Amazon NASDAQ Price Compared to Total Market

Stephan Schuller

October 9, 2023

1 Introduction

1.1 Goals

We hope to show, by creating a *multiple linear regression* model, whether the *Open Price* of *Amazon* and the *Open Price* of the *NASDAQ Composite* have some capacity to explain the daily *High Price* of *Amazon* stock.

1.2 Assumptions about the Data

We assume the data provided is accurate by date-line; this is because the variables from the two data sets are related by date. Therefore, the sample size of this population are measured in days, and we hope to understand by a simple linear regression model and a multiple linear regression model if an assumption can be made about dates which are not part of this data set.

1.3 Hypothesis

The Null Hypothesis can be stated as follows: there is no correlation between the *Open Price* of *Amazon*, the *Open Price* of the *NASDAQ Composite* and the daily *High Price* of *Amazon*. Additionally we are unable to reasonably say that the Open Price of Amazon and the Open Price of NASDAQ Composite are valuable for explaining the daily High Price of Amazon.

2 R-Script & Overview of Functions

2.1 Models

Linear Regression

Using a simple linear regression model we hope to measure if there is a better model compared to simple mean for describing the daily High Price of Amazon. According to the project notes, we can examine the correlation between the Open Price of Amazon and High Price of Amazon for a potentially improved model.

For simple linear regression, where Y is the dependent variable, X is the independent variable, m is the estimated slope, and b is the estimated intercept. This is summarized by the formula:

$$Y = mX + b$$

For this particular example Y is the Amazon High price and X is the value of the predictor variable, Amazon Open Price, the simple linear regression model will have to be made before the rest of the formula can be solved.

$$High.amzn = m \cdot Open.amzn + b$$

Multiple Linear Regression

Using Multiple Linear Regression we hope to know if there is a better model compared to the simple regression model described above. According to the project notes, we can examine the correlation between the Linear Regression of the Open Price of Amazon and the High Price of Amazon with the the Open Price of the NASDAQ Composite for a potentially improved model.

Multiple Linear Regression can be explained in the following formula. Where Y_i is the dependant variable, B_0 is the Y intercept, X is the explanatory variables B_k are the multiple slope coefficients for each explanatory variable, and E is the residual or error.

$$Y_i = B_0 + B_1 \cdot X_1 + B_2 \cdot X_2 + \varepsilon$$

For this particular example Y is the Amazon High price, X_1 is the Amazon Open Price and X_2 is the NASDAQ Composite Open Price, the multiple linear regression model will have to be made before the rest of the formula can be solved.

$$High.amzn = B_0 + B_1 \cdot Open.amzn + B_2 \cdot Open.ixic + \varepsilon$$

2.2 R-Studio Libraries Used

Tidyverse

Tidyverse is an R programming package that helps to transform and better present data. It assists with data import. This contains the package GGplot 2 which plotting, data manipulation, and data visualization easier for us.

2.3 Data Sets

Amazon (AMZN.csv)

The data contained in the AMZN.csv includes: Date, Open, High, Low, Close, Adjusted Close, and Volume. The Date range represented is between 2/18/2022 and 2/17/2023. Typically the recorded days are cluster of 5 or 4 consecutive dates separated by 2 or 3 "missing" days. Checking a calendar for the month and day will show that the recorded days are weekdays and missing days are weekends or holidays. The dates recorded int the AMZN.csv file include all weekdays(Monday, Tuesday, Wednesday, Thursday, Friday) except holidays, and also excluded weekends (Saturday and Sunday).

Nasdaq Composite (IXIC.csv)

The data contained in the IXIC.csv file includes: Date, Open, High, Low, Close, Adjusted Close, and Volume. The Date range is from 1/2/2015 to 2/16/2023. Typically the recorded days are cluster of 5 or 4 consecutive dates separated by 2 or 3 "missing" days. Checking a calendar for the month and day will show that the recorded days are weekdays and missing days are weekends or holidays. The dates recorded int the IXIC.csv file include all weekdays(Monday, Tuesday, Wednesday, Thursday, Friday) except holidays, and also excluded weekends (Saturday and Sunday).

