Name: Lê Nguyễn Phú Lộc

Student ID: K194141730

**PROGRAM PACKAGE IN FINANCE 2**

**Final Exam**

1. **Literature review.**

* My student id ends with an even number so I did a review on factors affect **Leverage**.
* Al‐Sakran, S. A. (2001) documented a negative relationship between **growth, return on assets** and **leverage**. However, **size** is documented to show a positive relationship with **leverage**.
* Jermias, J., & Yigit, F. (2019) shown **firm’s size** partially has a positive and significant effect on **leverage**. They also find that **profitability** and **growth opportunity** are negatively and significantly related to **leverage**.
* A Kadim, K., & Nardi, S. (2018) find that **firm size** are positively and significantly related to **leverage**. They also find that **profitability** is negatively and significantly related to **leverage**; **asset growth** partially has a positive and not significant effect on **leverage**; **liquidity** partially has a positive and not significant effect on **leverage**.
* Karim, S., Rabbani, M. R., & Khan, M. A. (2021) show that company’s **leverage** is determined by **profitability, size, liquidity, and growth opportunities.**
* Ali, L. (2011) show **firm size, growth, profitability** have strong significant influence on firm’s **leverage**. The positive effect of **firm size** and a negative effect of **firm growth**, and **profitability**, on **leverage** confirm the predictions of capital structure theories as well as previous research papers.
* Onofrei, M., Tudose, M. B., Durdureanu, C., & Anton, S. G. (2015) said the relationship between **liquidity** and **leverage** appears to be negative and significant. Our regression results show that **firm size** is negatively related to **leverage**. **Growth opportunities** have a negative relationship with **leverage**.
* The variables I’ll use is **Profitability (ROA), Firm Size, Liquidity, and Growth** to see if **Leverage** is affected by them.
* Here is how I calculate the variables:

1. **ROA** is available on the ratio in finance.vietstock.vn.
2. **Firm Size**  = ln(**Total assets)**.
3. **Liquidity** =
4. **Growth =**
5. **Data collection and input**

* I select VCS – Vicostone from finance.vietstock.vn. The data was collected from Quarter 4 2007 to Quarter 1 2022. I download the data I need to calculate the variables.

# Import library

library(readxl)

library(tidyverse)

library(SciViews)

library(dplyr)

library(tidyr)

library(zoo)

# Data collection and input

df <- read\_excel('D:/University/N3-II/UDCT2/K194141730.xlsx')

df <- transmute(df,Quarter=Quarter,

ROA=ROA,

GRO=(`Total assets`-lag(`Total assets`))/lag(`Total assets`),

LIQ=`Current assets`/`Short -term liabilities`,

Size= ln(`Total assets`),

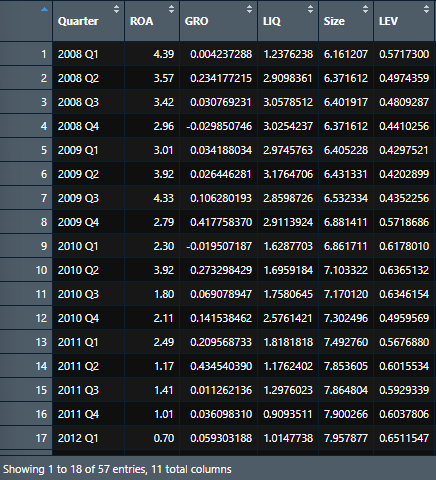
LEV=Liabilities/`Total assets`)

df <- df[complete.cases(df), ]

df <- df %>%

mutate(Quarter = as.yearqtr(format(Quarter), "%Y Q%q"))

* First, I import the library I need then import the excel into a dataframe.
* Then I calculate the variables, Growth first row will be NA because there is not enough data to calculate so I drop the NA.
* I change the Quarter column type to yearqtr for easier use of this column.
* Here the final df:



1. **Provide descriptive statistics of all the variables for BEFORE and AFTER periods**

