STATS 506 Final Project Report:

Stock Daily Return for Top 10 US Companies by Market Capitalization

1 Introduction

In our final project, we obtained the *20 Years of End-of-Day Stock Data for Top 10 US Companies by Market Capitalization*, called **stock data** in short, from Amazon Web Service (AWS) data product, and *US Market Information Factor Data*, called **factor data** in short, from 202009 CRSP database, and explored these two sets of data using `dplyr`, `data.table` tools and Bootstrap/permutation test method in R.

1.1 Components of Data

The **stock data** are consisted of 10 csv files with the name format like "daily_adjusted_STOCKNAME". This dataset contains 20 years of historical data for the top 10 US stocks by market capitalization as of September 5, 2020. The dataset contains the following 10 symbols (in alphabetical order):

Stock Name	Company	Data Start Date	Data End Date	Number of Sample
AAPL	Apple Inc.	2000-01-03	2020-09-01	5200
AMZN	Amazon, Inc.	2000-01-03	2020-09-01	5200
BRK-A	Berkshire Hathaway Inc.(Class A)	2000-01-03	2020-09-01	5200
FB	Facebook, Inc.	2012-05-18	2020-09-01	2086
GOOG	Alphabet Inc.	2014-03-27	2020-09-01	1621
JNJ	Johnson & Johnson	2000-01-03	2020-09-01	5200
MA	Mastercard Incorporated	2006-05-25	2020-09-01	3593
MSFT	Microsoft Corporation	2000-01-03	2020-09-01	5200
V	Visa Inc.	2008-03-19	2020-09-01	3137
WMT	Walmart Inc.	2000-01-03	2020-09-01	5200

There are 9 variables in each csv file, which are timestamp, open, high, low, close, volume, adjusted close, dividend payout and split ratio. The detailed meanings of these variables are shown as below.

Variable	Definition
timestamp	Date in the format of "2020-09-01" or "09/01/20"
open	As-traded opening price for the day
high	As-traded high price for the day
low	As-traded low price for the day
close	As-traded close price for the day
volume	Trading volume for the day
adjusted_close	Split & dividend adjusted closing price of the day
split ratio	Ratio of new number of shares to old on the effective date
dividend	Cash dividend payout amount to stockholder

The **factor** data records 5 variables, Date, Mkt-RF, SMB, HML and RF, from July, 1926 to September, 2020. And the detailed meaning of these variables are shown in the table below.

Variable	Definition
Date	Date in the format of "202009"
Mkt-RF	Average market risk rate minus risk-free rate in one month (market risk premium)
SMB	Average historic excess returns of small-cap companies over large-cap companies in one month (size premium)
HML	Average historic excess returns of value stocks with high book-to-price ratio over growth stocks with low book-to-price ratio in one month (value premium)
RF	Average risk-free Rate in one month

1.2 Variables of Interest

Especially, we were interested in two variables in **stock** data, volume and adjusted_close. For these two variables, we constructed other six variables to explore further, including Price.Increase, Vol.Increase, Return, Price.ifIncrease, Vol.ifIncrease and Vol.ifIncrease.1. The following table shows the meanings of these six variables and equations used to calculate them.

Variable	Definition	Calculating Equation
Price.Increase	Daily increase of adjusted_close	$\text{Price.Increase}_t = \text{adjusted\_close}_t - \text{adjusted\_close}_{t-1}$
Vol.Increase	Daily increase of trading volume	$\text{Vol.Increase}_t = \text{volume}_t - \text{volume}_{t-1}$
Return	Daily Return	$\text{Return}_t = \frac{\text{adjusted\_close}_t - \text{adjusted\_close}_{t-1}}{\text{adjusted\_close}_{t-1}}$
Price.ifIncrease	Indicator whether the adjusted_close price is increasing	1 if $\text{Price.Increase}_t > 0$; 0 otherwise
Vol.ifIncrease	Indicator whether the trading volume is increasing	1 if $\text{Vol.Increase}_t > 0$; 0 otherwise
Price.ifIncrease.1	Indicator whether the yesterday's price is increasing	1 if $\text{Price.Increase}_{t-1} > 0$; 0 otherwise

1.3 Data Manipulation

Firstly, we merged 10 csv files containing **stock** data information to one dataset and added one column called "Company" taking values from the 10 stock name (AAPL, AMZN, BRK-A, FB, GOOG, JNJ, MA, MSFT, V, WMT), which was extracted from the csv file name "daily_adjusted_STOCKNAME.csv" using `gregexpr` function in R.

Next, we calculated the variables of interest in section 1.2 for each stock csv file and choose these columns and the timestamp and Company column as our new **stock** data.