Reading the Data Sets

First the data sets must be read from their original format and brought into R-Studio in a format that is usable by our required functions. This is done through use of the *read.csv* function:

```
# Read in the .csv files
amzn_data <- read.csv(file.path(file_root, "AMZN.csv"))
ixic_data <- read.csv(file.path(file_root, "^IXIC.csv"))
```

2.4 Manipulating Data

The two data sets must be "cleaned up" and unified into a single data set for better use of the data. Because we intend to model the data by **date** be must first make sure the Date column in both data sets is representative of a real date. Otherwise they will be incompatible to merge. The date columns must be compared and merged together; unless all dates are formatted consistently there is risk of something unexpected happening to the data To accomplish this formatting we use the *as.Date* function:

```
# Convert the Date fields to Date type
amzn_data$Date <- as.Date(amzn_data$Date)
ixic_data$Date <- as.Date(ixic_data$Date)
```

After the Dates from both data sets have been formatted correctly, we can safely combine them into a single data set using the *merge* function:

```
# Combine the datasets by date
combined_data <- merge(amzn_data, ixic_data, by = "Date", suffixes = c(".amzn", ".ixic"))
```

This also has the added benefit of eliminating any dates from the two data sets which are not present in the other. This means that the much larger NASDAQ Composite data set is pruned. This is necessary since we are analyzing individual dates. So it would not be beneficial to model dates where some of the data is missing.

After the two data sets have been merged on the Date column we are able to removed any lines with missing values using the *omit* function:

```
# Remove rows with NA values
combined_data <- na.omit(combined_data)
```

We now have a single data set. It originated from two .csv files, was format corrected for date, merged, and pruned. We are ready to begin modelling with this data.

2.5 Modelling and Plotting Data

The most prudent place to begin is to make a simple regression model using only information originating from the Amazon data set. This is necessary because evaluating the standards for accuracy (especially for complex models) is best done comparatively.

Simple Linear Regression Model

The first model will be a linear regression model created by using the *High Price* from *Amazon* and the *Open Price* from *Amazon*. This is accomplished using the *lm* function:

```
# Create a linear regression model with just amazon data.
model <- lm(High ~ Open, data = amzn_data)
```

With the newly created model we are now able to easily create a helpful summary with the *summary* function:

```
# Get model summary and output to console
summary(model)
```

Finally we are able to create the most compelling part, the plot. This is created with the GGPlot2 library using the *ggplot* function as well as adjust the type of plot and what will be displayed:

```
# Generate plot of the model
plot_model <- ggplot(amzn_data, aes(x = Open , y = High)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Amazon High Price vs. Open Price",
       x = "Amazon Open Price",
       y = "Amazon High Price") +
  theme_bw()
```

Multiple Linear Regression Model

For the multiple linear regression model, the steps will be familiar. We begin by creating the model:

```
# Create a multiple linear regression model
model <- lm(High.amzn ~ Open.amzn + Open.ixic, data = combined_data)
```

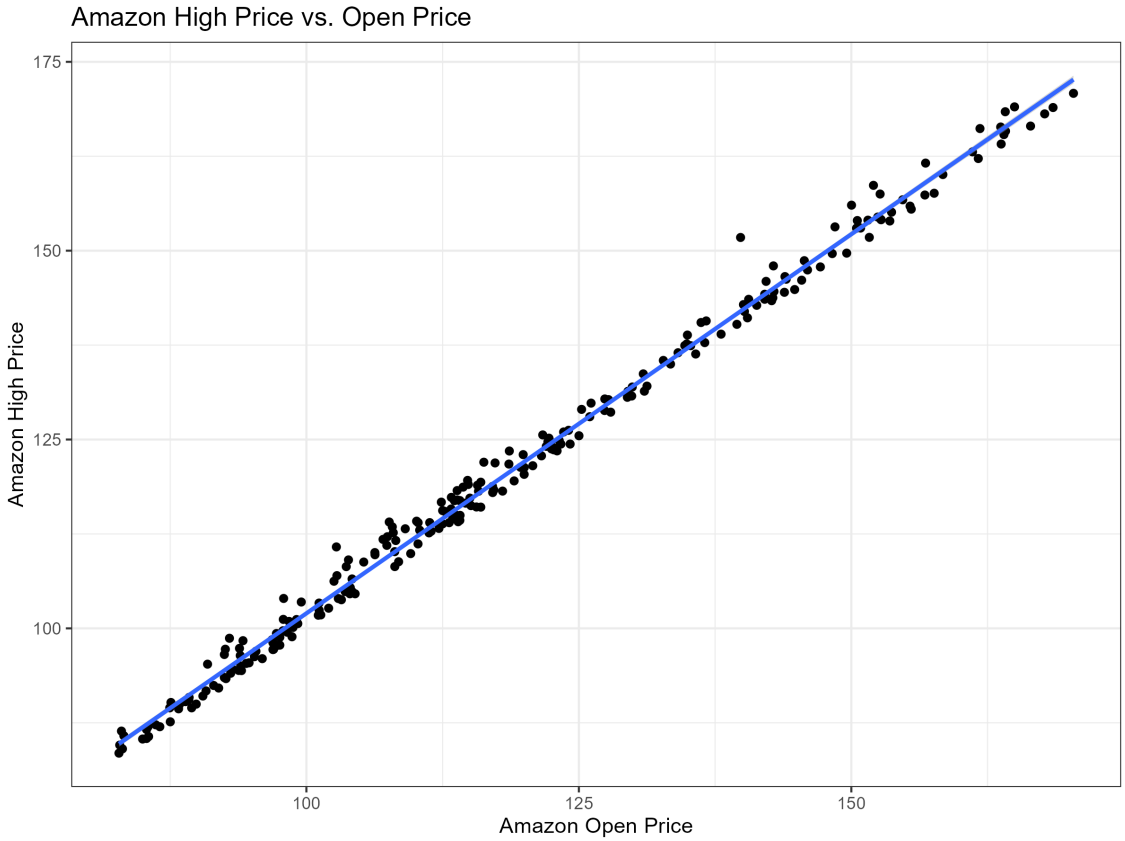
The *summary* function is the same for this model and the *plot* function will be configured only slightly differently:

```
# Generate plot of the model
plot_model <- ggplot(combined_data, aes(x = Open.amzn, y = High.amzn, color = Open.ixic)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Amazon High Price vs. Open Price with NASDAQ",
       x = "Amazon Open Price",
       y = "Amazon High Price") +
  theme_bw()
```

