0         1          1              0              0
## 5           1        0         1          1              0              1
## 6           1        1         1          1              0              1
## 7           1        1         1          1              1              1
## 8           1        1         1          1              0              1
## 9           1        1         1          1              1              1
##   NDAQ_High NDAQ_Low NDAQ_Open NDAQ_Close NDAQ_adj_Close NDAQ_Share_Vol
## 1         0        0         0          0              0              0
## 2         0        0         0          0              0              0
## 3         0        0         0          0              0              0
## 4         1        0         0          1              0              0
## 5         1        0         0          1              0              0
## 6         1        0         0          0              1              0
## 7         1        0         0          1              0              0
## 8         1        1         1          1              0              0
## 9         1        1         1          1              0              0
##       adjr2       BIC
## 1 0.9980190 -1695.197
## 2 0.9991569 -1924.656
## 3 0.9994054 -2015.736
## 4 0.9994981 -2057.610
## 5 0.9996345 -2139.865
## 6 0.9996516 -2148.446
## 7 0.9996521 -2144.206
## 8 0.9996559 -2142.683
## 9 0.9996555 -2137.763
```

In this scenario, models of particular interest, characterized by a high adjusted R-squared value and a low Bayesian Information Criterion (BIC) score, are Model Four-comprising AAPL_Low, AAPL_Open, AAPL_Close, NDAQ_High, and NDAQ_Close-and Model Six, which includes AAPL_Low, AAPL_Open, AAPL_Close, AAPL_Share_Vol, NDAQ_High, and NDAQ_adj_Close.

```
##           Model      MSE      MAE      VEcv
## mse4 Model Four 533.9389 5.064990 -55720.50
## mse6  Model Six 533.9466 5.065899 -55721.31
```

Although quite close in value, the lower MSE and MAE values indicate that Model 4 is the preferable choice. Furthermore, the higher VEcv value of Model 4 suggests that a larger percentage of the variance in Apple's high price is explained by its low, open, and close, as well as Nasdaq's high and close values. It's important to note that the program reordered the variables due to a detected linear dependency. This reordering contrasts with the original model described in the paper, which included only AAPL_Open and NDAQ_Open as regressors. This difference implies a potential removal of variables from the original model.

Moving on to the issue of the time series data, it has become apparent to me as to why not all of the typical assumptions present for a linear regression model were not tested in this paper. This is because time-series data, even data that appears to be very linear, has a different set of assumptions that we must satisfy. While the typical assumption of independence among residuals must be met (as previously confirmed in section 3), two additional assumptions need to be satisfied: temporal dependence, which signifies that current observations are affected by past ones, and stationarity, indicating that the statistical characteristics of the time series remain constant over time.[5] This means that an entirely different model that accounts for time series and satisfies these two addition points of testing should be used in order to form an accurate prediction of Apple's high price. To address this issue, I suggest employing ARIMA, an autoregressive integrated moving average model. While it doesn't solely rely on the linear regression framework, it encompasses several common features. These characteristics enable it to accommodate the linear relationship within the underlying data while considering the time series nature of the dataset.



Utilizing the model previously derived by regsubsets(), the ARIMA forecast line closely tracks the Apple high data, in contrast to the multiple linear regression plot previously displayed. However, the time frame displayed on the x-axis of the graph used might raise some concerns. Although the values themselves could be concerning, the forecast line's scaling remains consistent with the scale of the AAPL high data. Consequently, I maintain that the graph produces an accurate result. While the additional assumptions of this time series model can be verified, I believe that the concepts of verifying the assumptions of stationarity and temporal dependence lie outside of the scope of this paper.

## Section 5: A Discussion

While I believe my approach to the problem, which was partially unresolved by the authors Shruti Shakhla, Bhavya Shah, Niket Shah, Vyom Unadkat, and Pratik Kanani, is an improvement over their original solution, I acknowledge that, in striving for a more accurate answer, I might have made mistakes as well, particularly in the process of choosing a different model. While I maintain that the chosen regressors from the enhancement section resulted in better prediction outcomes, I also hold the opinion that conducting alternative model selection methods could have further improved the prediction metrics. However, I refrained from employing other model selection techniques as the initial assumptions would likely remain unverified due to the time series nature of the data in question. The existence of such issues has led me to the conclusion that while multiple linear regression can be utilized in stock analysis, its inability to retain its initial strengths when dealing with time series data indicates the need to use alternative methods for achieving improved prediction results. At the very least, this idea is supported by the graphs generated by both the MLR and ARIMA models. However, it's worth noting that there might be some discrepancies present in the ARIMA, potentially casting doubt on the precision of the model. Although it may seem odd to use ARIMA in the context of this paper, there is justification for its usage outside of the contant idea that time series data will not fit well with a regular linear model, multiple or simple otherwise. Although multiple linear regression has its place in the realm of stock data prediction, the flexibility of an ARIMA presents itself as a viable way of navigating the chaotic seas known as the United States financial market.

## References

[1]. Shruti Shakhla, Bhavya Shah, Niket Shah, Vyom Unadkat, & Pratik Kanani(2018). Stock Price Trend Prediction Using Multiple Linear Regression. *International Journal of Engineering Science Invention (IJESI)*, *Volume 7 Issue 10 Ver II*, PP 29-33

[2]. Lock Siew Han, & Md Jan Nordin(2017). Integrated Multiple Linear Regression One | Rule Classification Model for the Prediction of Stock Price Trend. *Journal of Computer Sciences*, DOI: 10.3844/jcssp.2017.422.429

[3]. Weisberg, S. (1980). Multiple Regression. In Applied linear regression (pp. 51). essay, John Wiley & Sons.

[4]. Schifano, E. (2023). STAT 3215Q: Note 3 Multiple Regression: Hypothesis Testing and Inference.

[5]. Complete Disertation. (2023). The Stationary Data Assumption in Time Series Analysis. Statistics Solutions. 2023