Stock Volatility Prediction

Dingyu Sun | Sharvil Turbadkar | Jian Jian | Rishabh Upadhyay

Dec/02/2020

Abstract

We investigated the long-term stock return of several individual companies including Apple Inc. (AAPL), Alphabet Inc. (GOOG), and Microsoft Corporation (MSFT), and created the time series model to predict the beta in the future. Also, we analyzed the relationship between the individual company stock return and the Dow Jones Industrial Average market return. Our project goal is to predict the beta value by building several time-series models and comparing the prediction with the beta value we computed from the CAPM formula. In this project, we used the backward elimination method to construct the model. The accuracy of these three models was not good enough, the adjusted R-square can be improved in the future by adding more features.

1.Introduction

1.1 Describe the purpose

Information and data play a central role in the current industry field. Companies related to the finance field use modeling techniques to compute and predict the market and companies' stock trends. "The stock market is a market that enables the seamless exchange of buying and selling of company stocks. Every Stock Exchange has its own Stock Index value. The index is the average value that is calculated by combining several stocks."[1]. The index can transfer the information of the whole stock market, and predict the future trend. The stock market has a huge impact on people, the company's development, the country's economy, and the pattern of world economic development. In that regard, predicting the stock trends based on the current situation can maximize the profit and minimize the risk.

There is a long time history of predicting the stock return, by predicting the stock return investors can revise their portfolio allocation, make a better investment decision, and understand the risk-return trade-off. Beta is an important measurement in the finance field and it can send substantial information to the investor. Beta measures the volatility of an individual stock in relation to the overall market. The beta of the individual stock measures the degree of its deviation from the market. A beta value larger than one represents the stock swings more than the market stock. If a stock's beta is less than one, it means that the stock has a lower movement than the market. In this project, we utilized different features to build the time series model and predict the beta value.

Another important concept is the risk premium, it is an expected return that investment return in excess of the risk-free rate of return. We used the quantmod package to download the finance data from Yahoo Finance. To get the respective risk premium, we imported the risk-free return data and substracted the difference between the market return and the risk free-return, in this project the risk premium helped us compute the beta value.

The dataset has 12 columns, the Dow Jones Industrial Average monthly market return, the Apple Inc. (AAPL), Alphabet Inc. (GOOG), Microsoft Corporation (MSFT) monthly market return, and their respective auto regression term. We also added the extended dataset of the 3-Month Treasury Bill to calculate the risk premium.

This project organized as follow, we built three time-series model. We used three models to predict the beta value and compare the predicted value with the true beta values which were computed by the CAPM formula. The first model is the full model with all features. We ruled out the insignificant variables first. The second model is a "time trend + autoregressive + seasonal category dummies" model with three lagged variables. Based on the result of the second model, the third model is a "time trend + autoregressive" model but with only one lagged variable. We applied backward elimination during the training process. The performance of the model improves every after each round of the feature evaluation.

1.2 Key findings of the project

- The adjusted R-square of all three models is less than 0.1, which means these three models do not have a good performance, thus to train a better model, we need to add more features into the model.
- From the result of the three models, we found that beta has a relationship with the time trend and did not show any seasonal pattern.

2. Dataset description

2.1 Document source and description

The data source: : Yahoo Finance& FRED

In this project, we used a finance package called "quantmod" to download the finance data from the website. By entering the symbol of the stock and website name also the data period, we can download the data from that website. In this project, we set the start date as 2010 and the end date as 2020, and download the monthly data from Yahoo Finance.

The final dataset contains data from a ten-year window. It contains 291 rows and 12 columns. It can be divided into 4 parts, the first part is the beta values of each company; the second part is the trends and the seasonal variables, and the autoregression variables; the third part is the returns of the market and companies; the last part is the dummy variables, we created two variables to represent three companies.

The true beta value was computed by using the CAPM formula. We computed the subtracting result of the market return and the risk-free return and the result of companies return and the risk-free return, after that we built a linear model to get the coefficient, and the coefficient

is the beta. And the CAPM believes no intercept in its model because the market will sell or buy the asset back to the CAPM model.

$$R_e - R_f = \beta * (R_m - R_f)$$

	1	2	3	4	5	6
beta	0.8068101	0.7877504	0.4482081	0.3996683	0.3602597	0.3601669
trend	1.0000000	2.0000000	3.0000000	4.0000000	5.0000000	6.0000000
spring	0.0000000	0.0000000	0.0000000	1.0000000	1.0000000	1.0000000
summer	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
fall	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
DJI	-0.0063121	0.0053312	0.0570253	0.0129900	0.0363870	0.0173208
return	-0.0177649	-0.0914429	-0.1447893	-0.0319337	0.0019546	-0.0003290
lag_1	-0.1086001	-0.0177649	-0.0914429	-0.1447893	-0.0319337	0.0019546
lag_4	0.0448219	0.0882004	0.0016960	-0.1086001	-0.0177649	-0.0914429
lag_12	-0.0558834	0.0595547	0.1268110	0.1874107	0.1044835	-0.0267695
GOOG	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
AALP	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

