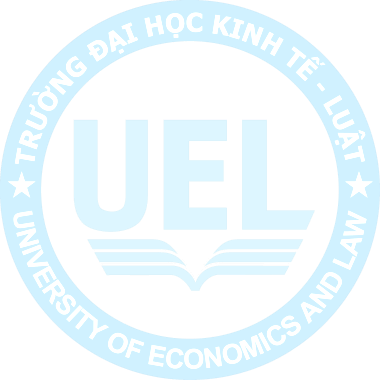
**END-OF-COURSE PROJECT**



**VIETNAM NATIONAL UNIVERSITY**

**UNIVERSITY OF ECONOMICS AND LAW**



**PROGRAM PACKAGE FOR FINANCE 2**

**Lecturer:** Ph.D Nguyễn Thanh Liêm

**Submitted by:**

|  |  |
| --- | --- |
| Bùi Nguyễn Thuỳ Như | K194141737 |

*Ho Chi Minh City, June 15th 2022*

**TABLE OF CONTENT**

[1. LITERATURE REVIEW 2](#_Toc106217882)

[1.1. Theories about cash holding 2](#_Toc106217883)

[1.2. Literature review 2](#_Toc106217884)

[2. DATA COLLECTION AND INPUT 4](#_Toc106217885)

[2.1. Data collection 4](#_Toc106217886)

[3. DESCRIPTIVE STATISTICS 6](#_Toc106217888)

[4. DATA VISUALIZATION OF CASH HOLDING 8](#_Toc106217889)

[4.1. Box & whisker plot of cash holding 8](#_Toc106217890)

[4.2. Histogram plot of cash holding 8](#_Toc106217891)

[5. MULTIPLE REGRESSION 9](#_Toc106217892)

[5.1. With the usual individual variables 9](#_Toc106217893)

[5.2. With the interaction between Covid-19 dummy variable 11](#_Toc106217894)

[5.3. Predict for all the quarters of the sample using Model 1 12](#_Toc106217895)

[6. PERFORM ARIMA MODEL TO PREDICT CASH HOLDING 13](#_Toc106217897)

[7. EXPLAIN HOW RANDOM FOREST CAN BE USED IN THIS CASE TO PREDICT 16](#_Toc106217898)

[REFERENCES 18](#_Toc106217899)

[APPENDIX 18](#_Toc106217900)

# LITERATURE REVIEW

## 1.1. Theories about cash holding

The amount of cash in a company's account represents its liquidity and asset structure. The amount of cash indicates if the manager is properly utilizing resources to carry out production and company tasks. A corporation with a lot of cash can be overspending and underutilizing its resources. If the company's cash flow is low, it may be difficult to pay its obligations and other expenses. Furthermore, the amount of cash on hand has an impact on the company's ability to deal with unexpected problems.

Traditional calculation according to Bates et al. (1999), Ozkan and Ozkan

(2004), the fomula is:

In the world, there have been many theories about the motivation for holding cash of enterprises. In this article, I focus on presenting two important theory and the basis for variable selection. That is, the Trade-off Theory (Keynes, 1936) states that firms determine their cash holdings by balancing the marginal costs and marginal benefits of holding cash. And, Pecking Order Theory (Myers, 1984) suggests that in order to reduce the financial costs associated with regarding asymmetric information, the sources of funding used by the firm are ranked class as follows: retained earnings, then debt, and finally stock issuance.

## 1.2. Literature review

|  |  |
| --- | --- |
| **Author** | **Conclusion** |
| Al-Najjar (2012) | Al-Najjar (2012) contributes to the study by exploring the financial determinants of cash holdings with data on 1212 listed non-financial firms in developing countries, namely Brazil, Russia, India, and China, and compare these results with those based on developed markets: USA and UK. The results in the study show that leverage, dividend only, liquidity, profitability, and firm size affect cash holdings. |
| Ferreira and Vilela (2004) | This study investigates the determinants of corporate cash holdings in EMU countries. The results suggest that cash holdings are positively affected by the investment opportunity set and cash flows and negatively affected by asset's liquidity, leverage and size. Bank debt and cash holdings are negatively related, which supports that a close relationship with banks allows the firm to hold less cash for precautionary reasons. Firms in countries with superior investor protection and concentrated ownership hold less cash, supporting the role of managerial discretion agency costs in explaining cash levels. Capital markets development has a negative impact on cash levels, contrary to the agency view. |
| Mai Daher (2010) | This article investigates the determinants of cash holdings of Indian manufacturing firms. The author investigate firm specific determinants such as Firm Size, Growth Opportunities, Leverage, Cash Flow, Dividend, Net Working Capital, R&D Expenditure, Assets Tangibility, Profitability, Interest Expenses, Cash Conversion Cycle, Inverse of Altman’s Z score, Firm Age and Cash Flow Volatility using a sample of 500 manufacturing firms for a period from 2005-2017. The study finds that Growth Opportunities, Leverage, Cash Flow, Dividend, Net Working Capital, R&D Expenditure and Profitability positively affect cash holdings whereas Firm Size, Assets Tangibility and Interest Expenses negatively affect cash holdings. Further, Firm Size and Growth Opportunities support the trade-off theory. Cash Flow and Profitability support the pecking order theory. |
| Mesfin, E. A. (2016) | This study is aimed to investigate the firms’ specific and macroeconomic variables of cash holdings of manufacturing share companies in Ethiopia over the period from 2009 to 2014 inclusive. In doing so, a multiple linear regression model is used for 15 randomly selected manufacturing share companies of Ethiopia. The findings of the study revealed that growth opportunity, cash flows and firm size are statistically significant and positively affect the cash holding of the manufacturing share companies. On the other hand, net working capital, capital expenditure and inflation have negative and statistically significant impact on cash holdings. Besides this, the leverage, profitability and real gross domestic product are statistically insignificant variables of cash holding decision for Ethiopian manufacturing share companies. |
| Uyar and Kuzey (2013) | This study is aimed to investigate ash flow, growth opportunities, amount of capital expenditure, liquid assets used as cash substitute, level of tangible assets, financial debt ratio and leverage effect to cash holding |