3 Plots and Summary

3.1 Amazon High Price vs. Open Price

When reviewing the plot of the simple linear regression model we created it is easy to see there is a strong correlation between the *Open Price* and the daily *High Price* of *Amazon*. Where Open Price is rising the daily high price is also rising consistently.



Residuals:

Min	1Q	Median	3Q	Max
-2.2163	-1.2698	-0.5290	0.8773	9.7695

Coefficients:

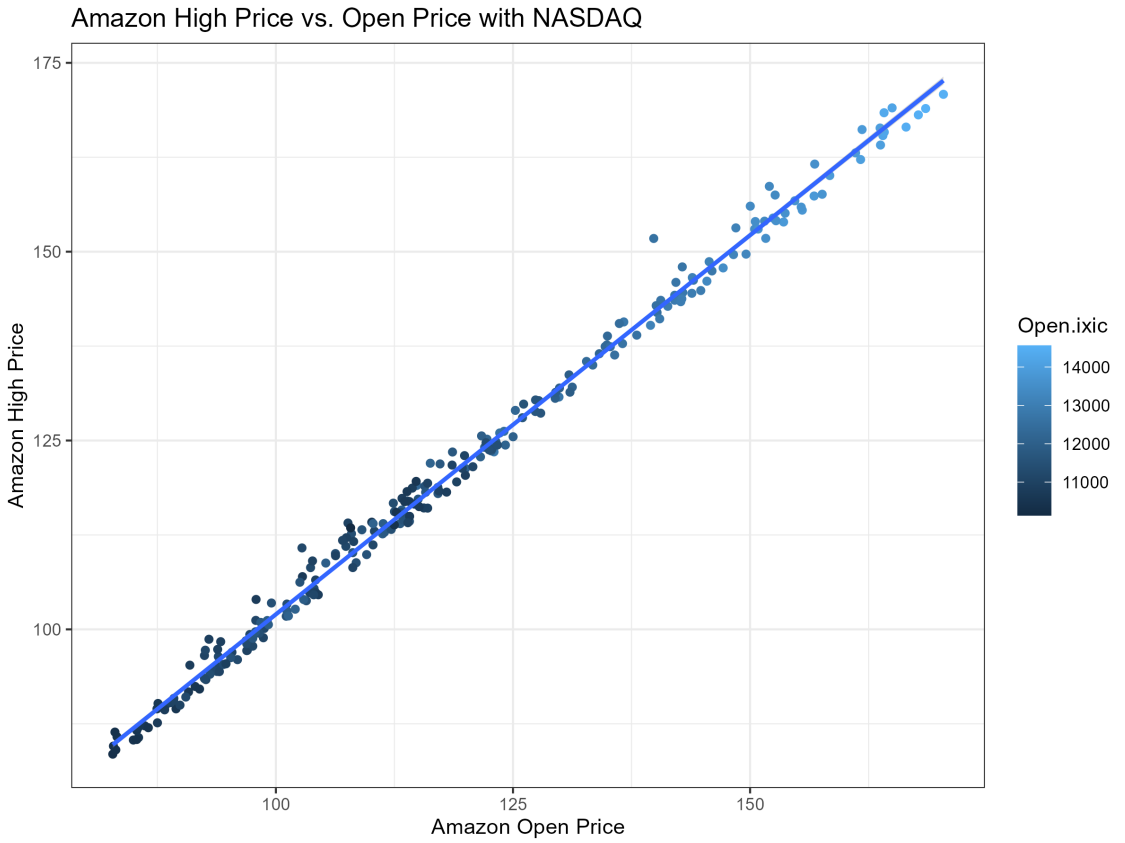
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.557136	0.572400	2.72	0.00698 **
Open	1.004183	0.004745	211.63	$< 2e^{-16}$ ***