Finally, to merge the **stock** and **factor** two datasets, we split the variable "timestamp" in **stock** to three variables "Year", "Month" and "Day", and the variable "Date" in **factor** to two variables "Year" and "Month". The variable, timestamp, with format "2020-09-01" was read as the Date format for some stock csv files, and for these files, we just split the timestamp using `format` function to extract corresponding part of the date; but timestamp with format "09/01/20" was read as the character format for some csv files, we wrote a function called `char2date` to transform and extract needed part of the date. For the date in **stock** data, we just treated it as string and used `substring` to extract year and month. And we finally merged the **stock** data and the **factor** based on the variables, Year and Month.

Our final data set called stock which contains all information we needed in this project is shown as below.

Year	Month	Day	Price.Increase	Vol.Increase	Return	Price.ifIncrease	Vol.ifIncrease	Price.ifIncrease.1	Company	Mkt.RF	SMB	HML	RF
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
2000	1	31	0.0164	2485100	2.0918367	1	1	0	AAPL	-4.74	5.62	-1.46	0.41
2000	1	28	-0.0646	742900	-7.6125383	0	1	0	AAPL	-4.74	5.62	-1.46	0.41
2000	1	27	-0.0015	-241200	-0.1764498	0	0	0	AAPL	-4.74	5.62	-1.46	0.41
2000	1	26	-0.0159	-1160600	-1.8360277	0	0	1	AAPL	-4.74	5.62	-1.46	0.41
2000	1	25	0.0463	502400	5.6484080	1	1	0	AAPL	-4.74	5.62	-1.46	0.41
2000	1	24	-0.0390	-491500	-4.5417492	0	0	0	AAPL	-4.74	5.62	-1.46	0.41
2000	1	21	-0.0169	-11921500	-1.9301051	0	0	1	AAPL	-4.74	5.62	-1.46	0.41
2000	1	20	0.0535	11013300	6.5077241	1	1	1	AAPL	-4.74	5.62	-1.46	0.41
2000	1	19	0.0202	1236300	2.5190173	1	1	1	AAPL	-4.74	5.62	-1.46	0.41
2000	1	18	0.0270	614300	3.4843206	1	1	1	AAPL	-4.74	5.62	-1.46	0.41

2 Scientific Questions & Methodology

Based on the data we obtained and processed in section 1, we proposed three questions on the correlation between price increase and volume increase, mean daily return, and two financial linear models respectively.

2.1 Relationship between Price Increase and Trading Volume Increase

Trading volume is the total number of shares of a stock that were traded during a given period of time. In finance, trading volume can legitimize the price action of a stock and investors usually use trading volume to decide either buy or sell the stock.[1]

In this problem, we aimed to explore **whether there exists a linear relationship between the trading volume increase and stock price increase by calculating the correlation coefficient between these two variables**. We calculated three Pearson correlation coefficients (ρ) and their confidence intervals here:

- ρ between Price.ifIncrease and Vol.ifIncrease

  We aimed at exploring the correlation between whether price increases or not and whether volume increase or not rather than their values through calculating this ρ.

- ρ between Price.ifIncrease.1 and Vol.ifIncrease

  There might exist a delay effect of price increase on volume increase because due to yesterday's price increase, investors might tend to buy or sell more stocks today. We calculated this ρ to see if there is any relation between yesterday's price increase/decrease and today's volume increase/decrease.

- ρ between absolute value of Price.Increase and absolute value of Vol.Increase

  Sharp increase or decrease of price might cause that investors buy or sell large amount of stocks. We aimed to examine this proposed relationship by calculating this ρ.

For **confidence intervals**, we used the equation for confidence interval of Pearson's correlation below[2]:

$$\rho_L = \frac{\exp\left(2(z_\rho - z_{1-\alpha/2}/\sqrt{n-3})\right) - 1}{\exp\left(2(z_\rho - z_{1-\alpha/2}/\sqrt{n-3})\right) + 1},\ \rho_U = \frac{\exp\left(2(z_\rho + z_{1-\alpha/2}/\sqrt{n-3})\right) - 1}{\exp\left(2(z_\rho + z_{1-\alpha/2}/\sqrt{n-3})\right) + 1},$$

where ρ_L is the lower bound of confidence interval, ρ_U is the upper bound of confidence interval for the correlation coefficient ρ, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of standard normal distribution, and z_ρ is the transformation of correlation coefficient ρ with $z_\rho = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$. We chose to calculate 95% confidence interval, so $\alpha = 0.05$ here.

All these calculations are accomplished with `data.table` and `dplyr` tools in R.