Based on the research profile, choose the following independent variables impact on the Cash holding variable: Leverage, Firm size, Net Working Capital, Growth

\* Correlation between variables according to the theory

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Measurement** | **Reference** | **Expectation** |
| Cash holding | Cash holding= Cash/Total assets | Opler et al. (1999) |  |
| Firm size | Firm size = ln(Total assets) | Al-Najjar (2012),  Mai Daher (2010), | - |
| Leverage | Leverage = Total debt/Total assets | Al-Najjar (2012),  Uyar and Kuzey (2013),  Mesfin, E. A. (2016) | - |
| Net Working Capital | Net Working Capital t-1 – Net Woriking Capital t | Mai Daher (2010) | + |
| Growth | Growth = (Sales t – Sales t-1)/  Sales t-1 | Mesfin, E. A. (2016) | + |

# DATA COLLECTION AND INPUT

## 2.1. Data collection

HPG (Hoa Phat Group): established in 1995. The company was listed and traded on HOSE in 8/2007. Hoa Phat Group holds the No. 1 market share in Vietnam for construction steel, steel pipes and Australian beef. Currently, Hoa Phat Group is in the Top 10 largest private enterprises in Vietnam.

By Refinitiv Eikon I have collected 38 fields of HPG data from quarterly financial statements from 2007 to Q1 2022.

The data covers 52 quarters before 2020 and 9 quarters after 2020.

The data collected is not perfect, so it is necessary to process the data for ease of calculation and visualization. The data is also only the Net Working Capital field available. Cash holding, Leverage, Growth, Firm size must be calculated.

- Use as.yearqtr from library (“zoo”) to convert dates to quarterly

- Use options(scipen = 999) to convert exponential notation to specific display, easier to see detailed number

- Because the data has string “NULL” for NA value so it is hard to determine NA value

### Preprocess data--------------------------------------------------------

data$Time = as.yearqtr(data$Time,format = "%Y-%m-%d") # Convert dates to quarterly

print(data$Time)

options(scipen = 999) # Convert exponential notation to specific display

data[data == "NULL"] = NA # Transform string "NULL" to NA value

view(data) # Check data

- Check needed data field to find NA value. As a result, there is no NA value

# Count NA values in the variables I will choose

sum(is.na(data$TotalAssets))

sum(is.na(data$Cash))

sum(is.na(data$TotalDebt))

sum(is.na(data$TotalRevenue))

sum(is.na(data$NetWorkingCapital))

- The data is from 2022 to 2007, so reverse the time for easy viewing and calculation

data = data[order(data$Time), ] # Reverse data transfer follow quarters

view(data) # Check data

- Create a function to calculate percentage change

# Create function calculate percentage change

pct <- function(x) {x / lag(x) - 1}

- Create and mutate needed variable

# Create variable Columns

as\_tibble(data)

data = data %>%

mutate(Cash\_Holding = Cash/TotalAssets) %>% # Create Cash Holding Column

mutate(Leverage = TotalDebt/TotalAssets) %>% # Create Leverage Column

mutate(Firm\_Size = log(TotalAssets)) %>% # Create Firm\_Size Column

mutate(Growth = pct(TotalRevenue)) # Create Liquidity Column

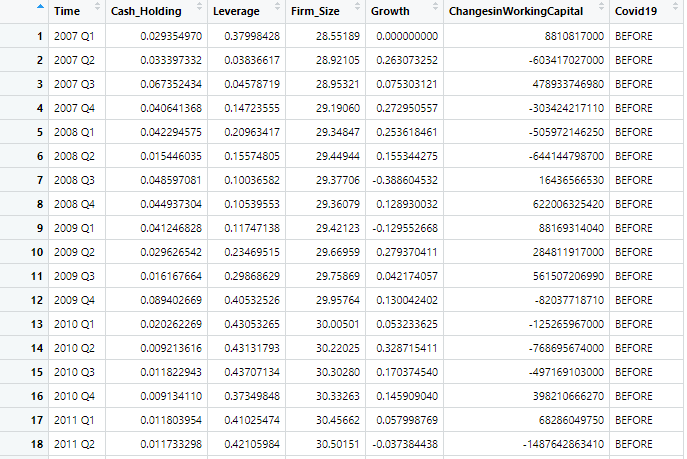
data$Growth[is.na(data$Growth)] = 0 # Fill the first cell of the Growth column as 0

view(data)

# 

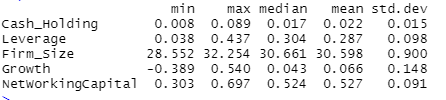
# DESCRIPTIVE STATISTICS

Create an empty column "Covid19," then classify to determine the years with COVID epidemics. Using the beginning of 2020 as the standard, quarters from previous years will receive the categorical value of "BEFORE" because COVID has not yet appeared, quarters from early 2020 and then with COVID will receive the categorical value "AFTER," and the result is:



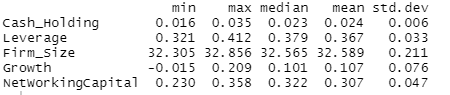
With the first 52 quarters receiving the “BEFORE” classification results and the following 8 quarters receiving the “AFTER” classification results.

- All variables in before Covid19 period:



According to the statistics described above, the years before the appearance of COVID did not have many abnormalities, there were not many fluctuations when the indexes all showed the growth of Hoa Phat group due to the demand to buy real estate during the period. This period is developing strongly, especially the Firm Size and Growth variables have quite large value ranges, leading to the largest standard deviation among the variables showing the development and expansion of the business during this period. The Growth variable also shows that the company's revenue is increasing and growing strongly. It can be concluded that, before Covid19, businesses thrived, the amount of cash holdings was maintained at a constant quarterly rate.

