Advanced Statistics Final Project:
Tesla Stock Data Multi Regression Analysis

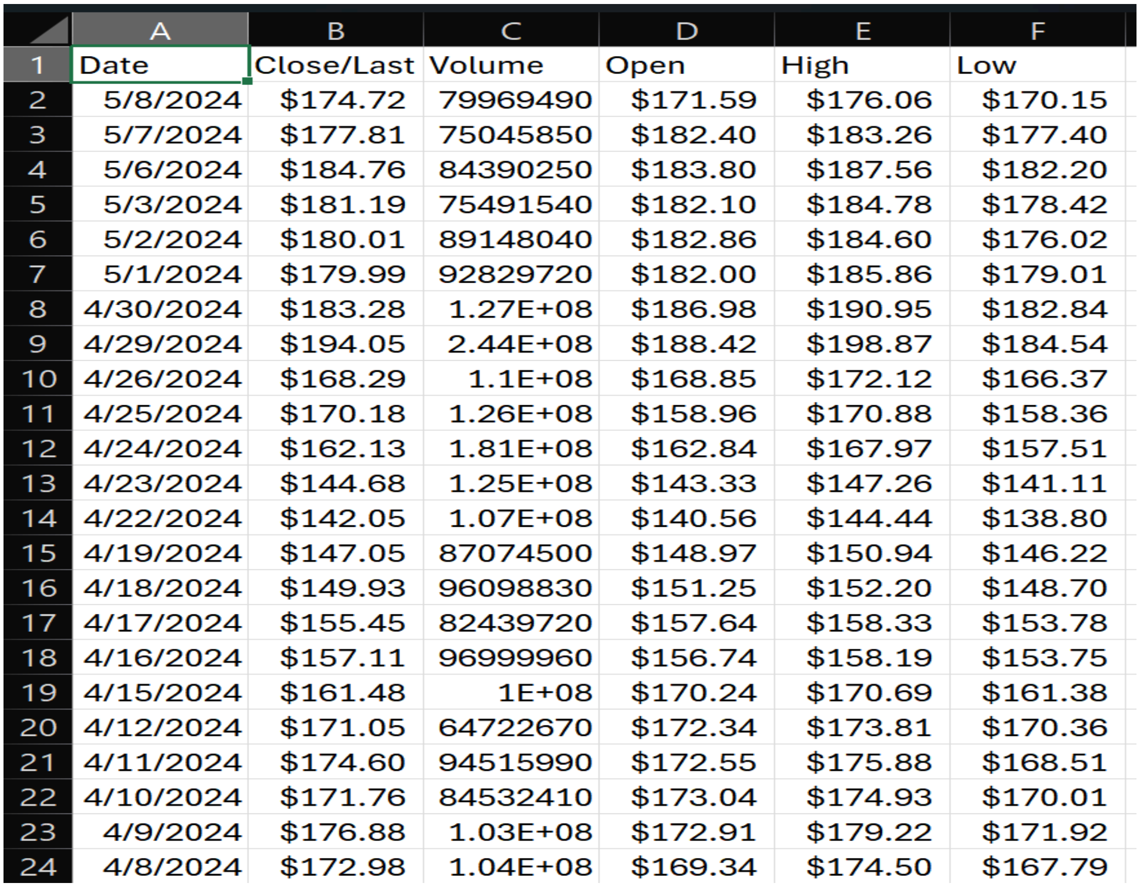
By: Vikram Oberai

May 15, 2024

Choice and Source of Data

Stock’s are a topic that has always fascinated me. One thing I have wanted to learn is how to use computer science to analyze stock data and make predictions about a stock. In this course (Advanced Statistics) I have gained a solid understanding of the R programming language and will be using it to analyze Tesla’s stock data as my final project. Tesla is one of my favorite companies and I wanted to see what conclusions I can find about Tesla’s stock from my analysis.

My data set is sourced from the NASDAQ website. Nasdaq is an American stock exchange and it holds information about stock data that can be downloaded as a .csv file to be analyzed.



	A	B	C	D	E	F
1	Date	Close/Last	Volume	Open	High	Low
2	5/8/2024	$174.72	79969490	$171.59	$176.06	$170.15
3	5/7/2024	$177.81	75045850	$182.40	$183.26	$177.40
4	5/6/2024	$184.76	84390250	$183.80	$187.56	$182.20
5	5/3/2024	$181.19	75491540	$182.10	$184.78	$178.42
6	5/2/2024	$180.01	89148040	$182.86	$184.60	$176.02
7	5/1/2024	$179.99	92829720	$182.00	$185.86	$179.01
8	4/30/2024	$183.28	1.27E+08	$186.98	$190.95	$182.84
9	4/29/2024	$194.05	2.44E+08	$188.42	$198.87	$184.54
10	4/26/2024	$168.29	1.1E+08	$168.85	$172.12	$166.37
11	4/25/2024	$170.18	1.26E+08	$158.96	$170.88	$158.36
12	4/24/2024	$162.13	1.81E+08	$162.84	$167.97	$157.51
13	4/23/2024	$144.68	1.25E+08	$143.33	$147.26	$141.11
14	4/22/2024	$142.05	1.07E+08	$140.56	$144.44	$138.80
15	4/19/2024	$147.05	87074500	$148.97	$150.94	$146.22
16	4/18/2024	$149.93	96098830	$151.25	$152.20	$148.70
17	4/17/2024	$155.45	82439720	$157.64	$158.33	$153.78
18	4/16/2024	$157.11	96999960	$156.74	$158.19	$153.75
19	4/15/2024	$161.48	1E+08	$170.24	$170.69	$161.38
20	4/12/2024	$171.05	64722670	$172.34	$173.81	$170.36
21	4/11/2024	$174.60	94515990	$172.55	$175.88	$168.51
22	4/10/2024	$171.76	84532410	$173.04	$174.93	$170.01
23	4/9/2024	$176.88	1.03E+08	$172.91	$179.22	$171.92
24	4/8/2024	$172.98	1.04E+08	$169.34	$174.50	$167.79