2.2 Mean Daily Return for Each Year

If the increase of a stock is steady, we expect that the mean daily return of this stock for every year will be similar. In this problem, we tried to observe **the trend for mean daily return for these 10 stocks over years, in other words, whether the mean daily return for each year is steady or fluctuate dramatically**. We calculated the mean daily returns for every stock and for every year in `data.table`.

We also calculated the **confidence intervals** for these mean daily returns. We chose to use **bootstrapping percentile method** (we set boostrapping times as 1000) to estimate the confidence interval, because there are approximately 250 records of returns in one year and the number of samples might not be large enough to use the equation below for the sample mean to calculate the confidence interval,

$$\Big(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\Big),$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of standard normal distribution, s is the sample standard deviation and n is the sample size.

2.3 Financial Linear Models

In this problem, we tried to **fit two classic models for returns in finance**, which are CAPM model and Fama and French three factor model.

2.3.1 CAPM Model

The Capital Asset Pricing Model (CAPM) describes the relationship between systematic risk and expected return for assets, particularly stocks. The formula for calculating the expected return of an asset given its risk is as follows:

$$ER_i - R_f = \beta_i(ER_m - R_f),$$

where ER_i is the expected return of stock, R_f is the risk-free rate, ER_m is the expected return of the total market and β_i is the beta of the stock, a measure of how much risk the stock will add to a portfolio that looks like the market. If a stock is riskier than the market, it will have a beta greater than 1. If a stock has a beta of less than 1, the formula assumes it will reduce the risk of a portfolio.[3]

We calculated β_i for each stock using `lm` function by taking the mean daily return for one month minus the risk-free rate in this month as Y and the market rate minus risk-free rate in this month as X. And we found confidence interval of β_i using **bootstrapping percentile method** (we set bootstrapping times as 500) to **compare the risk of ith stock relative to the stock market**. And to **examine whether the risks for these 10 stocks are the same**, we performed **permutation test**. We totally did $\binom{10}{2} = 45$ permutation tests to examine each pair of β's from these 10 stocks. For each permutation test, our null hypothesis is

$$H_0: \ \beta_i = \beta_j, \ i, j \in \{AAPL, AMZN, ..., WMT\}, \ i \neq j.$$

2.3.2 Fama and French Three Factor Model

The Fama and French Three-Factor Model (or the Fama French Model for short) is an asset pricing model developed in 1992 that expands on the capital asset pricing model (CAPM) by adding size risk and value risk factors to the

market risk factor in CAPM.The formula is:

$$ER_i - R_f = \alpha_i + \beta_{1i}(ER_m - R_f) + \beta_{2i}SMB + \beta_{3i}HML,$$

where ER_i is the expected return of ith stock, R_f is risk-free rate of return, ER_m is the expected return of the total market, SMB is the size premium and HML is the value premium.[4]

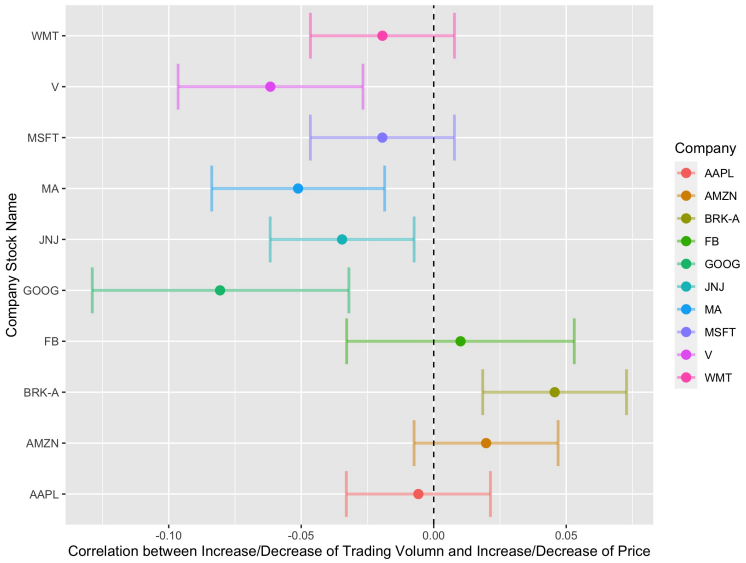
We aimed to find out **whether the market return, size premiumand value premium are significant for the return of stock, in other words, how well the three factor model works**. To achieve this goal, we calculated β_{1i}, β_{2i} and β_{3i} using `lm` function by taking the mean daily return for one month minus the risk-free rate in this month as Y, the market rate minus risk-free rate in corresponding month as X_1, the size premium in corresponding month as X_2 and the value premium in corresponding month as X_3 and the confidence intervals for β_{1i}, β_{2i} and β_{3i} using `confint` function, and then judged the significance of market return, size premium and value premium by seeing whether the confidence interval contains 0.