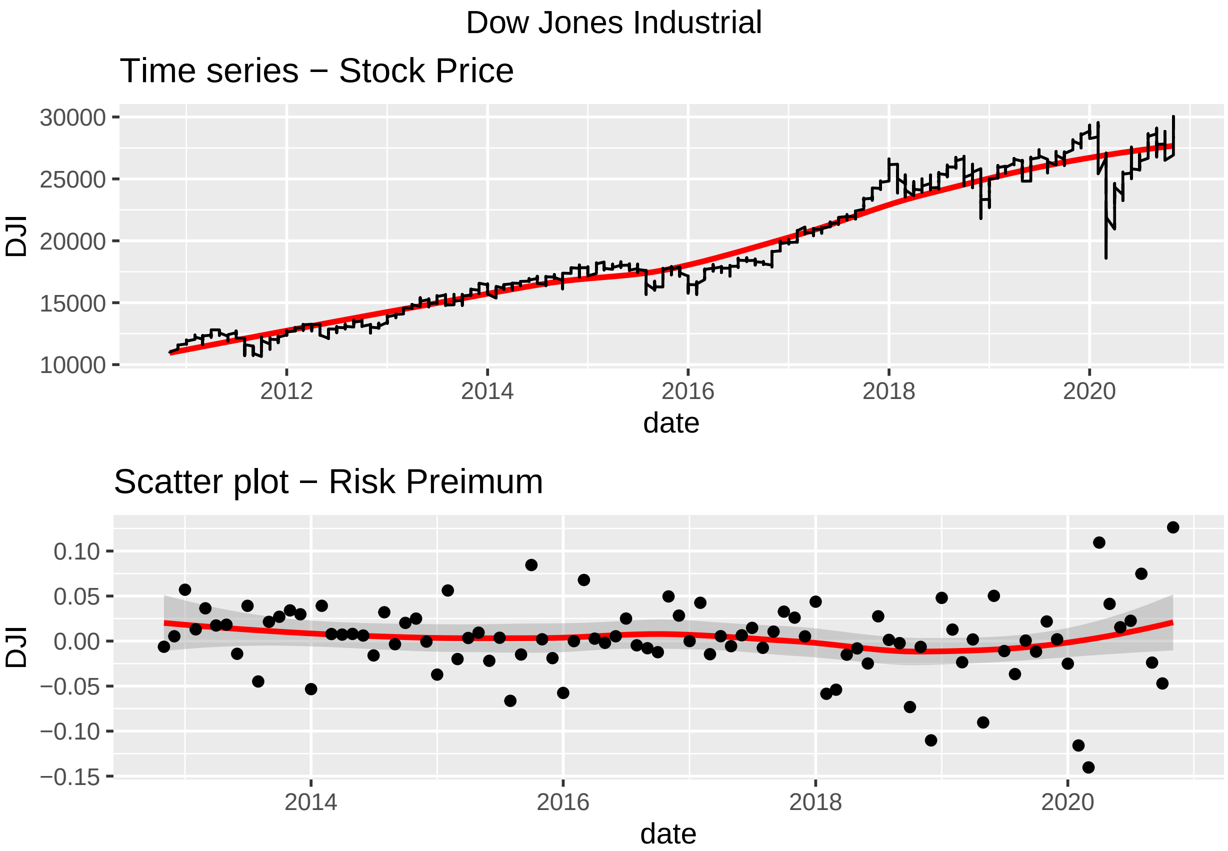
2.2 Time span and number of observations

The three lagged variables included the previous month, the previous 4 months, and the previous year. The three seasonal categorical variables included spring, summer, and fall.

2.3 Time series plot and scatter plots to indicate primary relationships existing in your data.

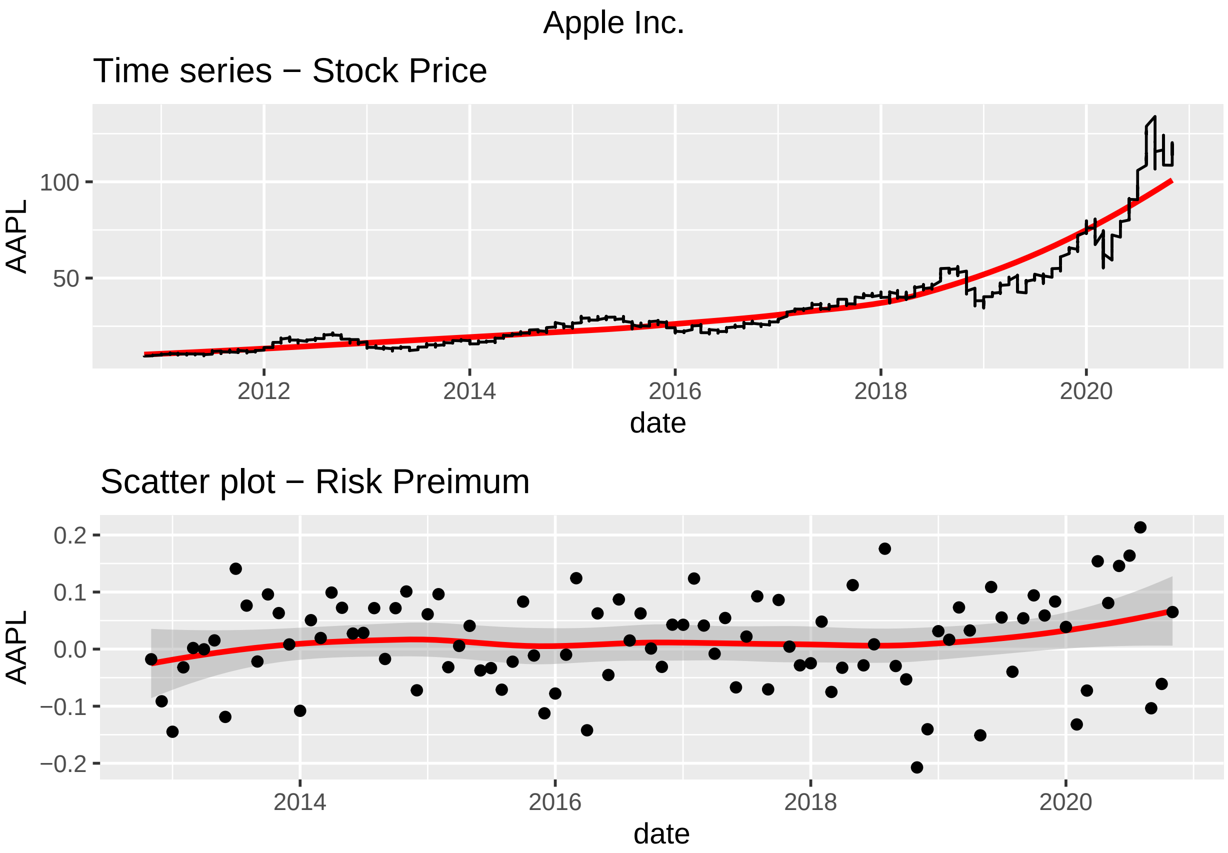
2.3.1 Time series& scatter plot of the Dow Jones Industrial

The time series plot shows a daily positive relationship to the market price of the Dow Jones Industrial. The relationship between the date and the risk premium seems flat.