Data Set Information and Variables

The data set contains information about Tesla's stock data for a 30 day time period. The information included in the data set is the Date | Volume | Opening Price | Closing Price | High Price | Low Price. The volume indicates the amount of shares of Tesla were traded on a specific trading day. The opening price represents the price of Tesla at the market open. The closing price represents the price of Tesla at the market close. The high price and low price represent the highest price Tesla reached and the lowest price Tesla reached during trading hours.

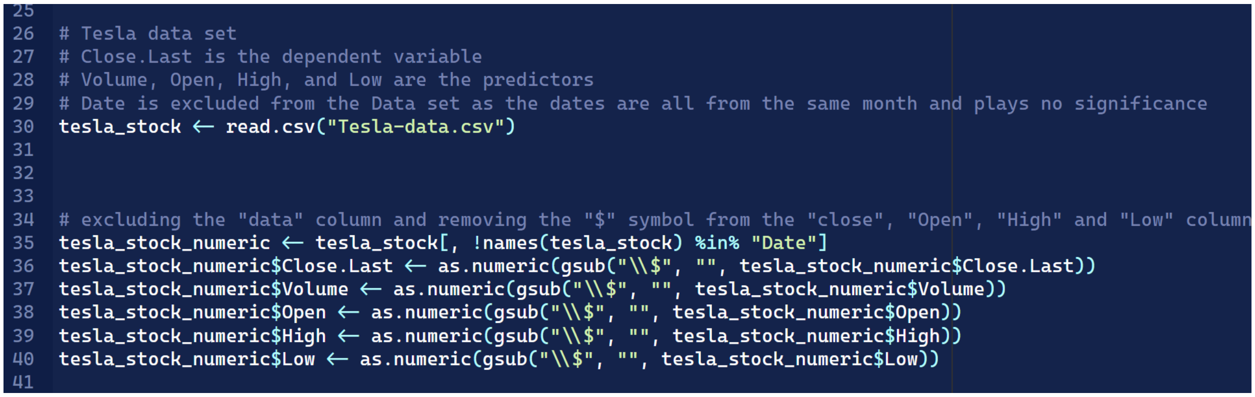
Independent Variables (Predictors):

- Opening
- High
- Volume
- Low

Dependent Variable:

- Close

To perform the multi regression analysis I needed to make changes to the format of my data set. I removed the "Date" column as it was insignificant to the analysis and did not contain numeric data. In the Opening, High, Low, and Closing columns, the price of Tesla was preceded by a dollar sign, which changed the data type of the values. To fix this issue, I removed the "$" sign and changed the values to numeric values.



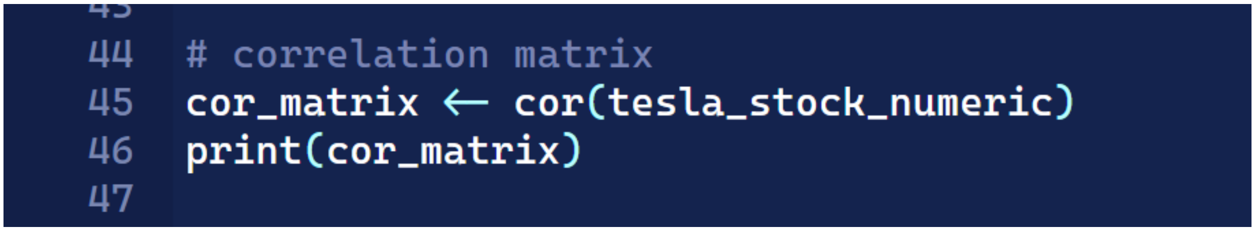
```
# Tesla data set
# Close.Last is the dependent variable
# Volume, Open, High, and Low are the predictors
# Date is excluded from the Data set as the dates are all from the same month and plays no significance
tesla_stock ← read.csv("Tesla-data.csv")



# excluding the "data" column and removing the "$" symbol from the "close", "Open", "High" and "Low" column
tesla_stock_numeric ← tesla_stock[, !names(tesla_stock) %in% "Date"]
tesla_stock_numeric$Close.Last ← as.numeric(gsub("\\$", "", tesla_stock_numeric$Close.Last))
tesla_stock_numeric$Volume ← as.numeric(gsub("\\$", "", tesla_stock_numeric$Volume))
tesla_stock_numeric$Open ← as.numeric(gsub("\\$", "", tesla_stock_numeric$Open))
tesla_stock_numeric$High ← as.numeric(gsub("\\$", "", tesla_stock_numeric$High))
tesla_stock_numeric$Low ← as.numeric(gsub("\\$", "", tesla_stock_numeric$Low))
```

Correlation Matrix

The correlation matrix shows the strength of the relationship between the variables. In a multi regression analysis it is preferred when there is a strong correlation between the predictors and the dependent variable, and a weak correlation between the predictors.