3 Results & Analysis

We showed the results of our solutions to three problems by printing the calculating results stored in `data.table` and plotting corresponding graphs by using `ggplot`. The number results in `data.table` are shown in Appendix.

3.1 Relationship between Price Increase and Trading Volume Increase

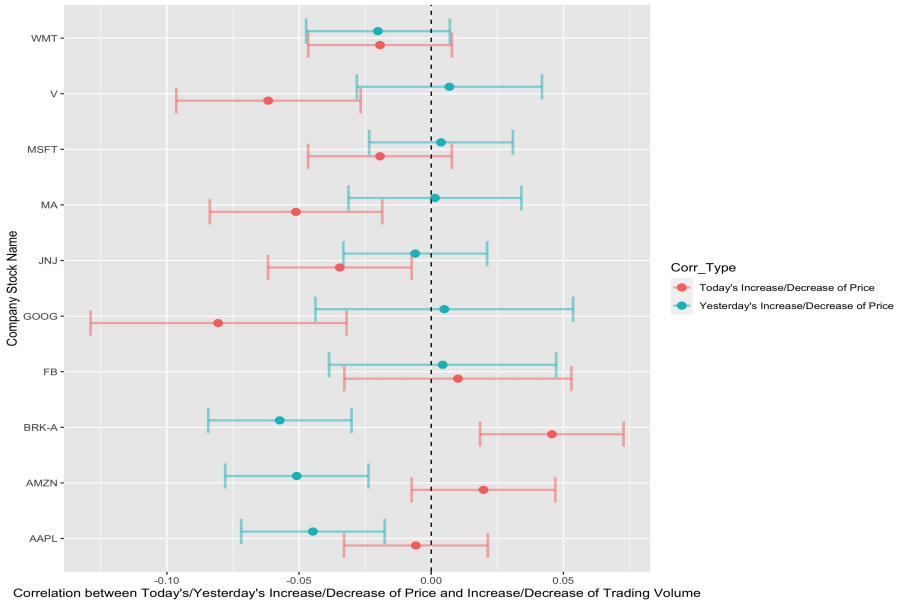
- ρ between Price.ifIncrease and Vol.ifIncrease



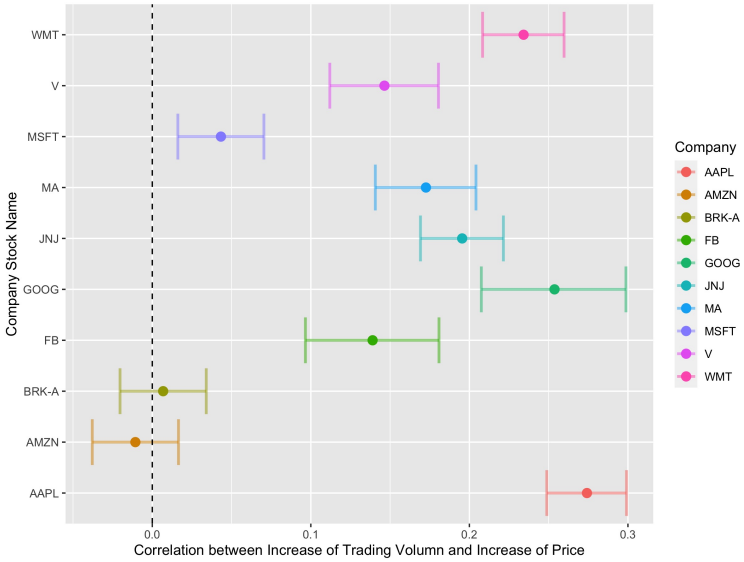
It can be seen that all correlation coefficients are in the level of 0.01, which are pretty small. Although some of the confidence intervals do not contain 0, the value is too small to think the Price.ifIncrease and Vol.ifIncrease positively/negatively linearly correlated.

- ρ between Price.ifIncrease.1 and Vol.ifIncrease

  The correlation coefficients become even much smaller and most of confidence intervals contain 0, which is not consistent to our guess in section 2.1. Price.ifIncrease.1 and Vol.ifIncrease are not significantly positively/negatively linearly correlated.