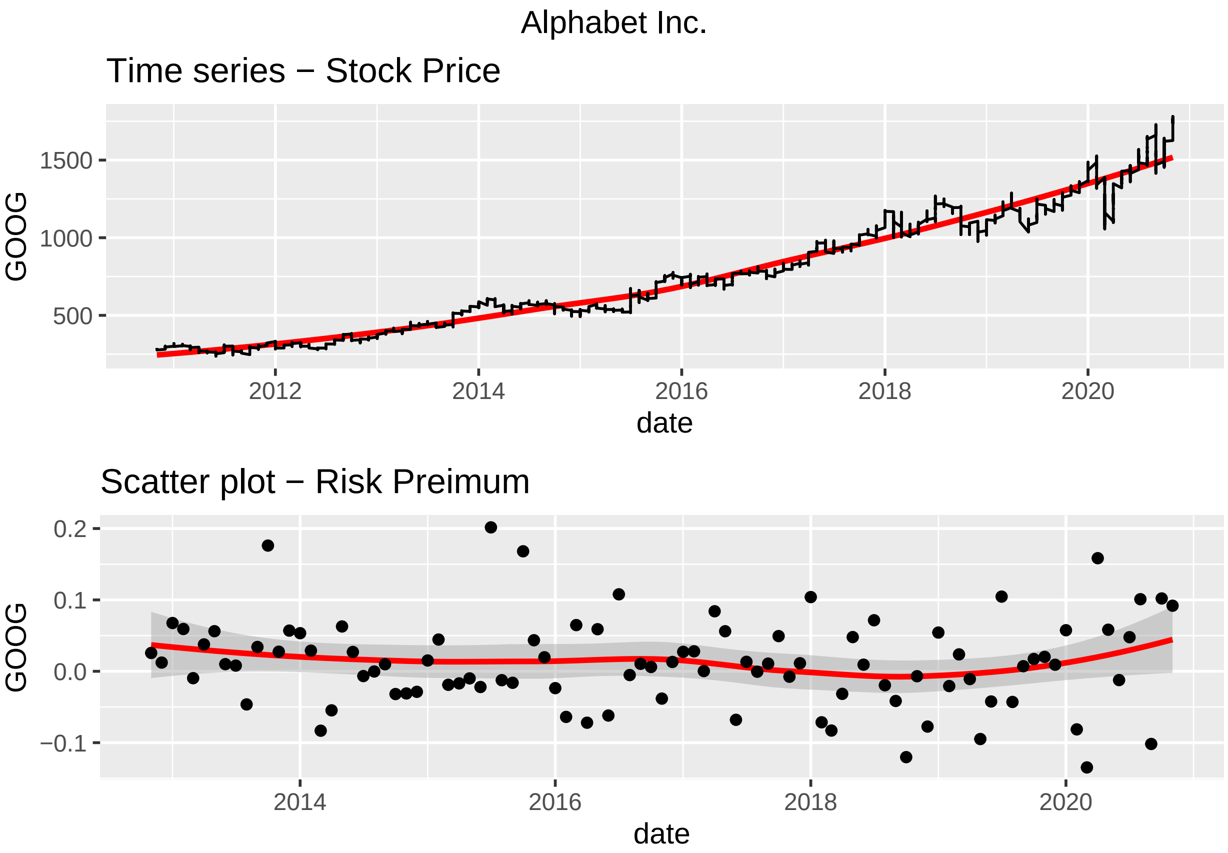
2.3.2 Time seriess& scatter plot of Apple Inc.

Compared with the Dow Jones Industrial plot, the time series plot of Apple Inc like a quadratic plot. With the date increasing, the stock price of Apple Inc increased at a fast speed. The scatter plot does not show much difference from the previous one.



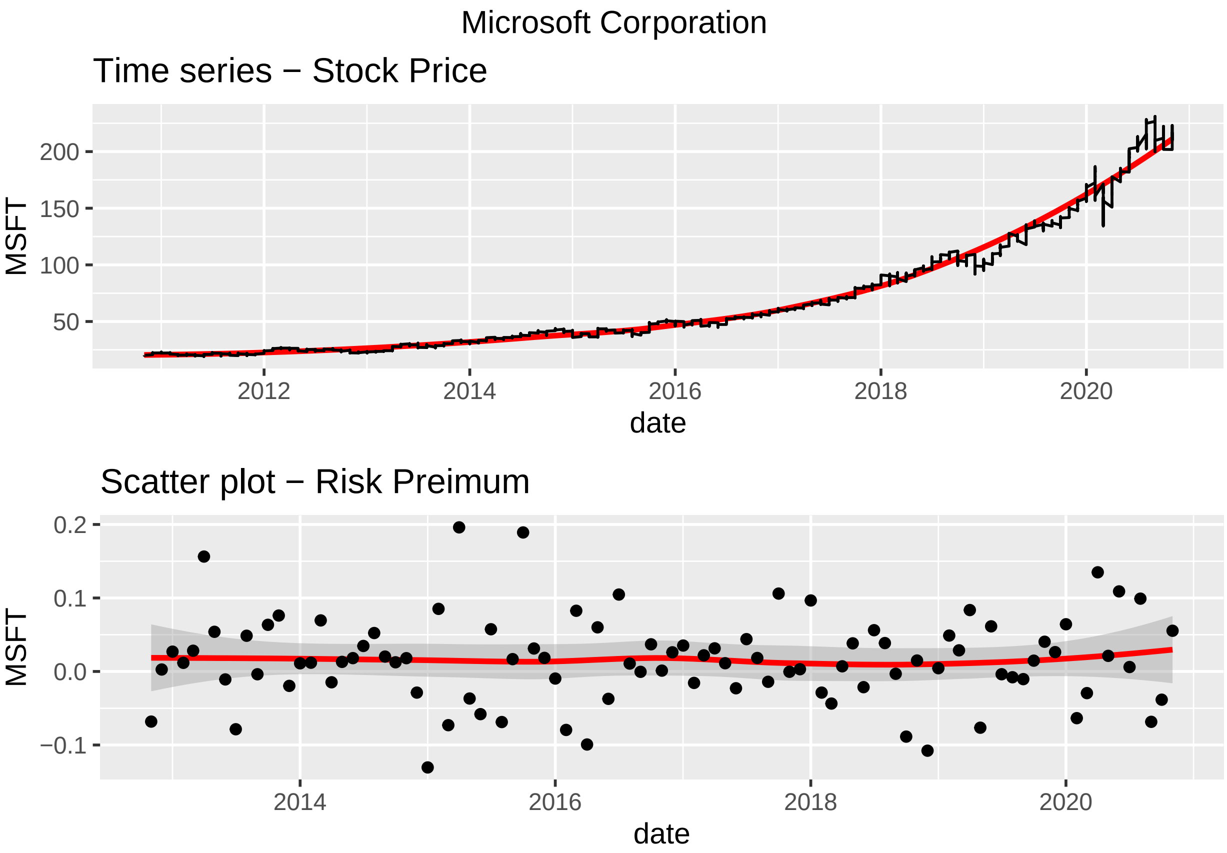
2.3.3 Time seriess& scatter plot plot of the Alphabet Inc.