# Descriptive stats BEFORE and AFTER Covid

df %>%

filter(`COVID-19`=='BEFORE') %>%

select(ROA) %>%

summarise(Median = median(ROA),

Mean = mean(ROA),

Min = min(ROA),

Max = max(ROA),

Standard\_Deviation = sd(ROA)) %>%

print()

df %>%

filter(`COVID-19`=='AFTER') %>%

select(ROA) %>%

summarise(Median = median(ROA),

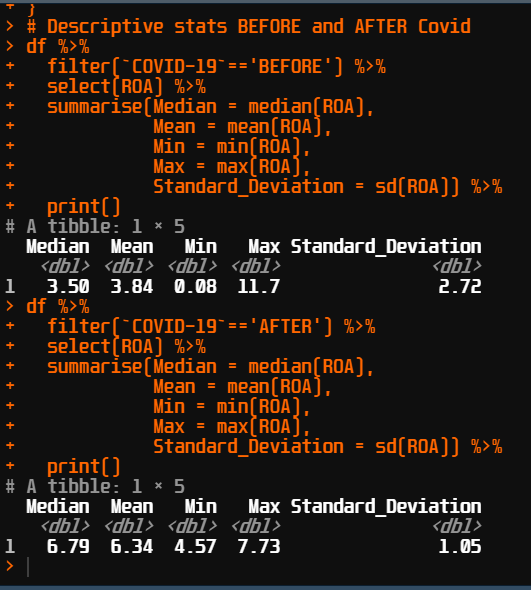
Mean = mean(ROA),

Min = min(ROA),

Max = max(ROA),

Standard\_Deviation = sd(ROA)) %>%

print()

* I create a Covid-19 column to call the data easier then I filter the variables by Covid-19 column and select the variables I want to describe with their Min, Max, Mean, Median and Standard Deviation.
* The first variable is ROA: 
* We can see that the Max have been decrease by almost 40% but the Min, Mean, Meadian increased so we could see that Covid-19 have no impact in the company’s profitability. The company still have the ability to maintain the return on assets during Covid-19.

df %>%

filter(`COVID-19`=='BEFORE') %>%

select(GRO) %>%

summarise(Median = median(GRO),

Mean = mean(GRO),

Min = min(GRO),

Max = max(GRO),

Standard\_Deviation = sd(GRO)) %>%

print()

df %>%

filter(`COVID-19`=='AFTER') %>%

select(GRO) %>%

summarise(Median = median(GRO),

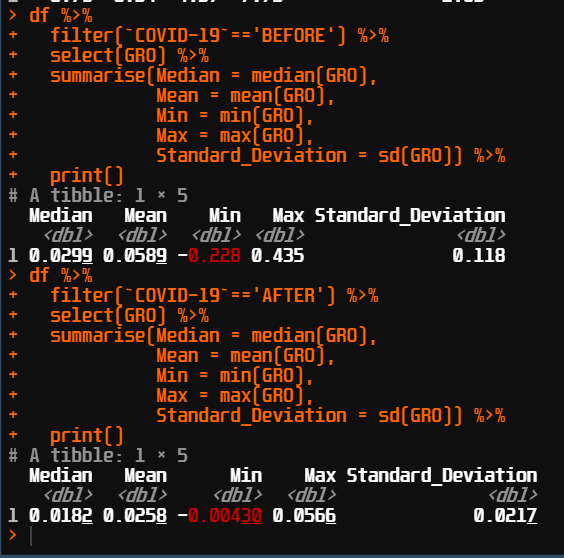
Mean = mean(GRO),

Min = min(GRO),

Max = max(GRO),

Standard\_Deviation = sd(GRO)) %>%

print()

* Next is the Growth: 
* Unlike the ROA, Covid-19 impacted Growth so much that it shrink every statistic of Growth after the Covid-19. It means the total asset after Covid-19 increased not much compare to before Covid-19.

df %>%

filter(`COVID-19`=='BEFORE') %>%

select(LIQ) %>%

summarise(Median = median(LIQ),

Mean = mean(LIQ),

Min = min(LIQ),

Max = max(LIQ),

Standard\_Deviation = sd(LIQ)) %>%

print()