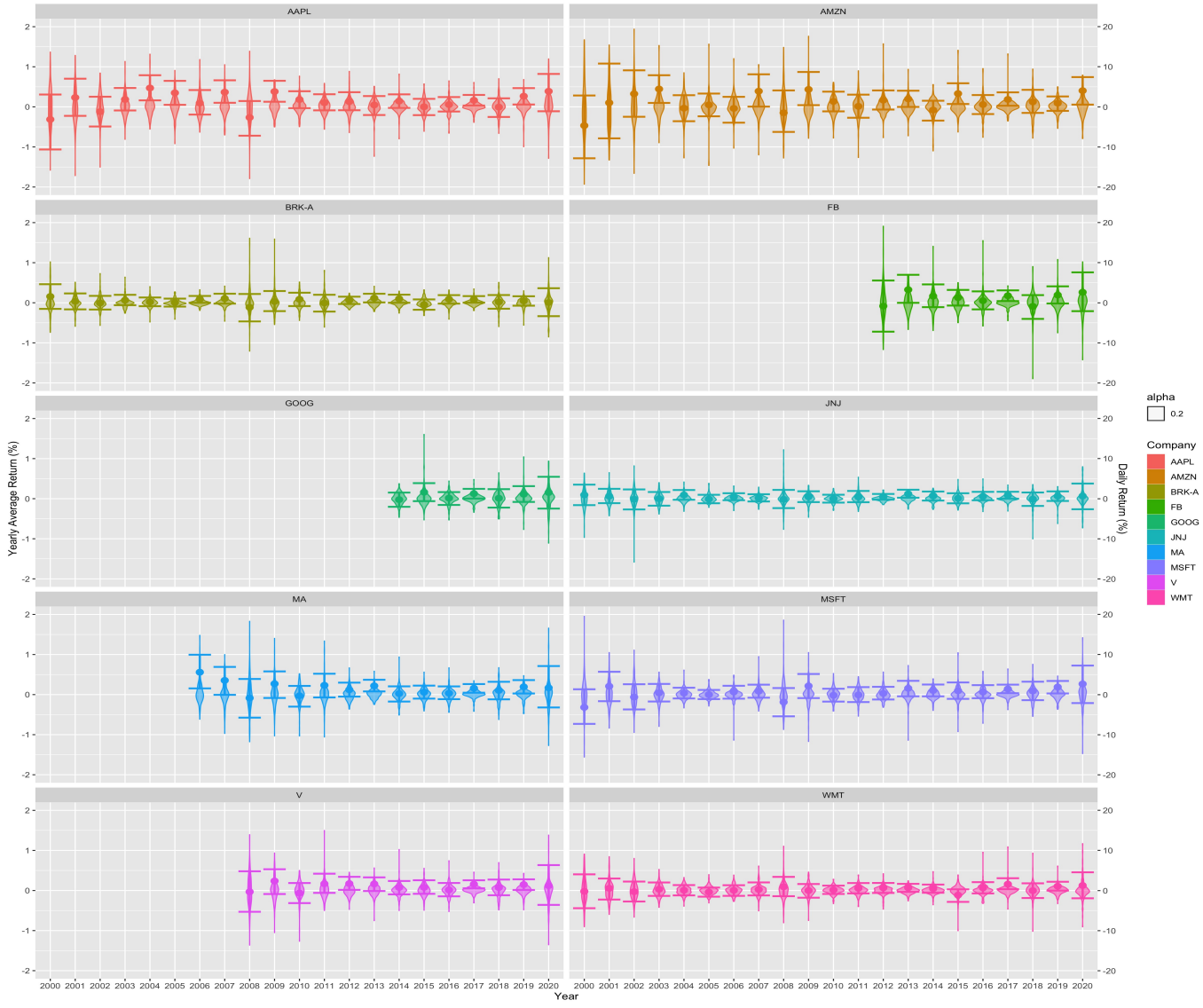
- ρ between absolute value of Price.Increase and absolute value of Vol.Increase



It can be found that there are a significantly positive linear correlation between the absolute value of Price.Increase and the absolute value of Vol.Increase for all stocks except AMZN and BRK-A.

3.2 Mean Daily Return for Each Year

We plotted the confidence interval with violin plot showing the distribution of the data in the following graph. As the graph shows, the mean daily return for each year does not fluctuate dramatically for every stock, in other words, the mean daily return for every year is steady. We also found that for the year with larger confidence interval for mean daily return, the distribution of daily return in this year are more dispersing; and on the contrary, the year with smaller confidence interval for mean daily return, the distribution of daily return in this year are more centralized. It makes sense statistically, because more dispersing distribution indicates larger variance which would lead to wider confidence interval.