The time series plot shows that the date seemingly has a linear relationship with the stock price of Alphabet Inc, and the point spread of the scatter plot is more spread than the previous two plots.



2.3.4 Time series plot of the Microsoft Corporation.

The trend of the time series plot is similar to Apple Inc, and the variation of the scatter plot is more spread than the previous two plots.



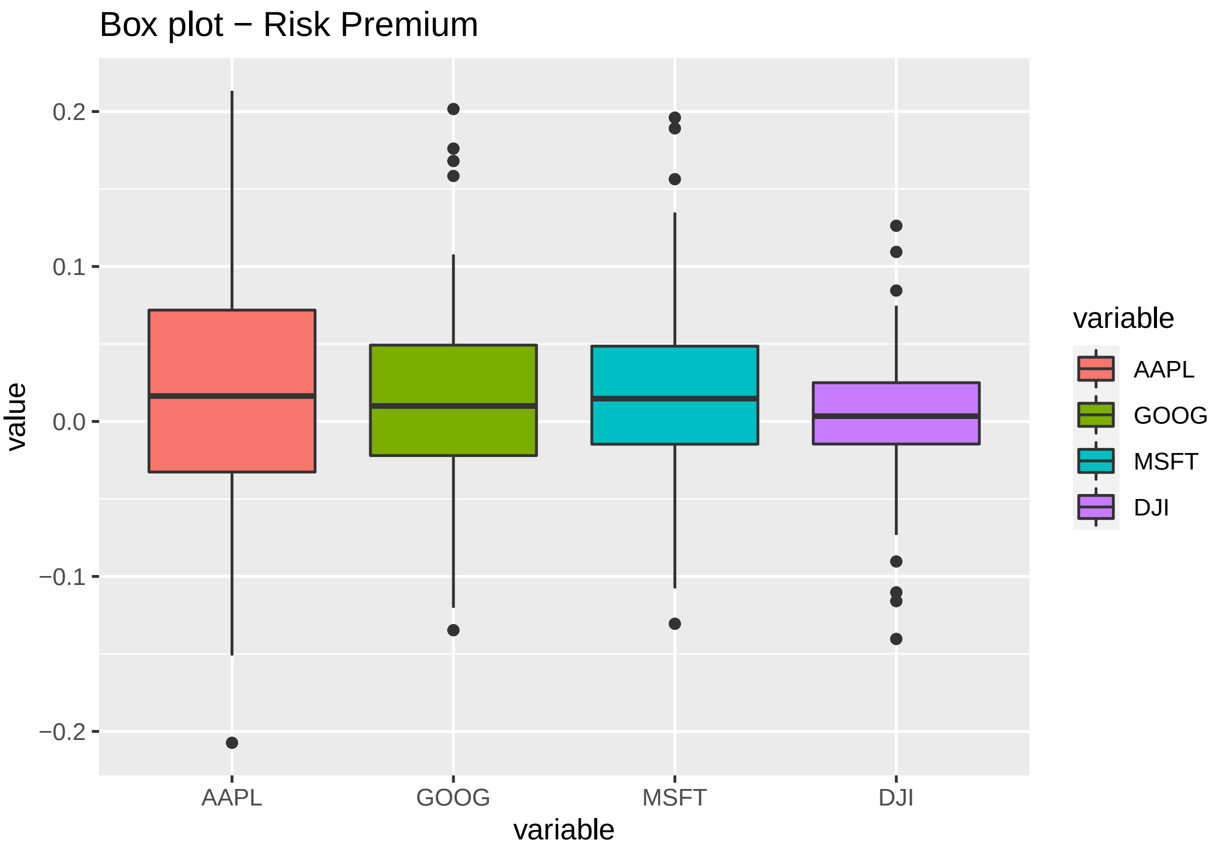
Microsoft Corporation

Time series − Stock Price

Scatter plot − Risk Preimum

2.4 Summary statistics(mean, standard deviation and correlation)

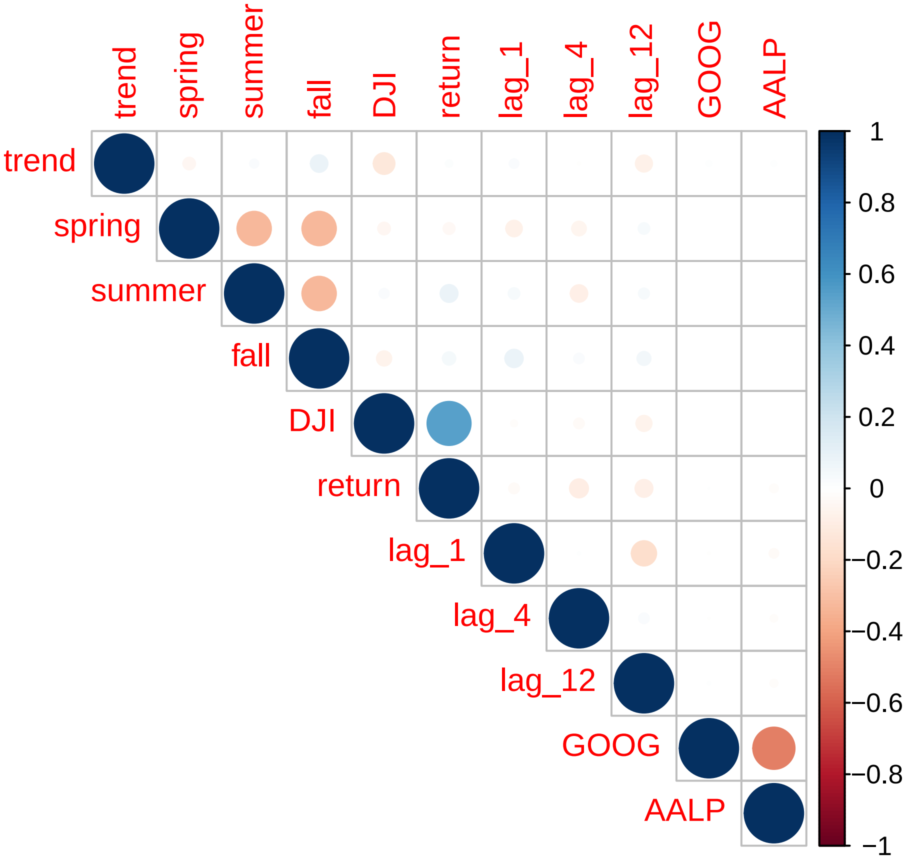