Residual S.E.	Mult. R^2	Adj. R^2	F-statistic	p-value:
1.692 on 249 D.o.F.	0.9945	0.9944	4.479e+04 on 1 and 249 D.o.F.	$< 2.2e^{-16}$

With a P-Value $< 2.2e^{-16}$. It is very unlikely that this model has arrived at this correlation accidentally. On days where the Amazon open price is increasing it is also very likely that the daily high price of Amazon will also be rising.

3.2 Amazon High Price vs. Amazon Open Price & NASDAQ Composite Open Price

It is difficult to see if the addition of the NASDAQ Composite Open Price is having any influence on the accuracy of the model. There is no visible change from the simple linear regression model.



Residuals:

Min	1Q	Median	3Q	Max
-2.4555	-1.2411	-0.3657	0.8351	9.6796

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.5983806	1.7197801	3.255	0.00129 **
Open.amzn	1.0288681	0.0111814	92.016	$< 2e^{-16}$ ***
Open.ixic	-0.0005881	0.0002388	-2.463	0.01446 *

Residual S.E.	Mult. R^2	Adj. R^2	F-statistic	p-value
1.674 on 247 D.o.F.	0.9946	0.9946	2.278e+04 on 2 and 247 D.o.F.	$< 2.2e^{-16}$

In combination with the plot, the summary does not reveal too much of difference in the accuracy of the Multiple Linear Regression model compared to the Simple Linear Regression model created before.

4 Conclusion

Linear Regression

We may now complete the Linear regression model equation we began before. Remember in this equation Y is the dependent variable (the Amazon High Price), X is the independent variable (The Amazon Open Price) m is the estimated slope, and b is the estimated intercept. This is completed by the formula:

$$High.amzn = 1.004183 \cdot Open.amzn + 1.557136$$

The p-values from the summary indicate the statistical significance of each predictor variable. In this case, the p-value of the Open variable is less than 0.05, which suggests that it is a statistically significant predictor of the response variable.

The R-squared value is 0.9945, indicating that 99.45% of the variability in the daily **High Price** can be explained by **Open Price**. The adjusted R-squared value is 0.9944, takes into account the number of predictor variables used in the model. The residual standard error is 1.692, which tells us the standard deviation of our data points. Most interesting the F-statistic is $4.479e^4$, which tests the overall significance of the model. The p-value for the F-statistic is less than 0.05, indicating that the model is statistically significant. We now have a benchmark to say comparatively how significant is Multiple Linear Regression model.

Multiple Linear Regression

Like before we can also finish the Multiple Linear Regression equation:

$$High.amzn = 5.598 + 1.029 \cdot Open.amzn - 0.001 \cdot Open.ixic + \varepsilon$$

The Adjusted R-squared value of 0.9946 indicates that 99.46% of the variability in the daily **High Price** can be explained by **Amazon Open Price** and **NASDAQ Composite Open Price** within this model.There is small improvement compared to the accuracy of the simple linear model, but there are a few things to consider. The NASDAQ Composite Open Price contains the Amazon Open Price. This means that some small part of this predictor variable *contains* the other. Additionally the other stock symbols which make up the NASDAQ Composite may cause it to increase or decrease to a degree not reflected purely by the performance of Amazon itself. Which means that in part, the composite price may go up or down due to factors not related to Amazon.

5 Bibliography

Pressman, R.S. & Maxim, B.R. Software Engineering: A Practitioner's Approach, 9th Ed. New York, NY: McGraw-Hill Education, 2020.

Stamer, J. StatQuest, Linear Regression, Clearly Explained University of North Carolina, Chapel Hill. statquest.org 2023