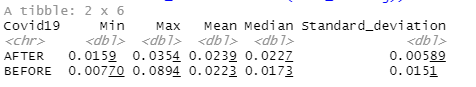
- All variables in after Covid19 period:



According to the above descriptive statistics, although the years after the appearance of Covid19 were not unusual, compared to the previous years, the above indicators all had a slower growth rate, even some even lower. Based on the standard deviation of Cash holding, it can be seen that the indexes during this period did not fluctuate too much, lower than previous years but the min value is higher, which proves that the company started to hold money, present a fixed amount of cash, this amount is more than in previous years to keep liquidity during Covid19. Due to the impact of Covid19, in order to maintain operations, the company has to take on more debt, leading to a higher minimum leverage value. Growth and Firm size variables are quite low compared to previous years. coupled with the relatively small min max range, this indicates a slowdown in the company's growth during this period.

It can be seen that the influence of Covid19 has negatively affected the development and operation of the company, the demand from customers has decreased, and especially the import and export activities of Hoa Phat Group's raw materials are limited. The proof is revenue (the Growth variable decreases). However, Hoa Phat Group still has a stable liquidity ratio

\* Cash holding:



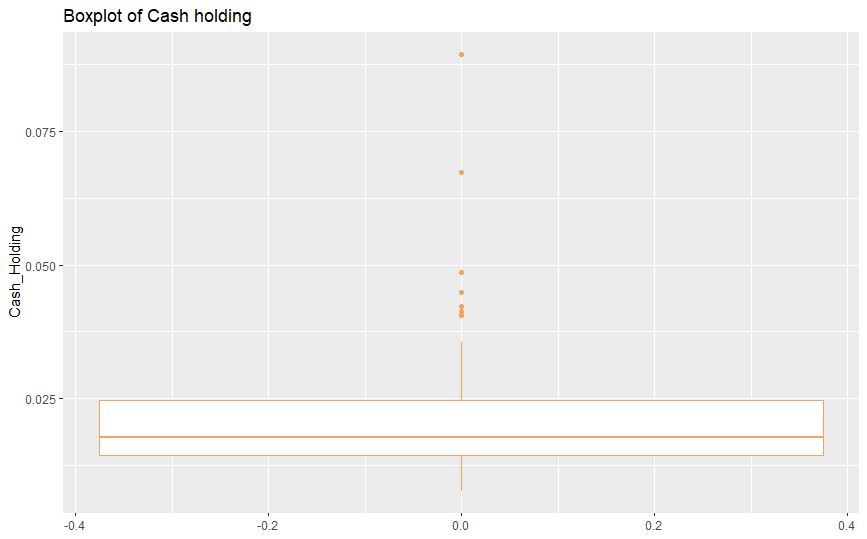
- In the pre-Covid19 period, HPG had quarters with a cash holding ratio: min = 0.00770, HPG had a quarter with the lowest cash holding ratio of 0.0077. Max = 0.0894, HPG has the highest cash holding quarter at 0.0894. The median and the mean for the quarters for this period are 0.894 and 0.223. Mean value > median value that the data has asymmetric distribution, the data is likely to have outliers. The standard deviation for the quarter for this period is as high as 0.151, indicating a change in cash holdings of 0.151.

- In the post-Covid19 period, HPG had quarters with a cash holding ratio: min = 0.0159, HPG has a quarter with the lowest cash holding ratio of 0.0159. max = 0.0354, HPG has the highest cash holding quarter at 0.0354. The median and mean of the quarters for this period are 0.0239 and 0.0227. Mean value > median value that the data has asymmetric distribution, the data is likely to have outliers. The standard deviation for these quarters is as low as 0.00589, indicating a change in cash holdings of 0.00589.

- In conclusion, in the years before Covid19, cash holdings were low, there were some quarters with suddenly high ratios (outliners). However, since the Covid19 period, the standard deviation has remained low, however, the distance between the min and the max has decreased, the company always holds a fixed cash to prevent liquidity risk.

# DATA VISUALIZATION OF CASH HOLDING

## 4.1. Box & whisker plot of cash holding



Median: data is concentrated in the 0.015 region

First quartile (Q1): about 0.007

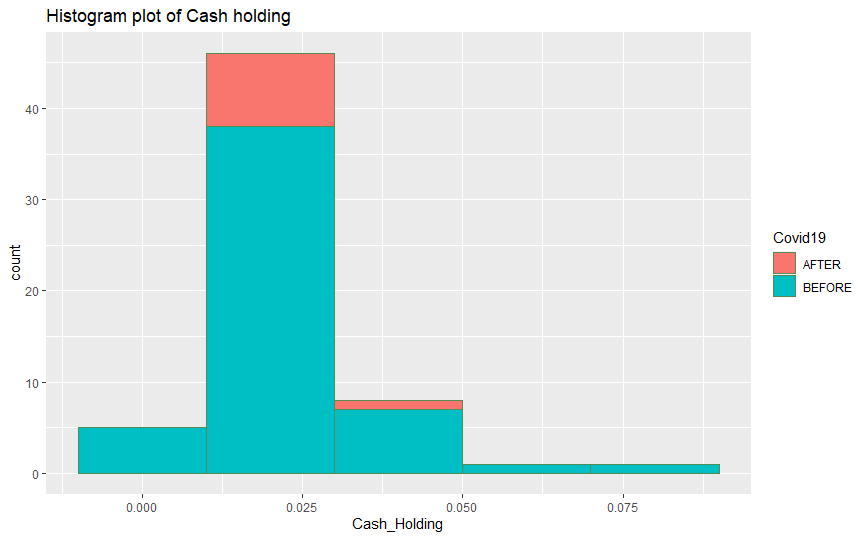
Third quartile (Q3): about 0.025

Minimum: about 0.007

Maximum: about 0.037

Cash holding ratio is not symmetrical and narrowly distributed, fluctuations in cash holdings quarterly are not much

## 4.2. Histogram plot of cash holding

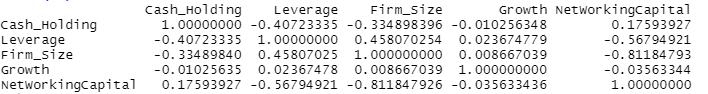


According to the chart above, most of the quarters, enterprises have cash holding ratio from 0.01 to 0.03, cash holding ratio from 0.03 to 0.05, very few quarters have cash holding ratio from 0-0.01 and only about 1-2 quarters where the company has a cash holding ratio above 0.05. Cash holding ratio for the most quarters is 0.01-0.03 with 46 quarters, 0.03-0.05 with 8 quarters and 0-0.01 with 5. During the Covid19 period, the company holds higher cash than in previous yearsIt can be seen that Hoa Phat Group tends to keep cash at extremely low levels and fixed quarterly.