2.4.1 Risk Premium distribution



Comparing with other companies, Apple Inc has the largest risk premium distribution, and Microsoft Corporation has the smallest risk premium distribution, which means Apple Inc might have the highest risk.

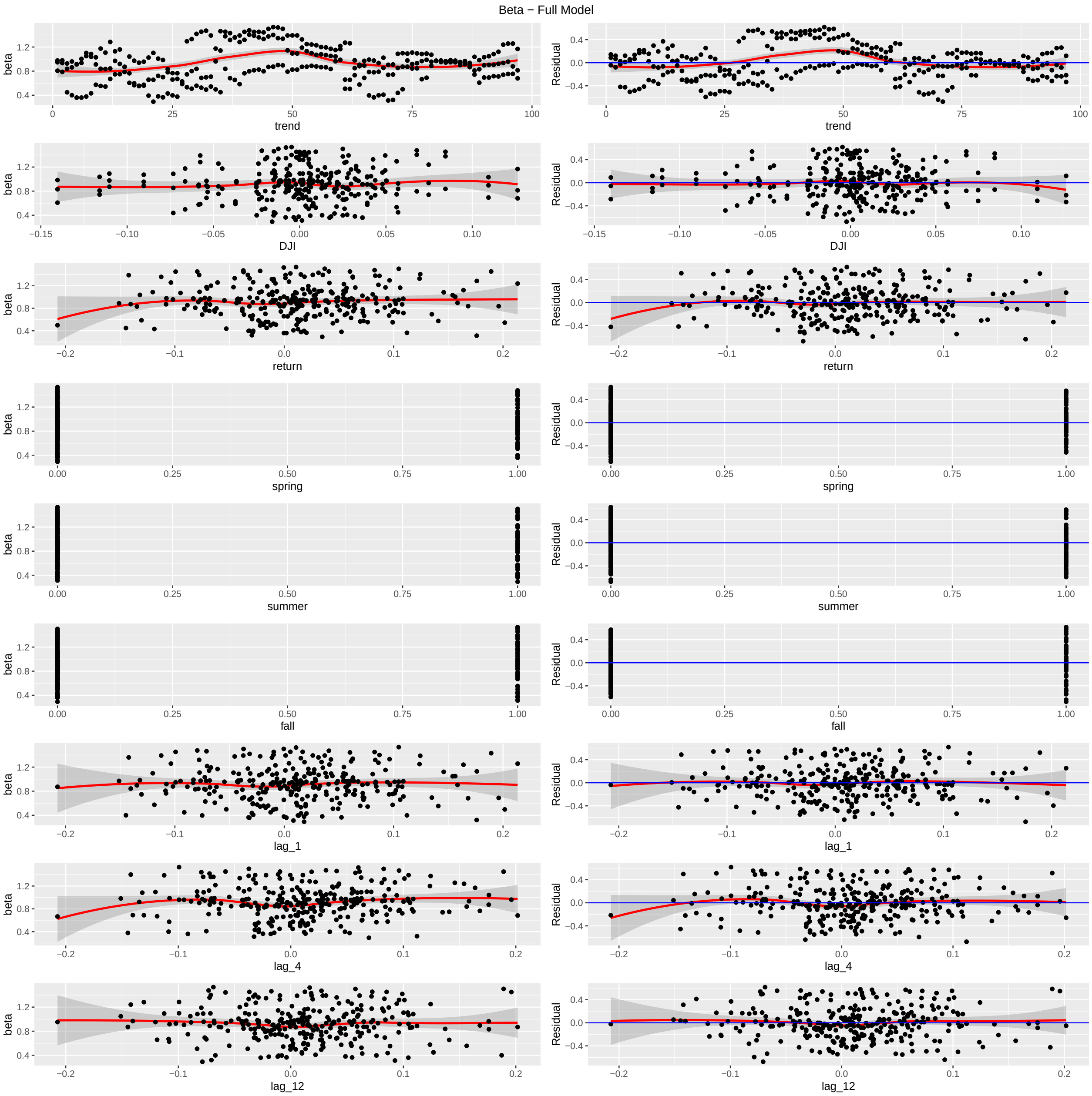
2.4.2 Correlations

From the correlation plot, we learned that the Dow Jones Industrial has a positive correlation with the three companies' returns. The cross-point of that pair shows dark blue. Summer and the three companies return have a positive, but their relationship is not as strong as the previous one, and their cross-point shows light blue.