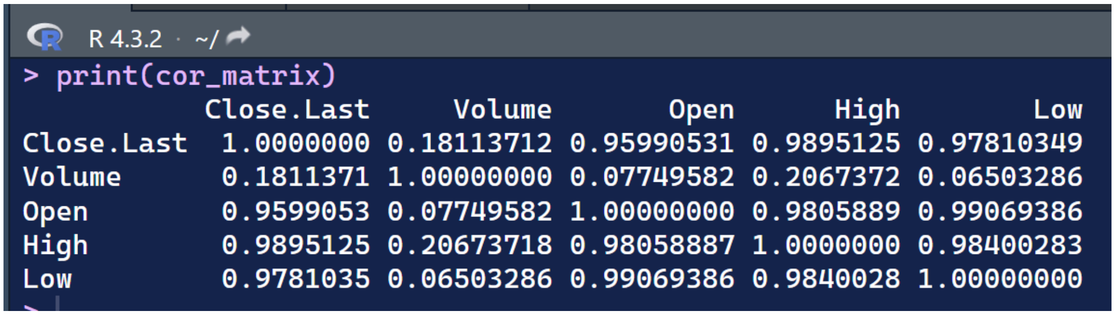


```r
# correlation matrix
cor_matrix <- cor(tesla_stock_numeric)
print(cor_matrix)
```

As seen in the correlation matrix for the Tesla data set, the dependent variable (Close.Last) has a strong positive correlation between the predictor variables – Open, High, and Low –, and the predictor variable – Volume – has a weak positive correlation with the other predictor variables.

However, there are a couple of problems that arise from this correlation matrix. The dependent variable has a weak positive correlation with the Volume, and the predictor variables – Open, High, and Low – have a strong positive correlation with one another. This correlation is not ideal as the predictor variables are supposed to be independent and therefore should not have a high correlation with other predictor variables.

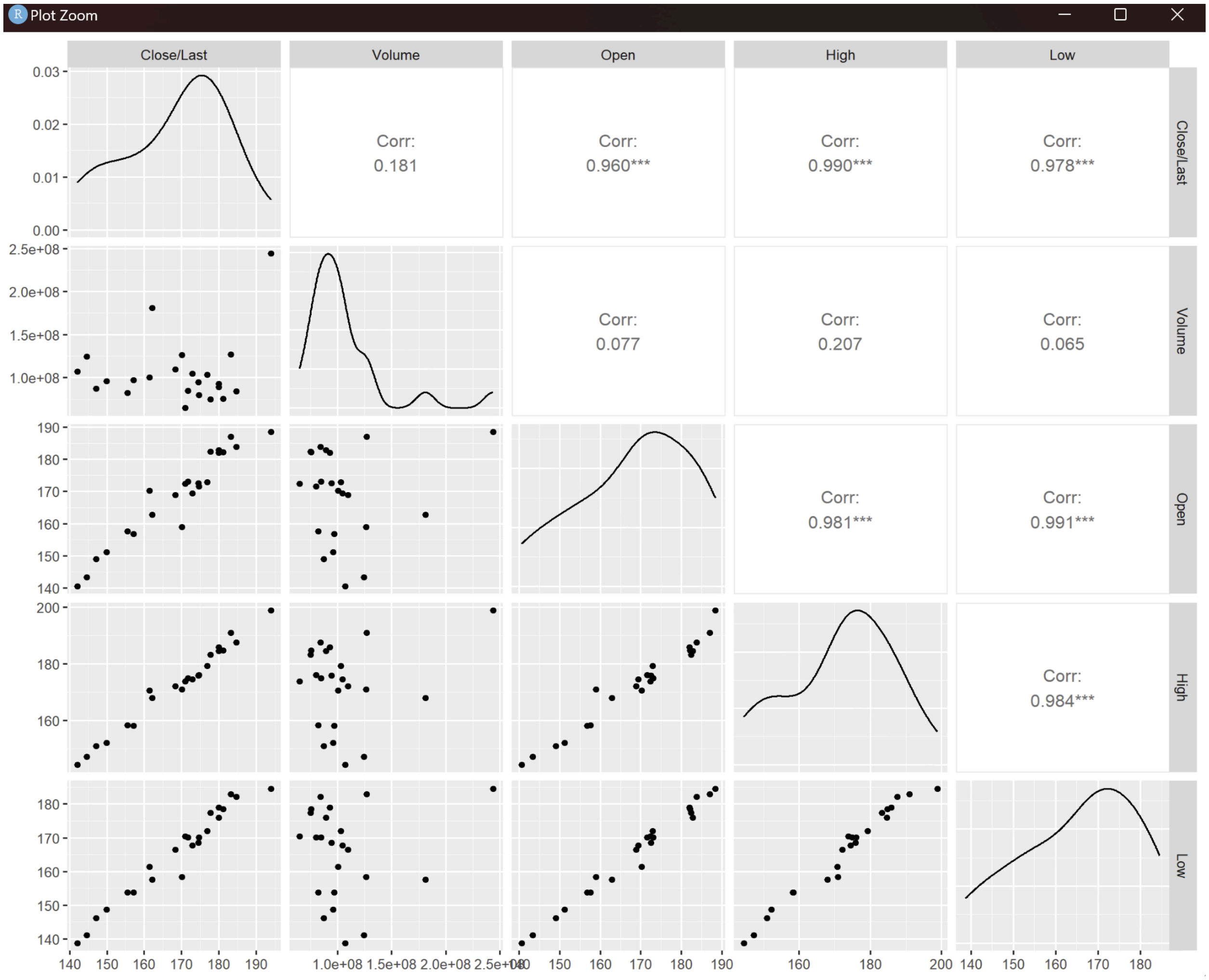
Since volume and closing price have a weak correlation, this correlation may indicate that volume is not the best predictor for the closing price.



```
> print(cor_matrix)
           Close.Last     Volume       Open      High        Low
Close.Last  1.0000000 0.18113712 0.95990531 0.9895125 0.97810349
Volume      0.1811371 1.00000000 0.07749582 0.2067372 0.06503286
Open        0.9599053 0.07749582 1.00000000 0.9805889 0.99069386
High        0.9895125 0.20673718 0.98058887 1.0000000 0.98400283
Low         0.9781035 0.06503286 0.99069386 0.9840028 1.00000000
```