# MULTIPLE REGRESSION

## 5.1. With the usual individual variables

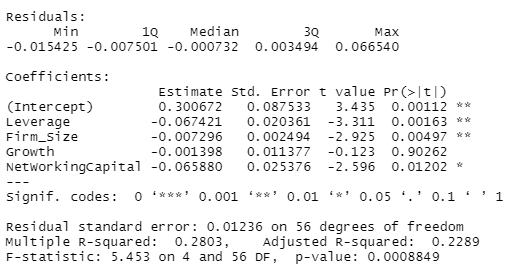
- Correlation matrix:



Of the 4 dependent variables, only the Firm size variable has a high correlation coefficient with the Net Working Capital variable of -0.812. The remaining variables have a rather low correlation with each other (the highest is only about -0.568).

There is suspected of having multicollinearity in the model.

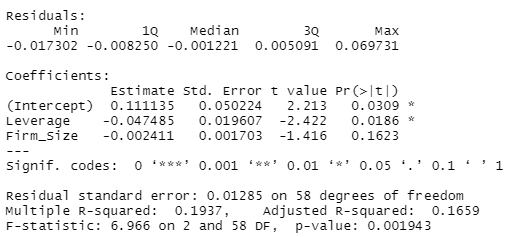
Regression of Cash holding variable with independent variables such as Leverage, Firm size, Growth, Net Working Capital the results are as follows:



Looking at the model, it is easy to see that the p-value of Growth = 0.90262 > 0.1, so the variable is not significant in the model, remove the variable.

Remove the variable NetWorkingCapital because multicollinearity is different from the original theory (expectation +) making the model can predict wrongly.

The model to remove the variable is as follows:



- Beta of Intercept = 0.11114 is significant in the condition that other variables do not affect Cash holding, the average Cash holding ratio is 0.11114 units.

- Beta Leverage = -0.04748 is significant in the condition that other variables do not change, when increasing by 1 unit Leverage, the Cash holding ratio will decrease on average 0.04748 units.

- Beta Firm\_Size = -0.00241 is significant in the condition that other variables remain unchanged, when increasing by 1 Firm\_Size unit, the Cash holding will decrease on average 0.00241 units.

- R-Squared = 0.1937 reflects the level of explanation of the independent variables for the dependent variable in the regression model, this is a rather low indicator that this model is difficult to explain the change of Cash holding. The independent variables explain 19.37% of the variation of the dependent variable. The reason is that to predict Cash holding, there need to be more influencing factors and not enough variables for analysis to reflect the change of Cash holding.

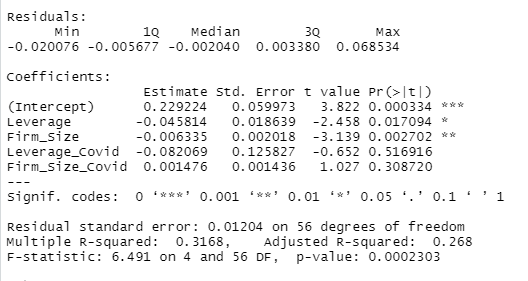
- At the 10% significance level, only 1 variable are statistically significant is Leverage having better predictive results.

## 5.2. With the interaction between Covid-19 dummy variable

Firstly, create 1 more column of dummy variable “Covid” with 2 values: “0” corresponds to the years before the appearance of Covid19 (from 2020 and earlier) and “1” corresponds to the years when Covid19 appeared (from 2020 and then).

Then, I establish three new columns to indicate the interaction between the independent variables and the dummy variable "Covid" by multiplying the values of the dependent variables by the new dummy column.

The model2 results is:



- Beta of Intercept = 0.22922 is significant in the condition that other variables do not affect Cash holding, the average Cash holding ratio is 0.22922 units.

- Beta Leverage = -0.04581 is significant in the condition that other variables do not change, when increasing by 1 unit Leverage, the Cash holding ratio will decrease on average 0.04581 units.

- Beta Firm\_Size = -0.00634 is significant in the condition that other variables remain unchanged, when increasing by 1 Firm\_Size unit, the Cash holding will decrease on average 0.00634 units.

- Beta Leverage\_Covid = -0.08207 is significant in the condition that other variables do not change, when increasing by 1 unit Leverage, the Cash holding ratio will decrease on average 0.08207 units.

- Beta Firm\_Size = 0.00148 is significant in the condition that other variables remain unchanged, when increasing by 1 Firm\_Size unit, the Cash holding will decrease on average 0.00148 units.

- R-Squared = 0.3168 reflects the level of explanation of the independent variables for the dependent variable in the regression model, this is a rather low indicator that this model is difficult to explain the change of Cash holding. The independent variables explain 19.37% of the variation of the dependent variable. The reason is that to predict Cash holding, there need to be more influencing factors and not enough variables for analysis to reflect the change of Cash holding.

- At the 10% significance level, only 2 variables are statistically significant is Leverage, Firm\_Size having better predictive results (p-value < 0.1).

Due to the small number of observations during the Covid-19 period, the impact of Covid on the Cash holding variable is not clear. In addition, Hoa Phat Group is an enterprise with a low cash holding rate before and after Covid, so the impact from Covid is negligible. Because the characteristics of Hoa Phat Group's products are quite specific, there is always a certain need. It can be concluded that Hoa Phat Group has not been too badly affected by Covid.