2.4.3 Scatter plot of the independent variables vs dependent variables

From the scatter plot we can see that relationship between "trend" and the beta is like a quadric, and the points of other plots spread randomly that no pattern exists in the graph.



3. Model Selection

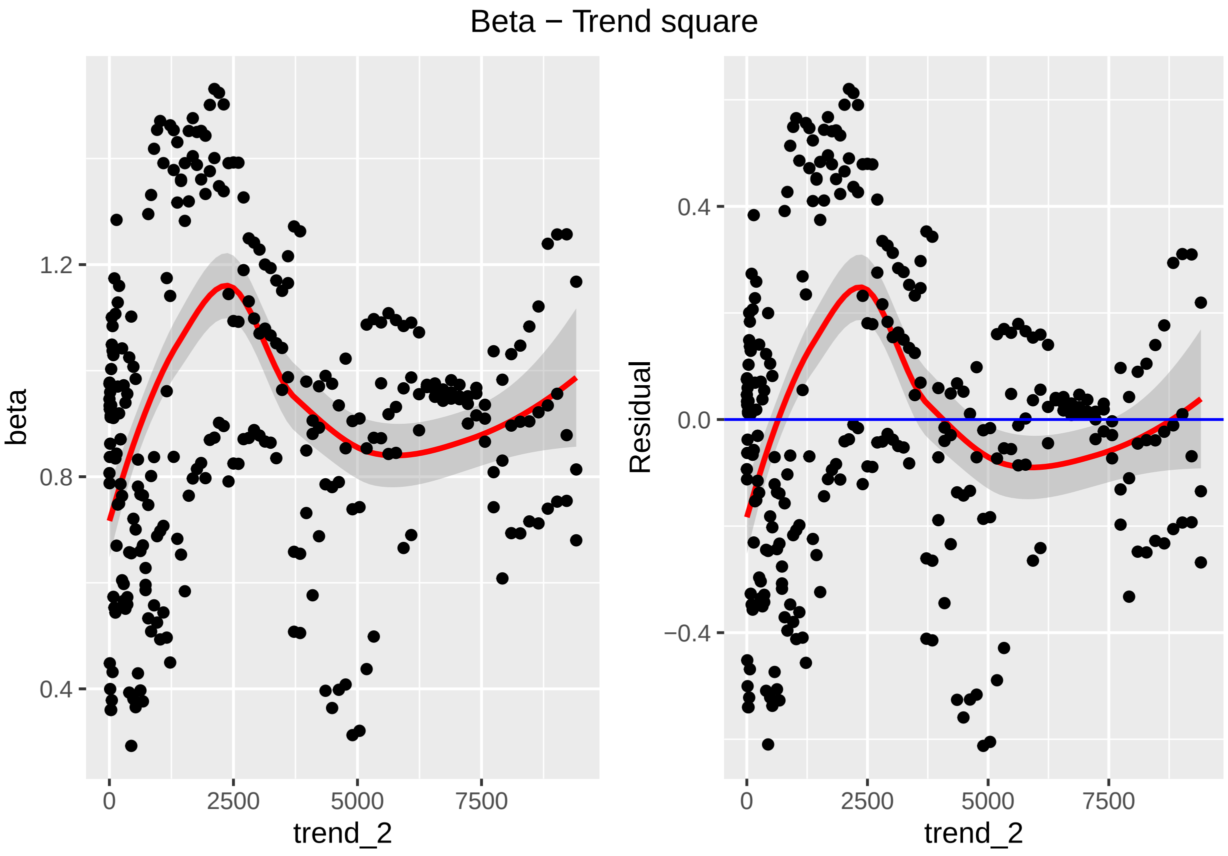
3.1 First Model

Use time series model to forescast the beta.

Time-series forecasting models are models that can predict the future stock return based on the historical observed values. The autoregressive model is the most common model used in value prediction, basically it depends on the previous values and created a linear model. In the first model, we used all variables and additional trend square to build a full model a predict the bata.

3.1.1 Data preprocessing

The data point randomly scattered after preprocessing the data.



3.1.2 Variables for model

Notations	
$Beta$	Response Variable (The cumputed beta value))
$trend$	Explanatory Variable (The trend index over time)
$trend^2$	Explanatory Variable (The quadratic trend index over time)
$spring$	Explanatory Variable (Dummy variable)
$summer$	Explanatory Variable(Dummy variable)
$fall$	Explanatory Variable (Dummy variable)
DJI	Explanatory Variable(Stock return of the Dow Jones Industrial)
$return$	Explanatory Variable (Stock return of the three companies)
lag_1	Explanatory Variable(The seasonal first lag of the beta)
lag_4	Explanatory Variable(The seasonal four lag of the beta)
lag_12	Explanatory Variable(The seasonal tweleve lag of the beta)
$GOOG$	Explanatory Variable(Dummy variable)
$AAPL$	Explanatory Variable (Dummy variable)

3.1.3 Build a model

1. Time series model

2. Model summary and evaluation

	Estimate St	d. Error	t value P	r(>|t|)
(Intercept)	0.6990328	0.0611980	11.4224846	0.0000000
trend	0.0103949	0.0023616	4.4016586	0.0000153
trend_2	-0.0000941	0.0000233	-4.0451796	0.0000678
spring	-0.0000273	0.0462801	-0.0005902	0.9995295
summer	-0.0016616	0.0467837	-0.0355159	0.9716939
fall	0.0305791	0.0470180	0.6503696	0.5159906
DJI	0.5959206	0.4753424	1.2536660	0.2110167
return	0.0343212	0.2923189	0.1174100	0.9066199
lag_1	0.0568659	0.2429291	0.2340844	0.8150918
lag_4	0.2846647	0.2468486	1.1531956	0.2498206
lag_12	-0.0281881	0.2586244	-0.1089923	0.9132872
GOOG	0.0218218	0.0396484	0.5503830	0.5824987
AALP	-0.0374889	0.0396862	-0.9446344	0.3456658