Scatter Plots of Pairwise Variables

The Scatter plot for pairwise variables shows the relationship between two variables. As seen in the Scatter Plot below, each variable is on the top and the right side, with a graph and correlation number given. From the scatter plot, there is a clear strong positive correlation between the Close/Last and the Open, High, and Low variables. This correlation is shown by the strong linear relationship between the points. However, there seems to be a weak positive correlation with the Volume variable. This correlation is clearly seen by the random scatter of the points between the Volume and the Close/Last.



Initial Model

y = B0 + B1 (x) + B2 (x) + B3 (x) + B4 (x)

Close.Last = B0 + B1 (Volume) + B2 (Open) + B3 (High) + B4 (Low)

In this model, y represents the dependent variable (Closing Price). This is what the model is trying to predict. The model depicts the relationship between the predictor variables and the dependent variable. The goal of the model is to predict the value of the dependent variable, based on a change in the predictor variable values.

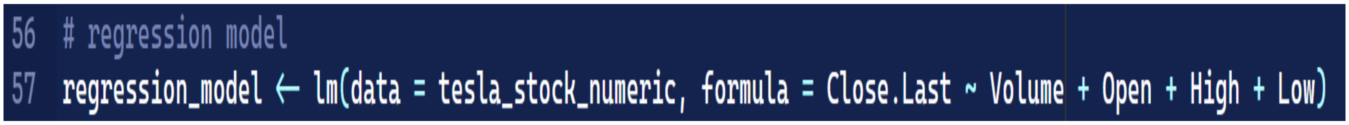
B0: y - intercept

B1: Coefficient of the Volume variable

B2: Coefficient of the Open variable

B3: Coefficient of the High Variable

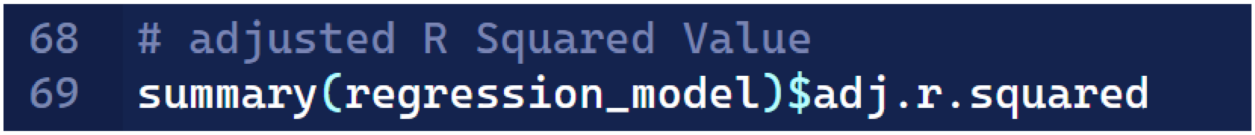
B4: Coefficient of the Low Variable



```
# regression model
regression_model <- lm(data = tesla_stock_numeric, formula = Close.Last ~ Volume + Open + High + Low)
```

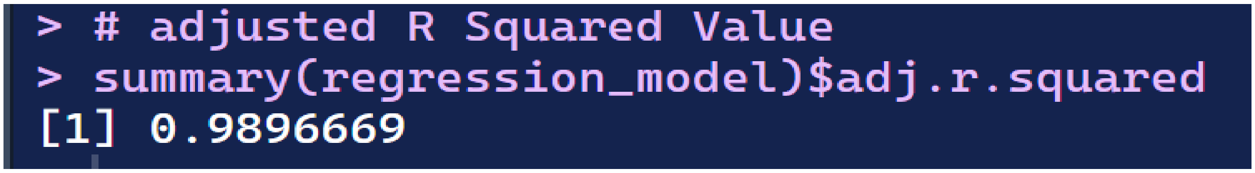
Analysis of the Overall Fit of the Model

The overall fit of the model is represented by the adjusted R squared value. An adjusted R squared value closer to 1 indicates that the majority of the variation can be explained by the model and an adjusted R squared value closer to 0 means that the model can not explain the majority of the variance.



```
# adjusted R Squared Value
summary(regression_model)$adj.r.squared
```

The adjusted R squared value for this data set is 0.989669 which is extremely close to 1. This adjusted R squared value means that 98.97% of the variation can be explained by the model. This value indicates that the model is a great fit and is effective in explaining variance based on the change in predictor values.



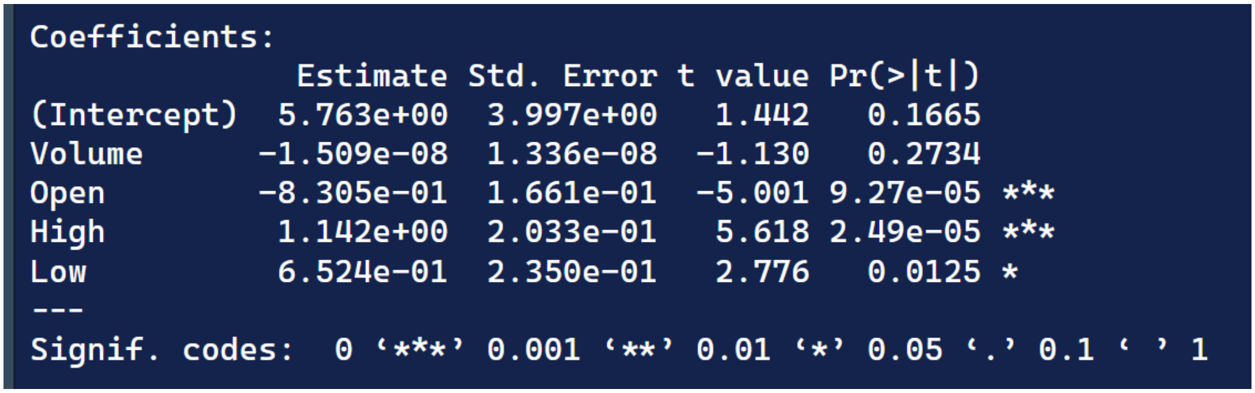
```
> # adjusted R Squared Value
> summary(regression_model)$adj.r.squared
[1] 0.9896669
```

Coefficients Test of Significance

The coefficient test signifies the strength of the linear relationship between the independent variable and the dependent variable. A coefficient that has a P - value less than 0.05 is significant, while a coefficient that has a P - value greater than 0.05 is not significant.