df %>%

filter(`COVID-19`=='AFTER') %>%

select(LIQ) %>%

summarise(Median = median(LIQ),

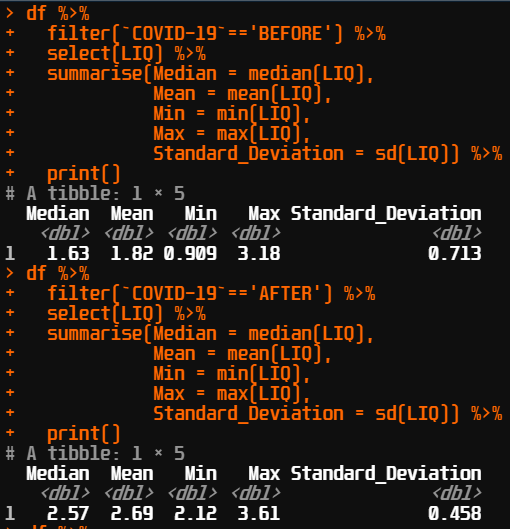
Mean = mean(LIQ),

Min = min(LIQ),

Max = max(LIQ),

Standard\_Deviation = sd(LIQ)) %>%

print()

* Liquidity: 
* The company’s Liquidtity has improved a lot after the Covid-19. The company could have increase its assets or reducing its liabilities making this ratio very high after Covid-19.

df %>%

filter(`COVID-19`=='BEFORE') %>%

select(Size) %>%

summarise(Median = median(Size),

Mean = mean(Size),

Min = min(Size),

Max = max(Size),

Standard\_Deviation = sd(Size)) %>%

print()

df %>%

filter(`COVID-19`=='AFTER') %>%

select(Size) %>%

summarise(Median = median(Size),

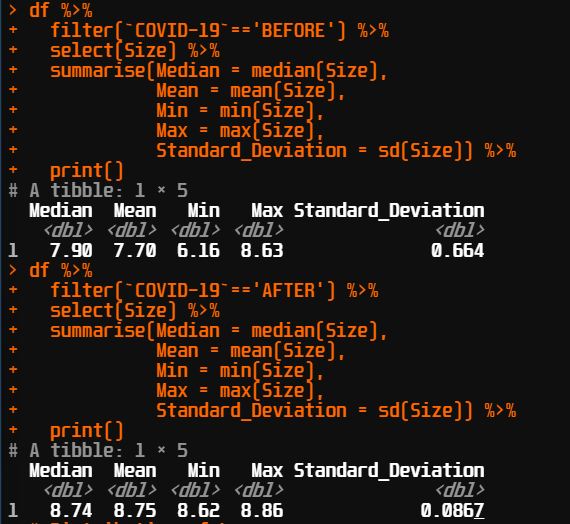
Mean = mean(Size),

Min = min(Size),

Max = max(Size),

Standard\_Deviation = sd(Size)) %>%

print()

* Finally, the firm’s Size:
* This ratio was calculated by log natural of the firm’s Total assets. Just like the Liquidity, the firm has an increasing in its Total assets making its Size stil rise despite the Covid-19.

1. **Provide box & whisker plot and histogram of the variable of assigned topic, i.e., Leverage/Cash holding (for the entire period)**

# Distribution of Leverage

df %>%

ggplot(aes(LEV,`COVID-19`,color= `COVID-19`))+

geom\_boxplot()+

theme\_minimal()+

labs(title = 'Boxplot of Leverage',x= 'COVID-19')