- Estimtaed regression line is:

$$\widehat{Beta} = 0.6990055 + 0.0103949 \times trend - 0.0000941 \times trend^2 - 0.0016342 \times spring + 0.0306064 \times summer$$

$$+0.0000273 \times fall + 0.5959206 \times DJI + 0.0343212 \times return + 0.0568659 \times y_t - 1 + 0.2846647 \times y_t - 4$$

$$-0.0281881 \times y_t - 12 + 0.0218218 \times GOOG - 0.0374889 \times AALP$$

- 8.37% of the variation in the Beta is accounted for by independent vairables.

3. P-value

	P.Value
(Intercept)	0.0000000
trend	0.0000153
trend_2	0.0000678
spring	0.9995295
summer	0.9716939
fall	0.5159906
DJI	0.2110167
return	0.9066199
lag_1	0.8150918
lag_4	0.2498206
lag_12	0.9132872
GOOG	0.5824987
AALP	0.3456658

- Only two variables's pvalue less than 0.05, $trend$ and $trend^2$.
- There is 0.0000153 chance that the relationship between $trend$ and $beta$ is due to chance.
- There is 0.0000687 chance that the relationship between $trend^2$ and $beta$ is due to chance.
- The p-value of the model is p = 0.01622
- There is almost 0.01622 that the overall model is due to chance.

4. Foecast

- Used the model to predict the lastest month's stock price and computed the MSE and MAE.

MSE	MAE
0.0472362	0.217339

5. Plot the prediction

- Plot the feature values and predictions.

Based on the result of model 1, we choose to drop variables with a high p-value, such as spring, summer, fall, one year lagged variable, and then observe the adjusted R-square at the next round of modeling. The prediction plot like the winsorized result of the original beta value.

3.2 Second Model

Use time series model to forescast the beta.

In the second model, we created a trend and autoregressive model to predict the beta.

3.2.1 Variables for model

Notations	
$Beta$	Response Variable (The cumputed beta value)
$trend$	Explanatory Variable (The trend index over time)
$trend^2$	Explanatory Variable (The quadratic trend index over time)
DJI	Explanatory Variable(Stock return of the Dow Jones Industrial)
$return$	Explanatory Variable (Stock return of the three companies)
lag_1	Explanatory Variable(The seasonal first lag of the beta)
lag_4	Explanatory Variable(The seasonal four lag of the beta)
$GOOG$	Explanatory Variable(Dummy variable)
$AAPL$	Explanatory Variable (Dummy variable)

3.1.2 Build a model

1. Create a model use the the trend variable and sesonal lagged variable to build the model

2. Model summary and evaluation

	Estimate St	d. Error	t value P	r(>|t|)
(Intercept)	0.7031216	0.0553548	12.7020841	0.0000000
trend	0.0104464	0.0023433	4.4579216	0.0000120
trend_2	-0.0000943	0.0000231	-4.0804995	0.0000586
DJI	0.5633199	0.4673081	1.2054571	0.2290374
return	0.0569356	0.2855657	0.1993782	0.8421106
lag_1	0.0786199	0.2355450	0.3337785	0.7387948

	Estimate St	d. Error	t value P	r(>|t|)
lag_4	0.2921390	0.2435466	1.1995197	0.2313328
GOOG	0.0219564	0.0394117	0.5571049	0.5778976
AALP	-0.0372091	0.0394388	-0.9434651	0.3462507

- Estimtaed regression line is:

$$\widehat{Beta} = 0.7031216 + 0.0104464 \times trend - 0.0000943 \times trend^2 + 0.5633199 \times DJI + 0.0569356 \times return$$

$$+0.0786199 \times y_t - 1 + 0.2921390 \times y_t - 4 + 0.0219564 \times GOOG - 0.0372091 \times AALP$$

- 8.143% of the variation in the Beta is accounted for by the independent variables.

3. P-value

	P.Value
(Intercept)	0.0000000
trend	0.0000120
trend_2	0.0000586
DJI	0.2290374
return	0.8421106
lag_1	0.7387948
lag_4	0.2313328
GOOG	0.5778976
AALP	0.3462507

- Only two variables's pvalue less than 0.05, $trend$ and $trend^2$.
- There is 0.0000120 chance that the relationship between $trend$ and $beta$ is due to chance.
- There is 0.0000586 chance that the relationship between $trend^2$ and $beta$ is due to chance.
- The p-value of the model is p = 0.002117
- There is almost 0.002117 that the overall model is due to chance.

4. Forecast

- Used the model to predict the lastest month's stock price and computed the MSE and MAE.

MSE	MAE
0.0494296	0.2223277

5. Plot the prediction

- Plot the feature values and predictions.

From the p-value result of model 2 we will drop variables with a high p-value, companies return, company dummy variable, one month lagged variables, and then observe the R-square of the next model.

3.3 Third Model

Use time series model to forescat the Beta.

In the third model, we created a trend and autoregressive model to predict the beta.