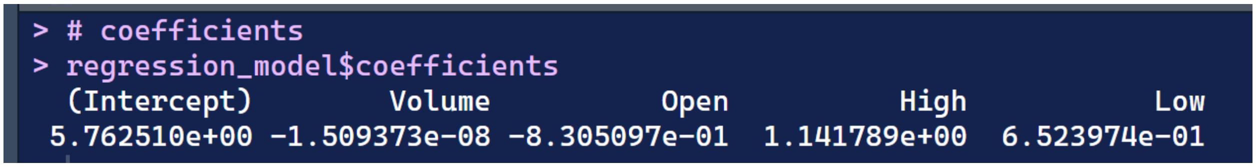
In the case of this model, the Open, High, and Low coefficients are significant as their P - Values are less than 0.05. Due to their P - Values being less than 0.05, these independent variables have a significant impact on the Closing price (Dependent variable).

On the other hand, the Volume and y - intercept have P - Values greater than 0.05, indicating that they are insignificant, and it can be concluded that there is not enough evidence to support that these coefficients have an impact on the Closing price.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.763e+00  3.997e+00   1.442   0.1665    
Volume      -1.509e-08  1.336e-08  -1.130   0.2734    
Open        -8.305e-01  1.661e-01  -5.001 9.27e-05 ***
High         1.142e+00  2.033e-01   5.618 2.49e-05 ***
Low          6.524e-01  2.350e-01   2.776   0.0125 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Individual Coefficient Interpretation



```
> # coefficients
> regression_model$coefficients
  (Intercept)        Volume          Open          High           Low 
 5.762510e+00 -1.509373e-08 -8.305097e-01  1.141789e+00  6.523974e-01 
```

B0 (y - intercept): 5.763e+00

- The estimated closing price of Tesla when all other predictor variables = 0 (x = 0).

B1 (Coefficient for Volume): -1.509e-08

- The closing price will decrease by $-1.509e-08 for every increase of 1 share traded of Tesla when all other independent variables stay constant.

B2 (Coefficient for Open): -8.305e-01

- The closing price will decrease by $-8.305e-01 for every $1 increase in the opening price while all other independent variables stay constant.

B3 (Coefficient for High): 1.142e+00

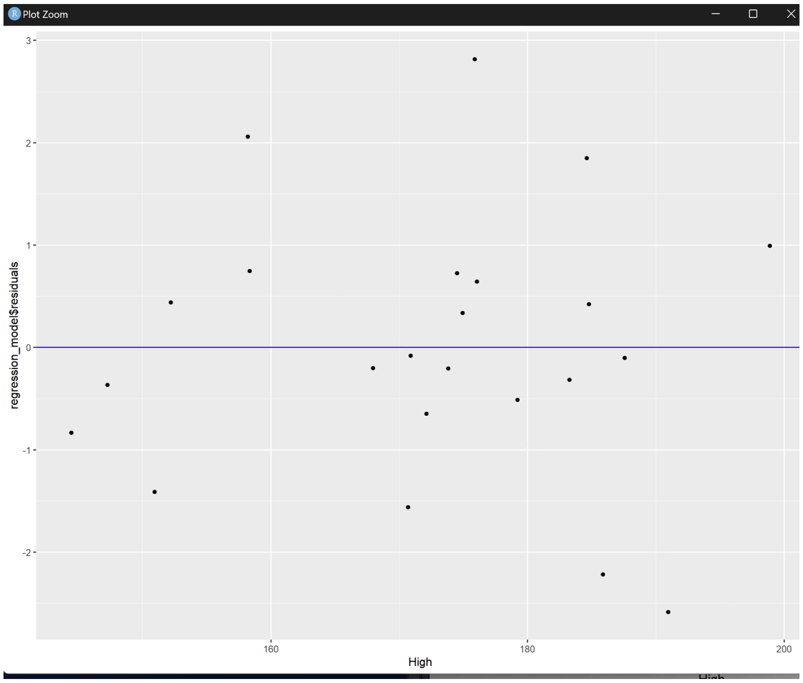
- The closing price will increase by $1.142e+00 for every $1 increase in the highest price of Tesla during a trading day while all other independent variables stay constant.