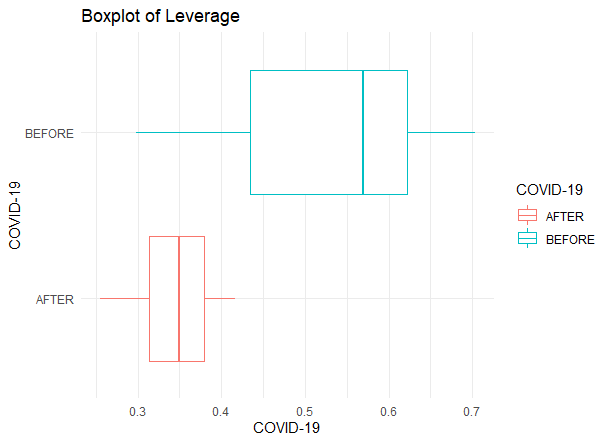
df %>%

ggplot(aes(LEV))+

geom\_histogram()+

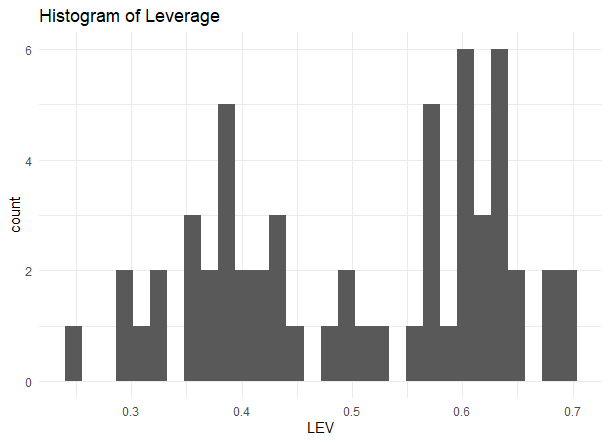
theme\_minimal()+

labs(title = 'Histogram of Leverage')

1. First, the boxplot:

* You can easily see that after the Covid-19, the firm’s Leverage has reduced in its range and mean. This can be explained like the Liquidity and Size, the company could have lessened their Liabilities or enhanced its Total assets making this ratio smaller than before the Covid-19.

1. The histogram:



* The distribution of the Leverage shows that the company leverage kinda evenly distributed.

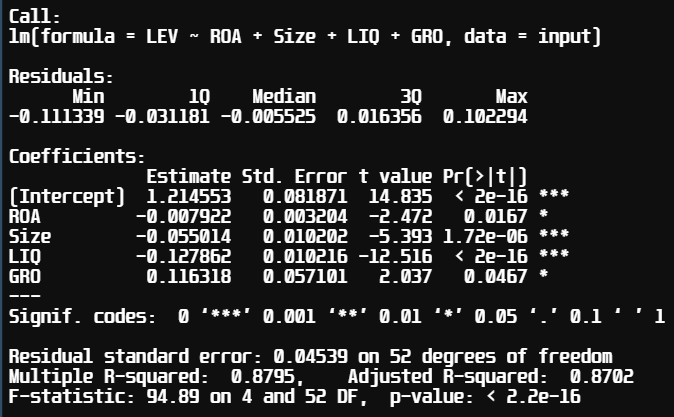
1. **Perform multiple regression to determine the significant determinants of the variable of assigned topic. The significance level is 10%.**
2. **With the usual individual variables (model 1)**

# Model 1

input <- df[c('LEV','ROA','Size','LIQ','GRO')]

model <- lm(LEV~ROA+Size+LIQ+GRO,data=input)

* First, I create a df to store the variables for the model. Then I input the variables to the model and run it. Here is the results:



* The estimate effect of ROA on Leverage is -0.007. The estimate effect of Size on Leverage is -0.055. The estimate effect of Liquidity on Leverage is -0.127. There is a significant negative relationship between ROA, Size, and Liquidity to Leverage (p-value <0.1).

- The estimate effect of Growth on Leverage is 0.116. There is a significant positive relationship between ROA and Leverage (p-value <0.1).

- The R2 shows that this model is good. And there is significant effect between the independent variables and the dependent variable.

1. **With the usual individual variables and the interaction between Covid-19 dummy variable and the independent variables (model 2)**

# Model 2

df$`COVID-19`[df$`COVID-19`=="BEFORE"] <- "0"

df$`COVID-19`[df$`COVID-19`=="AFTER"] <- "1"

df$`COVID-19`<- as.numeric(df$`COVID-19`)

df$ROAC <- df$ROA\*df$`COVID-19`

df$GROC <- df$GRO\*df$`COVID-19`

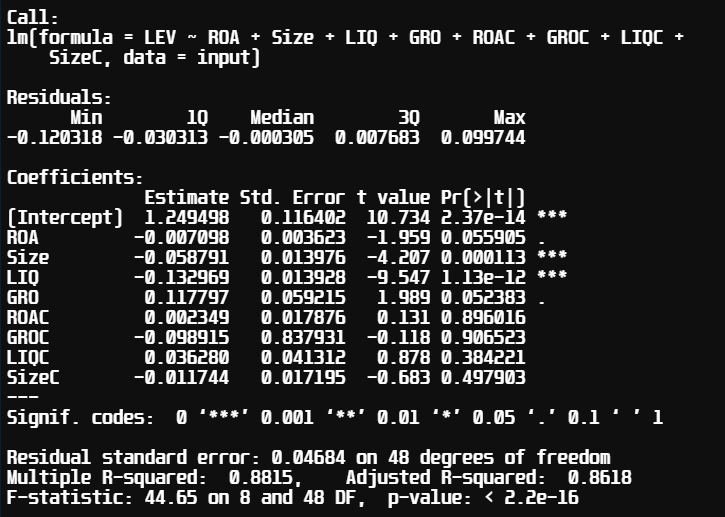
df$LIQC <- df$LIQ\*df$`COVID-19`

df$SizeC <- df$Size\*df$`COVID-19`

input <- df[c('LEV','ROA','Size','LIQ','GRO','ROAC','GROC','LIQC','SizeC')]