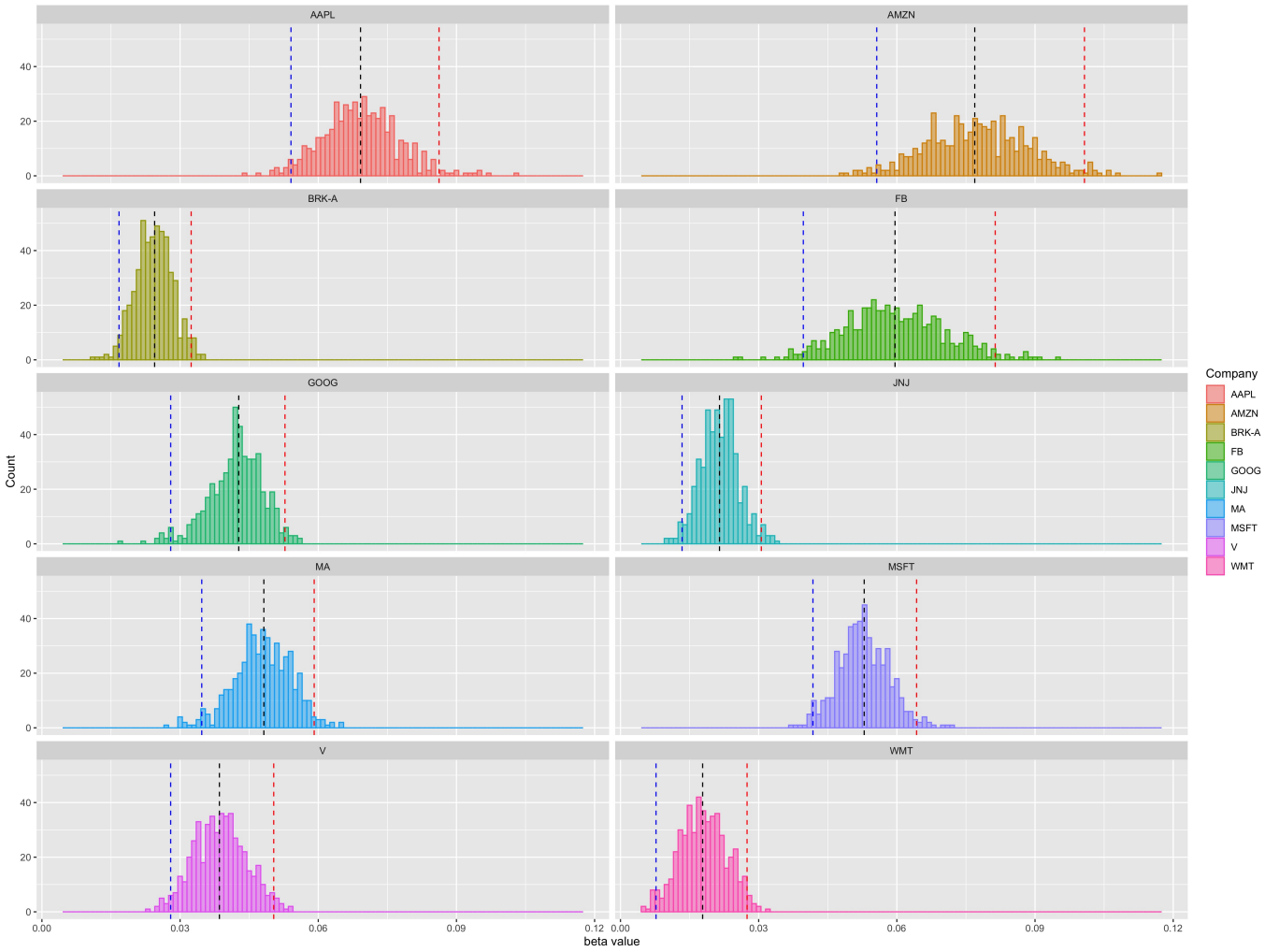


3.3 Financial Linear Models

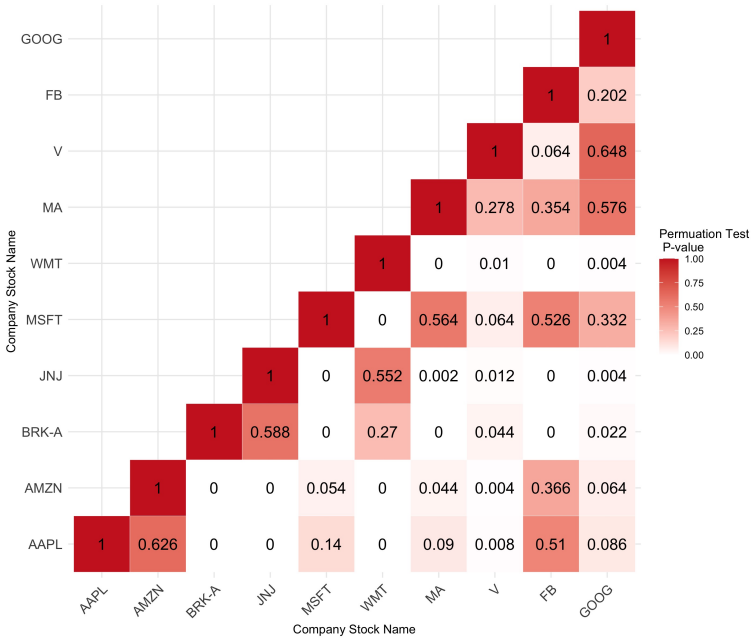
3.3.1 CAPM Model

We visualized β, which is the coefficient of linear regression for CAPM model for each stock, the result of our bootstrapping and 95% confidence interval using `ggplot`. (Blue line stands for lower bound of confidence interval, red line stands for upper bound of confidence interval and black line stands for the estimated value for β.)

As we expected, all β's for 10 stocks lie inside the confidence interval. Moreover, no matter how large or small the β is, all β's are larger than 0, which means that the expected daily return and the expected market return are positively related for these 10 stocks, in other words, if the market return