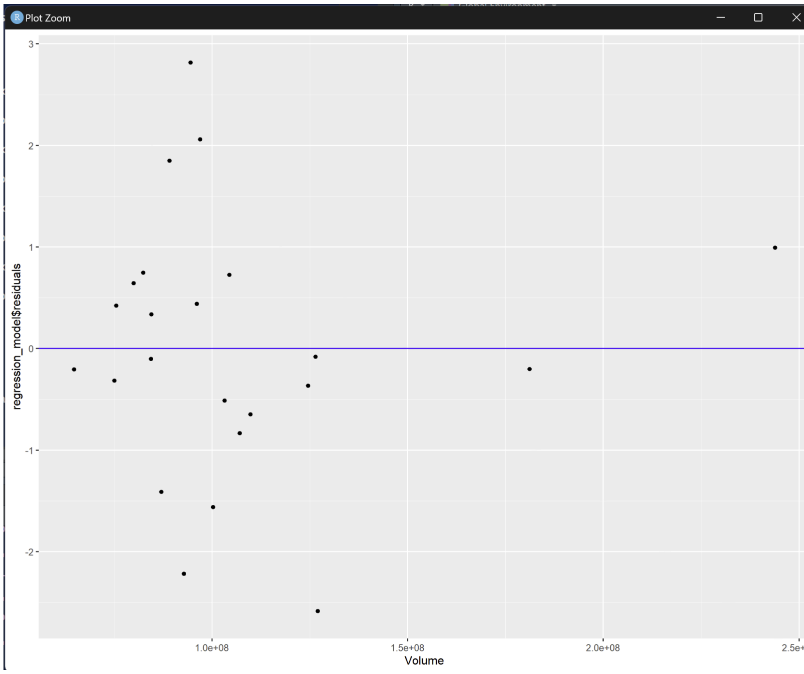
B4 (Coefficient for Low): 6.524e-01

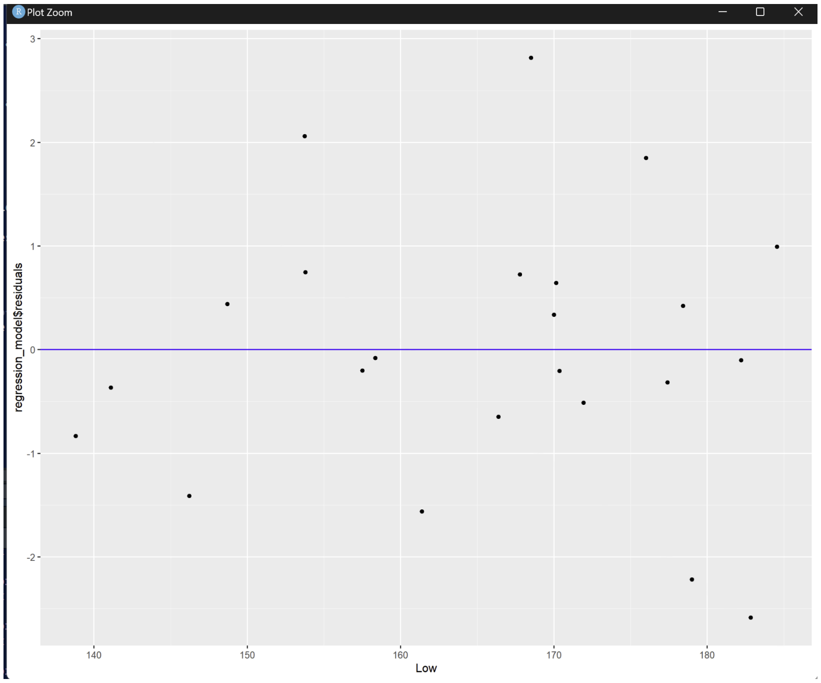
- The closing price will increase by $6.524e-01 for every $1 increase in the lowest price of Tesla during a trading day while all other independent variables stay constant.

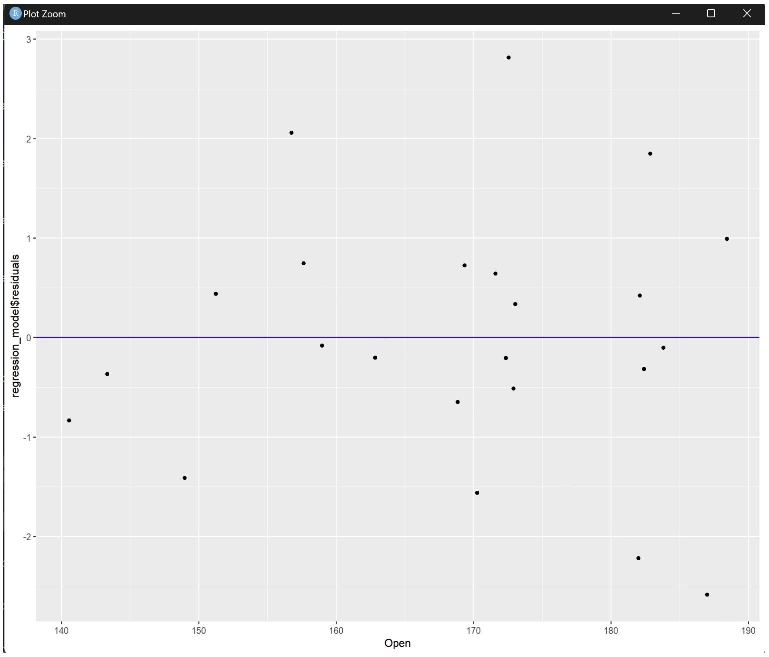
Residual Plot for Each Non-Dummy Predictor

The residual plots help show whether or not there is constant variance. Constant variance is desired in residual plots as it means there is no pattern of variance and the variance of the residuals are the same for different values of the independent variables. From the plots below it is clear that the High, Low, and Open variables have constant variance as the points are approximately split evently with half on top and half on bottom with no clear pattern. With the Volume residual plot, there seems to be non constant variance as a pattern seems to be forming where the majority of the residual points are clustered to the left of the plot. This can have impacts on the analysis and cause the coefficient to be insignificant.