## 5.3. Predict for all the quarters of the sample using Model 1

Use predict.lm() to forecast prices based on pre-built model1. Then, create a dataframe with 3 columns "Time",'Actual', 'Predicted' to visualize the forecast. The result is:

Time Actual Predicted

1 2007 Q1 0.029354970 0.02426419

2 2007 Q2 0.033397332 0.03959591

3 2007 Q3 0.067352434 0.03916602

4 2007 Q4 0.040641368 0.03377652

5 2008 Q1 0.042294575 0.03043296

6 2008 Q2 0.015446035 0.03274833

7 2008 Q3 0.048597081 0.03555263

8 2008 Q4 0.044937304 0.03535300

9 2009 Q1 0.041246828 0.03463390

10 2009 Q2 0.029626542 0.02846886

11 2009 Q3 0.016167664 0.02521548

12 2009 Q4 0.089402669 0.01967217

13 2010 Q1 0.020262269 0.01836100

14 2010 Q2 0.009213616 0.01780487

15 2010 Q3 0.011822943 0.01733266

16 2010 Q4 0.009134110 0.02027950

17 2011 Q1 0.011803954 0.01823524

18 2011 Q2 0.011733298 0.01761395

19 2011 Q3 0.007703368 0.01907513

20 2011 Q4 0.014333042 0.02021745

21 2012 Q1 0.012915308 0.02149059

22 2012 Q2 0.015740489 0.02213518

23 2012 Q3 0.009992296 0.02250000

24 2012 Q4 0.009006633 0.02168084

25 2013 Q1 0.014132862 0.02238278

26 2013 Q2 0.010194521 0.02299125

27 2013 Q3 0.022220702 0.02193387

28 2013 Q4 0.012253358 0.02137303

29 2014 Q1 0.020224802 0.02123171

30 2014 Q2 0.024632054 0.02292287

31 2014 Q3 0.020892947 0.02211359

32 2014 Q4 0.022271129 0.02256067

33 2015 Q1 0.030855366 0.02300607

34 2015 Q2 0.018176262 0.02354790

35 2015 Q3 0.014265194 0.02371118

36 2015 Q4 0.013698176 0.02395652

37 2016 Q1 0.028737386 0.02469747

38 2016 Q2 0.021108711 0.02648415

39 2016 Q3 0.019336435 0.02711274

40 2016 Q4 0.016761375 0.02684964

41 2017 Q1 0.017703541 0.02345590

42 2017 Q2 0.017377301 0.02379438

43 2017 Q3 0.017535719 0.02396381

44 2017 Q4 0.014416550 0.02333124

45 2018 Q1 0.015555695 0.02410345

46 2018 Q2 0.017124700 0.02195129

47 2018 Q3 0.013374745 0.01935670

48 2018 Q4 0.023296242 0.01926368

49 2019 Q1 0.019194444 0.01678075

50 2019 Q2 0.016759110 0.01610738

51 2019 Q3 0.016368809 0.01488969

52 2019 Q4 0.016490270 0.01627045

53 2020 Q1 0.024729649 0.01493172

54 2020 Q2 0.021447972 0.01515346

55 2020 Q3 0.017145909 0.01474640

56 2020 Q4 0.015924960 0.01321700

57 2021 Q1 0.028126985 0.01364993

58 2021 Q2 0.022736018 0.01565847

59 2021 Q3 0.026636217 0.01547812

60 2021 Q4 0.035437761 0.01679079

61 2022 Q1 0.022661682 0.01655524

Use rmse function to calculate Root Mean Square Error (RMSE) is a measure of how well the model is performing. It does this by measuring the difference between predicted and actual values.

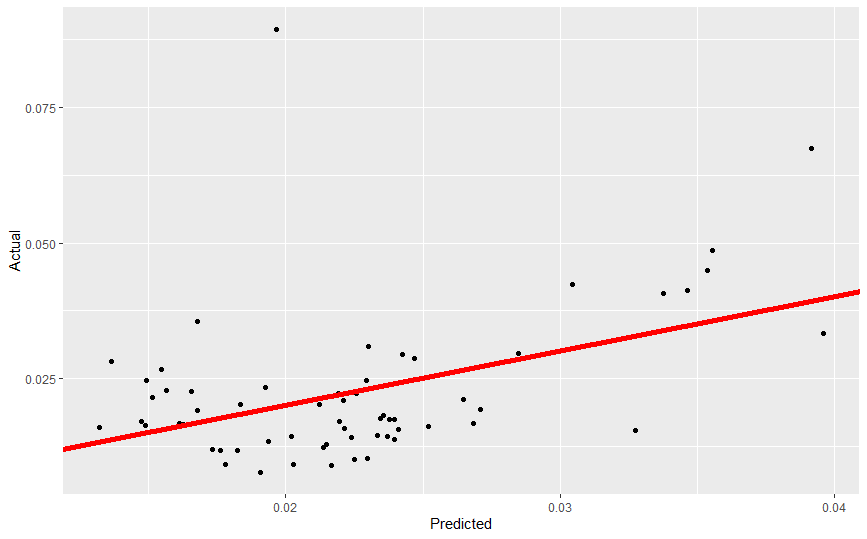
# RMSE

rmse(result\_predicted$Actual,result\_predicted$Predicted)

0.01253421

The RMSE = 0.01253421 is quite small, so that, the high level of estimation that shows the reliability of the model can be achieved.

Create a plot of predicted vs. actual values:



## The x-axis shows the predicted values ​​from the model and the y-axis shows the actual values ​​from the data set. The diagonal in the middle of the graph is the estimated regression line. The real values ​​are dispersed centered on this line.