3.3.1 Variables for model

Notations	
$Beta$	Response Variable (The cumputed beta value)
$trend$	Explanatory Variable (The trend index over time)
$trend^2$	Explanatory Variable (The quadratic trend index over time)
DJI	Explanatory Variable(Stock return of the Dow Jones Industrial)
lag_1	Explanatory Variable(The seasonal first lag of the beta)
lag_4	Explanatory Variable(The seasonal four lag of the beta)

3.3.2 Build a model

1. Create a model use the the trend variable and sesonal lagged variable to build the model
2. Model summary and evaluation

	Estimate St	d. Error	t value P	r(>|t|)
(Intercept)	0.6993305	0.0501034	13.957733	0.0000000
trend	0.0104363	0.0023368	4.465985	0.0000115
trend_2	-0.0000940	0.0000230	-4.082327	0.0000579
DJI	0.6131895	0.3899692	1.572405	0.1169623
lag_4	0.2909621	0.2416958	1.203836	0.2296484

- Estimtaed regression line is:

$$\widehat{Beta} = 0.0000000 + 0.0000115 \times trend + 0.0000579 \times trend^2 + 0.1169623 \times DJI + 0.2296484 \times y_t - 4$$

- 7.336% of the variation in the Beta is accounted for by the independent variables.

3. P-value

	P.Value
(Intercept)	0.0000000
trend	0.0000115
trend_2	0.0000579
DJI	0.1169623
lag_4	0.2296484

- Only two variables's pvalue less than 0.05, $trend$ and $trend^2$.
- There is 0.0000115 chance that the relationship between $trend$ and $beta$ is due to chance.
- There is 0.0000579 chance that the relationship between $trend^2$ and $beta$ is due to chance.

- The p-value of the model is p = 0.0002132
- There is almost 0.0002132 that the overall model is due to chance.

4. Forecast

- Used the model to predict the lastest month's stock price and computed the MSE and MAE.

MSE	MAE
0.0511617	0.2261896

5. Plot the prediction

- Plot the feature values and predictions.

4. Final Model Interpretation

4.1 Model assumption

1: $E(\epsilon|x_i) = 0$

2: $\epsilon \sim N[0, \epsilon^2]$

3: ϵ are independent variables

4.2 Estimated regression line

$$\widehat{Beta} = 0.0000000 + 0.0000115 \times trend + 0.0000579 \times trend^2 + 0.1169623 \times DJI + 0.2296484 \times y_t - 4$$

4.3 Interpretation of the slope, intercept and R2

4.3.1 Interpret the slope for trend.

As each month passes, beta is expected to increase 0.0000694, regardless of the beta in the same month of the previous four months.

4.3.2 Interpret the slope for DJI.

The beta is expected to increase 0.1169623 on average, holding the effect of time fixed.

4.3.3 Interpret the slope for seasonal variable.

The beta is expected to increase 0.2296484 on average, holding the effect of time and market return fixed.

4.3.4 Interpret the intercept.

At time = 0 and market return = 0 and beta in the same month of the previous 4 months = 0, the beta is expected to be 0 on average.

4.3.5 Interpret the R^2

The R-squared of model 3 is 0.07336, adjusted R-squared of model 3 is 0.0604. The adjusted R-squared in model 1 is 0.0441, the adjusted R-squared in model 2 is 0.05537, compared with the two models, model 3 is the better one. Even though model 3 is the best one, the R-square of model 3 is still small enough, and we need to add more features to improve the model performance.

4.4 Hypothesis test

Does the beta change has the seasonal pattern?

- step 1:

  $H_0 : \beta_4 = 0$ $\qquad$ H_a: $\beta_4 \neq 0$

- step 2:

  Find the p-value for this coefficient, 0.2296484

- step 3:

  Reject the null when p-value $< 5\%$

- step 4:

Since p-value $= 22.96\% > 5\%$. We can't reject the null. We can't prove that the beta has a seasonal pattern.

4.5 Residual analysis

5. Summay and Conclusion Remarks

5.1 Summarize the results of the report in a concise fashion

To sum up, in this project our goal is to predict the beta value based on the time series model. We used seasonal variables including 4 season and lagged variables, we also used the linear trend variable and its quadratic variable, we separated the stock return into the market and the three companies. The first model is the full model after we analyze the result. We constructed the second model, dropped some variables that have a large p-value, and the third model building also follows the same method. Thus, by evaluating the adjusted R-square and p-value we used the backward elimination method to organize our project, and this method can provide good guidance for our project development.

The interesting is the adjusted R-square of the three models is very small, all of them lower than 0.1, thus, all of these three models have not fully explained the economic phenomenon. In order to get a better prediction in the future, we need to add more features to our model.

From the p-value perspective to analyze the result, we only found that two variables can provide better influence to the model prediction, trend, and its quadratic pattern.

5.2 Comments on the quality of data and reliability of concomitant inferential statements(what type of data would improve the reliability of your statements)?

The different industries may have different effects on model performance. In this project, we selected three companies all of which are technology companies, and we choose the market return from the Dow Jones Industrial. To improve the reliability we can select the stock return from other companies or select the other market stock return data.

5.3 Include ideas that you have about future investigations

- Since the natures of each industry vary from one to another, we propose to collect more features and build the model industry by industry.
- We plan on making a much more robust model by adding returns of other companies from myriad domains and get a holistic idea of which features explain the variation in beta most.

6. Appendix

- quantmod package

6.1 Additional graphs and tables

6.2 Reference

[1]"CAPM: Capital Asset Pricing Model: CAPM Formula: Capital Market Line." M1 Finance, 7 Nov. 2019, www.m1finance.com/articles-2/capm-capital-asset-pricing-model/.

[2] https://www.wallstreetmojo.com/capm-beta-definition-formula-calculate-beta-in-excel/ - The following article discusses about the impact of myriad factors on the value of beta and which industries tend to have higher betas and which industries tend to have a lower beta .The value of beta depends upon the nature of the business (cyclic) along with the financial debt the company owes

[3] https://finvert.home.blog/2019/02/28/how-helpful-is-beta/- The following article discusses about the applications of beta and also dives deep into its limitations

[4] https://tgwhite.github.io/R_training/

[5] Chen, J. (2020, August 28). Market Risk Premium. Retrieved December 02, 2020, from https://www.investopedia.com/terms/m/marketriskpremium.asp

[6] http://www.quantmod.com/examples/

[7] https://www.kdnuggets.com/2020/01/stock-market-forecasting-time-series-analysis