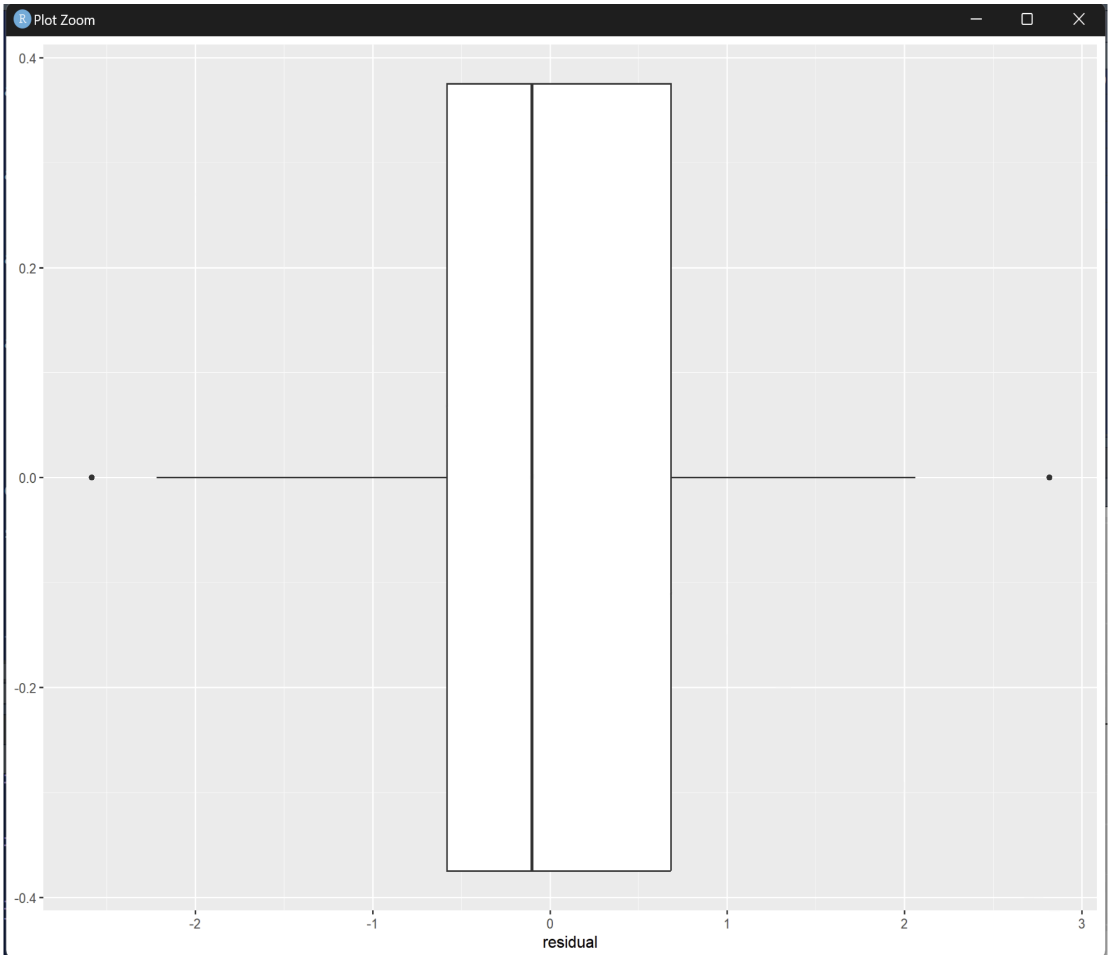
Boxplot of the Residuals

The boxplot for the residuals help provide multiple insights on the data. Since there is a big box plot, this indicates that there is a higher variance and spread between the residuals, which may indicate that the model is not capturing the variance in its entirety.

Furthermore, the box plot shows that there are residuals that are outliers. These outliers are represented by the point to the left of the whisker, and the point to the right of the whisker.

Based on these results, it can be concluded that there are some residuals that are too small and too large, and therefore deviate away from the normal distribution of the model.

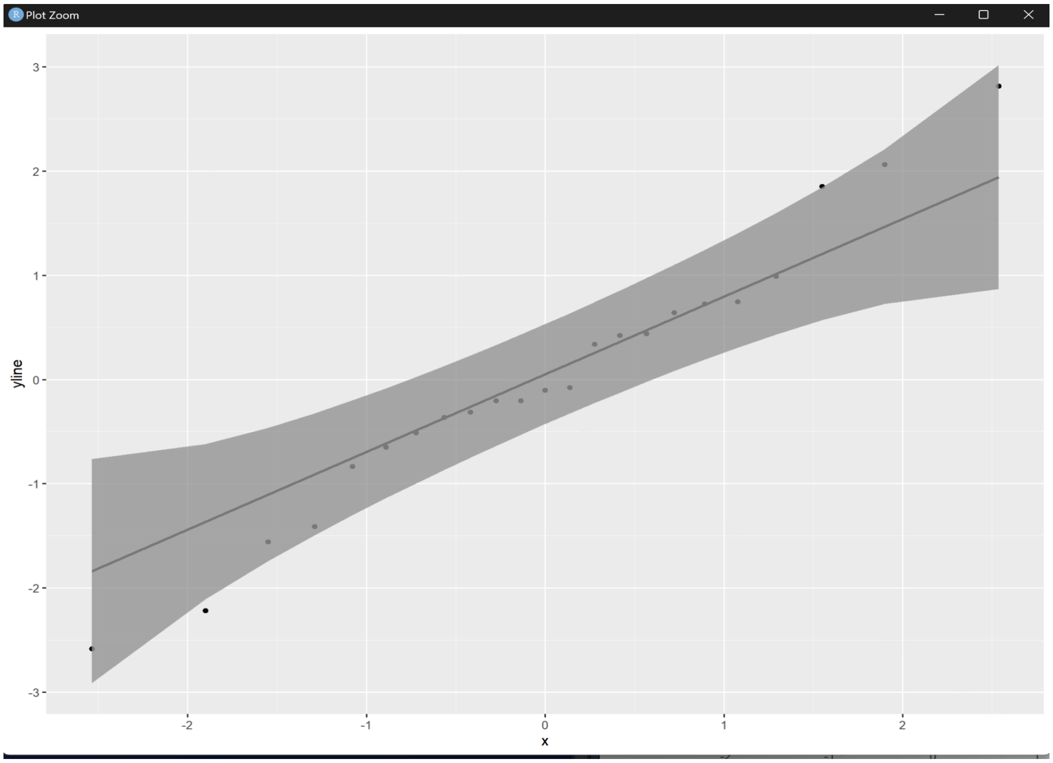
Since there are residual points that deviate from the rest of the data, it is important to conduct more tests to check for normality.