# PERFORM ARIMA MODEL TO PREDICT CASH HOLDING

Firstly, convert quarters to date, then convert the available data into time series data format. To do so we need to run the following command:

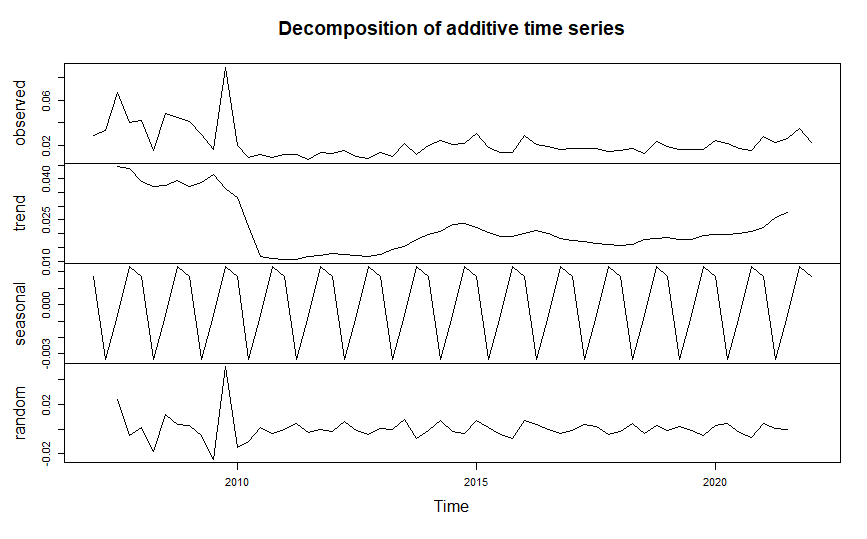
# 6. Perform ARIMA model to predict Cash holding---------------------------

as.Date(as.yearqtr(data$Time, format = "Q%q/%y"), frac = 1) # Convert quarters to date

tsData = ts(data$Cash\_Holding, start = c(2007, 1), frequency = 4) #Set timeseries data

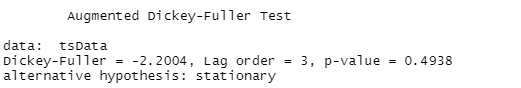
Start gives the starting time of the data, it is Jan 2007. As it is a quaterly data so ‘frequency=4’

Find out the components of this time series:



The results of the time series data decompose chart above show the following elements of the data: there is an overall quite uptrend in the data (based on the trend chart), the data is seasonal, the The seasonal law that repeats on a regular basis is quite clear (based on the seasonal chart)

\* Check stationary:



H0: Data is not stationary

H1: Data is stationary

Because p-value = 0.4938 > 0.05 so accept H0, can conclude Data is not stationary

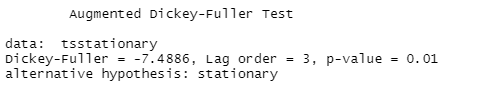
\* Removed the seasonal element of the data and fixed the stationarity of the data

# Find out the components of this time series

components.ts = decompose(tsData)

plot(components.ts)

\* Now check whether Cash\_holding is stationary



H0: Data is not stationary

H1: Data is stationary

Because p-value = 0.01 so accept H1, can conclude Data is stationary

\* Use auto.arima function to determine best P, D, Q (removed seasonal of data)

ARIMA(2,0,2)(1,0,1)[4] with non-zero mean : Inf

ARIMA(0,0,0) with non-zero mean : -321.8475

ARIMA(1,0,0)(1,0,0)[4] with non-zero mean : -335.7239

ARIMA(0,0,1)(0,0,1)[4] with non-zero mean : -348.5452

ARIMA(0,0,0) with zero mean : -323.8446

ARIMA(0,0,1) with non-zero mean : -349.4997

ARIMA(0,0,1)(1,0,0)[4] with non-zero mean : -348.4513

ARIMA(0,0,1)(1,0,1)[4] with non-zero mean : -347.2148

ARIMA(1,0,1) with non-zero mean : -347.8908

ARIMA(0,0,2) with non-zero mean : -348.0642

ARIMA(1,0,0) with non-zero mean : -336.7946

ARIMA(1,0,2) with non-zero mean : -347.0799

ARIMA(0,0,1) with zero mean : -351.1261

ARIMA(0,0,1)(1,0,0)[4] with zero mean : -350.1021

ARIMA(0,0,1)(0,0,1)[4] with zero mean : -350.184

ARIMA(0,0,1)(1,0,1)[4] with zero mean : -348.1982

ARIMA(1,0,1) with zero mean : -349.5656

ARIMA(0,0,2) with zero mean : -349.7671

ARIMA(1,0,0) with zero mean : -338.7883

ARIMA(1,0,2) with zero mean : -348.9613

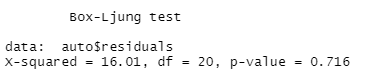
Best model: ARIMA(0,0,1) with zero mean

Best model is ARIMA(0,0,1)

\* Autocorrelation test in residuals: Ljung-Box test

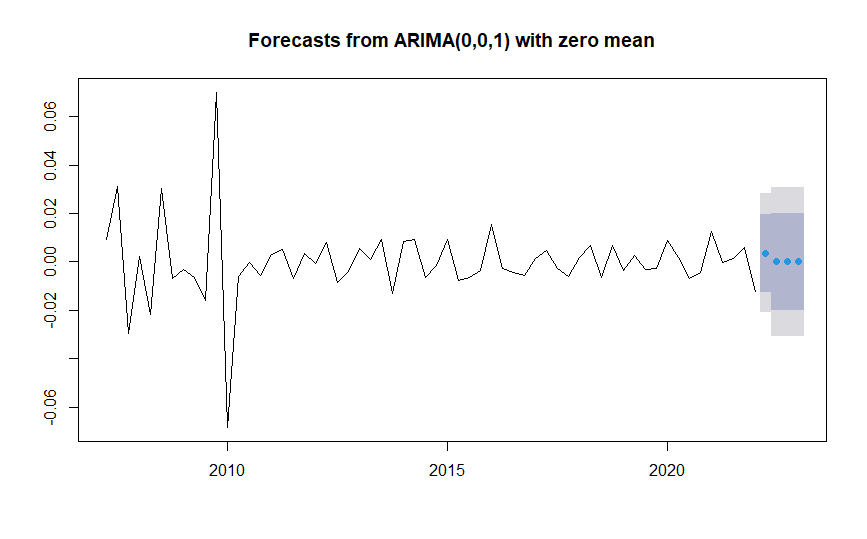
H0: Autocorrelation does not exist

H1: Autocorrelation exists



Result: p-value = 0.716 > 0.05, accept the hypothesis H0 does not exist correlation phenomenon.