increases, then the stock return increases. Then, the AMZN has the largest β value, so AMZN has the largest risk and as a result it has the largest return compared with other 9 stocks; similarly, WMT has the smallest β value, which means that WMT has the smallest risk and as a result the smallest return compared with other 9 stocks. Lastly, we found that the β distributions for AMZN and FB are flatter. We thought it might be caused by the more dispersing data for AMZN and FB.



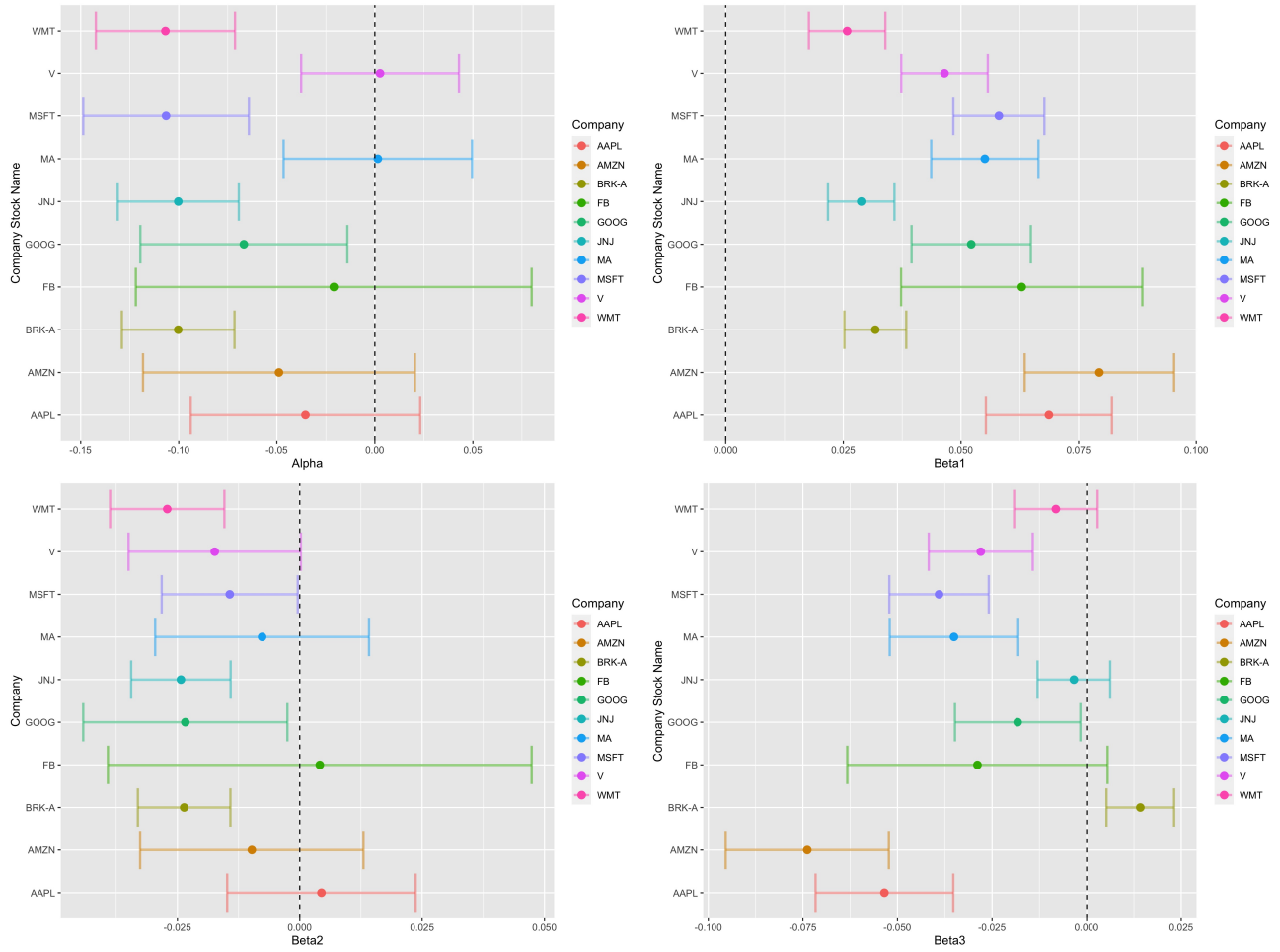
Observing the graph above, we found that some stocks' β's are pretty similar, so we performed the permutation test for these 10 stocks' β's. And our results are shown in the following graph. The number stands for the P-value of permutation test.