QQ-Plot of the Residuals

The QQ-Plot is used to check the normality of the Residual points. To have a normal distribution the points must follow the 45 degree reference line and fall within the gray confidence bands. As seen below, in the QQ-Plot the vast majority of the residual points fall within the gray bands and follow the 45 degree reference line, therefore it can be concluded that the residuals have a normal distribution.

The only problem identified in the QQ-Plot is at the beginning of the plot, one of the residual points does not fall within the gray bands. However, since it is only one point that has a small deviation from the gray bands, it does not significantly impact the overall normal distribution of the model.



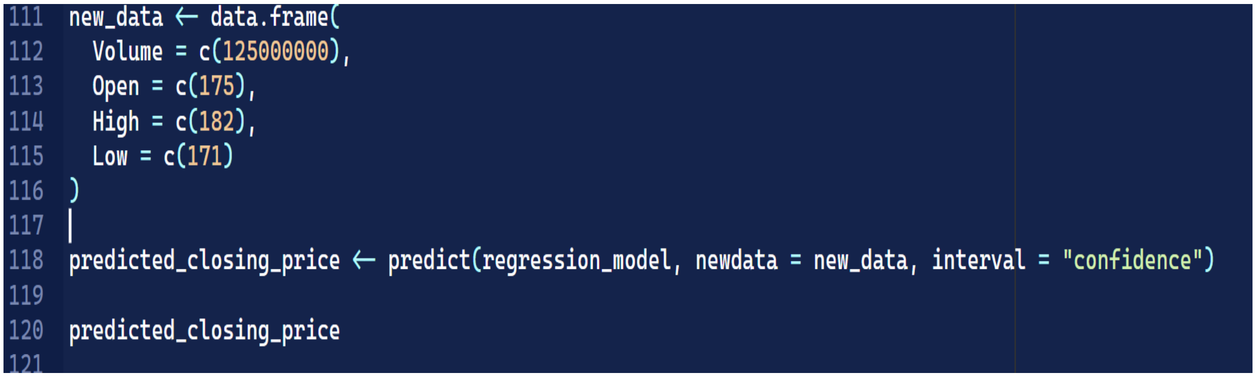
Predictor Predictions

For my Predictors, I choose values that were in a reasonable range based on the data provided in the data set. This allows me to compare the results of the closing price given by the prediction, to the closing prices of data points with similar values for the independent variables.

New Values:

- Volume: 125,000,000
- Opening Price: $175
- High Price: $182
- Low Price: $171

Result: The price of Tesla's stock with a trading volume of 125,000,000, an opening price of $175, a trading day high of $182, and a trading day low of $171, would result in a closing price of $177.90 with a 95% confidence of a closing price inbetween $176.52 and $179.28.



```
new_data <- data.frame(
  Volume = c(125000000),
  Open = c(175),
  High = c(182),
  Low = c(171)
)

predicted_closing_price <- predict(regression_model, newdata = new_data, interval = "confidence")

predicted_closing_price
```

Summary of Analysis

Not all regression models are a good fit for a data set. When deciding whether a model is a good fit for the data, it is important to consider the distribution, the adjusted R squared value, the collinearity between variables, and the significance of coefficients.

Positives:

The model had a high adjusted R squared value of 98.97%. This is a positive indicator in deciding whether or not the model fits the data. A high adjusted R squared value of 98.97% means that almost all of the variance in the data was able to be explained by the mode. Another positive aspect of the model was the strong collinearity between the Closing price and the majority of the Independet variables – Open, High, and Low. This shows that the Closing price is impacted by the change in these variables. Finally, the data was mostly normally distributed. Having a normal distribution is important for trying to predict natural phoemens such as the price of a stock.

Drawbacks:

There were several drawbacks to the model. The majority of the predictors, besides Volume, all had strong positive correlations with one another. This relationship is generally not ideal as the predictors are supposed to be independent variables and not heavily influenced by changes in other variables. Also, the Volume variable seemed to be a constant problem in the model as it had a weak correlation with the Closing price, it had a non constant variation for its residuals, and its coefficient was not significant.

Conclusion:

Overall, the model seems like a good fit for the data. The residual points were mostly normally distributed with 3 of the 4 Coefficients being significant. Furthermore, the majority of the predictors had a high strong positive correlation with the dependent variable indicating that changes in those predictors played a significant impact in the value of the dependent variable. Finally, the high adjusted R squared value is another key factor for making this model a great fit for the data, as most of the variance was captured and explained by the model. To improve the model, I would consider taking out the Volume variable, as it was affecting the normal distribution and its coefficient was not significant.