\* Prediction: for the next 4 quaters



In conclution, it can be seen that the model predicts that Cash Holds will fluctuate in a horizontal line because the model finds that the data near the predicted value is gradually stabilizing and not having many unusual fluctuations. In fact, Hoa Phat Group does not have any changes in the cash holding ratio over the years, the cash holding ratio is usually very low.

# Count the number of firms in an industry and with cash holding above a certain value

count\_firm **=** 0

indus **=** "Industrials"

value **=** mean**(**data**$**leverage**)** #or any given number

**for** **(**i **in** 1**:**nrow**(**data**))** **{**

**if** **(**data**$**industry**[**i**]** **==** industry **&** data**$**leverage**[**i**]** **>** value **){**

count\_firm **=** count\_firm **+** 1

**}**

**}**

print**(**count\_firm**)**

# 7. EXPLAIN HOW RANDOM FOREST CAN BE USED IN THIS CASE TO PREDICT

Random forest can be used in this case for many reasons:

Firstly, Since a random forest is an ensemble of decision trees, it has lower variance than the other machine learning algorithms and it can produce better results

In addition, for financial time series data like this, there is a discrepancy between the units of variables (e.g. ROA in percent, Cash in billion VND), and other machine learning algorithms that require data to be scaled, whereas Random Forest does not. There is very little pre-processing that needs to be done.

A highlight of the random forest is that when collecting many independent variables, in the training set, the random forest will determine which independent variable has the greatest influence on the dependent variable. Therefore if using random forest, you can collect as many independent variables as possible, it will automatically analyze the variable with the greatest impact (expressed through “feature importance”).

# REFERENCES

Al‐Najjar, B., & Anfimiadou, A. (2012). Environmental policies and firm value. Business Strategy and the Environment, 21(1), 49-59.

Basil Al-Najjar (2012), "The financial determinants of corporate cash holdings:  
Evidence from some emerging markets", International Business Review, vol 22,  
page 77-88

Bates, T. Kahle, K. Rene (2009). Why do US firms hold so much more cash than they  
used to? Journal of Finance, volume 54, page 1995 – 2021.

Charles Yuji Horioka, Akiko Terada Hagiwara (2013), "Corporate cash holding in  
Asia", ADB economics working paper series No. 381

Daher, M. (2010). The determinants of cash holdings in UK public and private firms. Department of Accounting and Finance.

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. Expert systems with applications, 40(10), 3970-3983.

Ferreira, M. A., & Vilela, A. S. (2004). Why do firms hold cash? Evidence from EMU countries. European financial management, 10(2), 295-319.

John Keynes (1936), The general theory of employment, interest and money, DemandPublishing.

Mesfin, E. A. (2016). The factors affecting cash holding decisions of manufacturing share companies in Ethiopia. International Journal of Advanced Research in Management and Social Sciences, 5(3).

Myers, S.C., Majluf, N., 1984. Corporate financing and investment decisions  
when firms have information that investors do not have

Opler, T., Pinkowitz, L., Stulz, R., & Williamson, R. (1999). The determinants  
and implications of corporate cash holdings. Journal of financial economics, 52(1), 3-  
46

Ozkan and Ozkan, 2004. Corporate cash holdings: An empirical investigation  
of UK companies

Wenfeng Wu, Oliver M. Rui, Chongfeng Wu (2012). Trade credit, cash holdings, and  
financial deepening: Evidence from a transitional economy. Journal of Banking&Finance.

# APPENDIX

# Name: Bui Nguyen Thuy Nhu

# Student ID: K194141737

###-----------------------------END-OF-COURSE PROJECT-------------------------------

### Import library------------------------------------------------------------------

library(tidyverse)

library(readxl)

library(dplyr)

library(forcats)

library(ggplot2)

library(scales)

library(zoo)

library(pastecs)

library(forecast) #forecast

library(tseries) #adf.test

library(lmtest) #coeftest

library(stats) #Box.test

library(Metrics) #RMSE

### Import data---------------------------------------------------------------------

data = read\_excel('K194141737.xlsx')

### Preprocess data-----------------------------------------------------------------

data$Time = as.yearqtr(data$Time,format = "%Y-%m-%d") # Convert dates to quarterly

print(data$Time)

options(scipen = 999) # Convert exponential notation to specific display

data[data == "NULL"] = NA # Transform string "NULL" to NA value

view(data) # Check data

# Count NA values in the variables I will choose

sum(is.na(data$TotalAssets))

sum(is.na(data$Cash))

sum(is.na(data$TotalDebt))

sum(is.na(data$TotalRevenue))

sum(is.na(data$NetWorkingCapital))

data = data[order(data$Time), ] # Reverse data transfer follow quarters

view(data) # Check data

# Create function calculate percentage change

pct <- function(x) {x / lag(x) - 1}

# Create variable Columns

as\_tibble(data)

data = data %>%

mutate(Cash\_Holding = Cash/TotalAssets) %>% # Create Cash Holding Column

mutate(Leverage = TotalDebt/TotalAssets) %>% # Create Leverage Column

mutate(Firm\_Size = log(TotalAssets)) %>% # Create Firm\_Size Column

mutate(Growth = pct(TotalRevenue)) # Create Liquidity Column

data$Growth[is.na(data$Growth)] = 0 # Fill the first cell of the Growth column as 0

view(data)

# Create variables to distinguish the period with Covid19

data$Covid19 = 0

data$Covid19[as.numeric(format(data$Time, format="%Y"))<2020] = "BEFORE"

data$Covid19[as.numeric(format(data$Time, format="%Y"))>=2020] = "AFTER"

data = data %>%

select (Time,Cash\_Holding,Leverage,Firm\_Size,Growth,NetWorkingCapital,Covid19)

view(data)

### 3. Descriptive statistics ------------------------------------------------------

# Descriptive statistics all variable before Covid19--------------------------------

Before\_Covid19 = data %>%

filter(Covid19 == "BEFORE") %>%

select(Cash\_Holding,Leverage,Firm\_Size,Growth,NetWorkingCapital) %>%

stat.desc() %>%

t() %>%

subset(select = c(min,max,median,mean,std.dev))

print(round(Before\_Covid19,3))