Based on the permutation test results, we cannot reject the null hypothesis that, $\beta_{AAPL} = \beta_{AMZN}$, $\beta_{AAPL} = \beta_{MSFT}$, $\beta_{AAPL} = \beta_{FB}$, $\beta_{AMZN} = \beta_{FB}$, $\beta_{BRK-A} = \beta_{JNJ}$, $\beta_{BRK-A} = \beta_{WMT}$, $\beta_{JNJ} = \beta_{WMT}$, $\beta_{MSFT} = \beta_{MA}$, $\beta_{MSFT} = \beta_{FB}$, $\beta_{MSFT} = \beta_{GOOG}$, $\beta_{MA} = \beta_{V}$, $\beta_{MA} = \beta_{FB}$, $\beta_{MA} = \beta_{GOOG}$, $\beta_{V} = \beta_{GOOG}$, $\beta_{FB} = \beta_{GOOG}$ under 10% significance (with 90% confidence). So, we can say that under 10% significance, the risks of two stocks in each of these pairs are the same. It can be roughly said that the companies of similar type tend to have similar risks. Take MA (Mastercard) and V(Visa) as example. It makes sense, to some extent, that their risks are pretty similar, because they are companies of the same type.

3.3.2 Fama and French Three Factor Model

We showed the point estimates for coefficients β_1, β_2, β_3 and their 95% confidence intervals for 10 stocks using `ggplot`.



None of 95% confidence intervals of β_1's for all 10 stocks contain 0. So, the market return is very significant for the return of stock for every stock. But confidence intervals of β_2's for AAPL, AMZN, MA, V and FB contain 0, which means that the size premium (SMB) is not significant for the return of these 5 stocks; on the contrary the size premium (SMB) is significant for the return of BKR-A, GOOG, JNJ, MSFT and WMT. Lastly, the confidence intervals of β_3's for WMT, JNJ and FB contain 0, so the value premium is not significant for the return of these 3 stocks; but the value premium is significant for the return of the other 7 stocks including AAPL, AMZN, BRK-A, GOOG, MA, MSFT and V. (α is the intercept of the linear regression. Here, we did not discuss about the significance of α.)

4 Conclusion

Based on the detailed analysis in section 3, it can be concluded that

- There are a significantly positive linear correlation between the absolute value of daily price increase and the absolute value of daily trading volume increase;
- The mean daily return for each year does not fluctuate dramatically for every of 10 stocks, which is pretty steady;
- According to CAPM model, every of 10 stocks' return increases as the market return increase; and AMZN has the largest risk among 10 stocks, WMT has the largest risk among 10 stocks. According to permuation test, the companies of similar type tend to have similar risks. According to Fama and French three factor model, market return is significant for the return of all 10 stocks, size premium is significant for the return of BKR-A, GOOG, JNJ, MSFT and WMT, and value premium is significant for the return of AAPL, AMZN, BRK-A, GOOG, MA, MSFT and V.

(All codes, data.tables and graphs are shown in the Appendix.pdf and code.ipynb.)

References

[1] M. F. Staff, "Does trading volume affect stock price?" 2016. [Online]. Available: https://www.fool.com/knowledge-center/does-trading-volume-affect-stock-price.aspx

[2] NCSS, "Confidence intervals for pearson's correlation." [Online]. Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Confidence_Intervals_for_Pearsons_Correlation.pdf

[3] W. Kenton, "Capital asset pricing model (capm)," 2021. [Online]. Available: https://www.investopedia.com/terms/c/capm.asp

[4] A. Hayes, "Fama and french three factor model," 2021. [Online]. Available: https://www.investopedia.com/terms/f/famaandfrenchthreefactormodel.asp