## Detail descriptive statistics

# Descriptive statistics all variable after Covid19---------------------------------

After\_Covid19 = data %>%

filter(Covid19 == "AFTER") %>%

select(Cash\_Holding,Leverage,Firm\_Size,Growth,NetWorkingCapital) %>%

stat.desc() %>%

t() %>%

subset(select = c(min,max,median,mean,std.dev))

print(round(After\_Covid19,3))

# Descriptive statistics: Cash holding ---------------------------------------------

data %>%

group\_by(Covid19) %>%

summarize(Min = min(Cash\_Holding),

Max = max(Cash\_Holding),

Mean = mean(Cash\_Holding),

Median = median(Cash\_Holding),

Standard\_deviation = sd(Cash\_Holding))

# Descriptive statistics: Leverage -------------------------------------------------

data %>%

group\_by(Covid19) %>%

summarize(Min = min(Leverage),

Max= max(Leverage),

Mean=mean(Leverage),

Median = median(Leverage),

Standard\_deviation = sd(Leverage))

# Descriptive statistics: Firm\_Size ------------------------------------------------

data %>%

group\_by(Covid19) %>%

summarize(Min = min(Firm\_Size),

Max= max(Firm\_Size),

Mean=mean(Firm\_Size),

Median = median(Firm\_Size),

Standard\_deviation = sd(Firm\_Size))

# Descriptive statistics: Growth ---------------------------------------------------

data %>%

group\_by(Covid19) %>%

summarize(Min = min(Growth),

Max= max(Growth),

Mean=mean(Growth),

Median = median(Growth),

Standard\_deviation = sd(Growth))

# Descriptive statistics: Change in Working Capital --------------------------------

data %>%

group\_by(Covid19) %>%

summarize(Min = min(NetWorkingCapital),

Max= max(NetWorkingCapital),

Mean=mean(NetWorkingCapital),

Median = median(NetWorkingCapital),

Standard\_deviation = sd(NetWorkingCapital))

### 4. The box & whisker plot and histogram of Cash holding-------------------------

# Box & whisker plot----------------------------------------------------------------

# For entire

data %>% ggplot(aes(y = Cash\_Holding)) +

labs(title = "Boxplot of Cash holding") +

geom\_boxplot(color = 'sandybrown')

# For 2 periods

data %>% ggplot(aes(x=Covid19, y = Cash\_Holding,fill = Covid19)) +

geom\_boxplot()

# Histogram plot---------------------------------------------------------------------

ggplot(data, aes(x= Cash\_Holding, fill=Covid19)) +

labs(title = "Histogram plot of Cash holding ") +

geom\_histogram(binwidth=0.02, color = 'palegreen4')

### 5. Perform multiple regression--------------------------------------------------

# 5.1. With the usual individual variables------------------------------------------

df = data %>% select(Cash\_Holding,Leverage,Firm\_Size,Growth,NetWorkingCapital)

cor(df)

model0 = lm(Cash\_Holding~Leverage+Firm\_Size+Growth+NetWorkingCapital, data=data)

summary(model0)

model1 = lm(Cash\_Holding~Leverage+Firm\_Size, data=data)

summary(model1)

# 5.2. With the interaction between Covid-19 dummy variable-------------------------

# Create dummy variable

data$Covid = 0

data$Covid[as.numeric(format(data$Time, format="%Y"))<2020] = 0

data$Covid[as.numeric(format(data$Time, format="%Y"))>=2020] = 1

data = data %>%

mutate(Leverage\_Covid = Leverage\*Covid) %>%

mutate(Firm\_Size\_Covid = Firm\_Size\*Covid) %>%

view(data)

# Perform multiple regression with dummy variable

model2 = lm(Cash\_Holding~Leverage+Firm\_Size

+Leverage\_Covid+Firm\_Size\_Covid, data=data)

summary(model2)

# 5.3 Predict for all the quarters of the sample using Model 1-----------------------

# Make predict

predict\_model = predict.lm(model1)

# Create result\_predicted data frame

result\_predicted = data.frame(matrix(ncol = 3, nrow = 61))

colnames(result\_predicted) = c("Time",'Actual', 'Predicted')

result\_predicted$Time = data$Time

result\_predicted$Actual = data$Cash\_Holding

result\_predicted$Predicted = predict\_model

# Convert factor to numeric

result\_predicted$Actual = as.numeric(result\_predicted$Actual)

result\_predicted$Predicted = as.numeric(result\_predicted$Predicted)

result\_predicted

# RMSE

rmse(result\_predicted$Actual,result\_predicted$Predicted)

# Plot predicted values

ggplot(result\_predicted,

aes(x = Predicted,

y = Actual)) +

geom\_point() +

geom\_abline(intercept = 0,

slope = 1,

color = "red",

size = 2)

# 6. Perform ARIMA model to predict Cash holding------------------------------------

as.Date(as.yearqtr(data$Time, format = "Q%q/%y"), frac = 1) # Convert quarters to date

tsData = ts(data$Cash\_Holding, start = c(2007, 1), frequency = 4) #Set timeseries data

# Find out the components of this time series

components.ts = decompose(tsData)

plot(components.ts)

# Check stationary

par(mfrow=c(1,1))

acf(tsData,main='ACF for Cash holding')

pacf(tsData,main='PACF for Cash holding')

adf.test(tsData)

timeseriesseasonallyadjusted <- tsData- components.ts$seasonal

tsstationary <- diff(timeseriesseasonallyadjusted)

# Now check whether Cash\_Holding is stationary

acf(tsstationary)

pacf(tsstationary)

adf.test(tsstationary)

# Use auto.arima function to determine best P, D, Q

auto = auto.arima(tsstationary,trace = T,max.order=4,

ic='aic')

summary(auto)

# Ljung-Box test

coeftest(auto.arima(tsstationary,seasonal=F))

acf(auto$residuals)

pacf(auto$residuals)

Box.test(auto$residuals,lag=20,type='Ljung-Box')

# Prediction

term= 4

predict(auto,h=term)

fcastauto=forecast(auto,h=term)

fcastauto

plot(fcastauto)