model <- lm(LEV~ROA+Size+LIQ+GRO+ROAC+GROC+LIQC+SizeC,data=input)

summary(model)

* I change the value in the Covid-19 column for the model. From before-after to 0-1, set its type as numeric. Then I create the variables with the interactions between the Covid-19 dummy and the independents. The result of the model 2:
* The old individual variables remain their significances. But the p-value of the Covid-19 and independents are unacceptable. This means the Covid-19 have no impact to the firm Leverage.
* The R2 of the model are high so this model is good.

1. **Predict the value of the variable of assigned topic for all the quarters of the sample using Model 1.**

# Predict model 1

input <- df[c('LEV','ROA','Size','LIQ','GRO')]

model <- lm(LEV~ROA+Size+LIQ+GRO,data=input)

pred <- predict(model, data = input, type='response')

* I built the model for the prediction and predict the Leverage.

# Create df for predicted data

result\_predicted <- data.frame(matrix(ncol = 3, nrow = 57))

colnames(result\_predicted) <- c("Quarter", 'Actual', 'Predicted')

result\_predicted$Quarter <- df$Quarter

result\_predicted$Actual <- df$LEV

result\_predicted$Predicted <- pred

# Convert to numeric

result\_predicted$Actual <- as.numeric(result\_predicted$Actual)

result\_predicted$Predicted <- as.numeric(result\_predicted$Predicted)

result\_predicted

- Then I create a df for the prediction and the actual data. After that I convert them to numeric to compare between predictions and actual data.

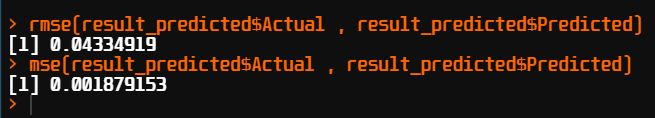
# RMSE, MSE

library("Metrics")

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

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

* I use RMSE and MSE to compare the prediction



* The model predicted very good. The RMSE, MSE are too small between the predictions and the actuals.

1. **Perform ARIMA model to predict the variable of interest for the 4 quarters in 2022**

library(forecast) #forecast, accuracy

library(tseries) #adf.test

library(lmtest) #coeftest

library(stats) #Box.test

# Leverage ADF test

lev <- diff(df$LEV,differences = 2)

# lev <- df\_train$LEV

# lev <- log(df$LEV)

acf(lev,main= "ACF for Leverage")

pacf(lev,main="PACF for Leverage")

adf.test(lev)# Leverage is not stationary

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

auto=auto.arima(lev,seasonal=F,trace = T,max.order=4,

ic='aic')

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

acf(auto$residuals)

pacf(auto$residuals)

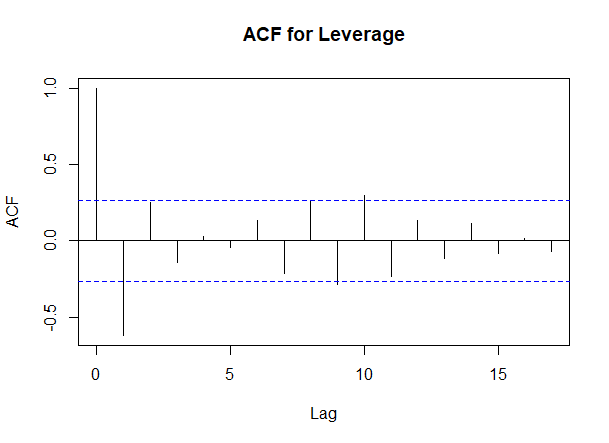
Box.test(auto$residuals,lag=20,type='Ljung-Box')

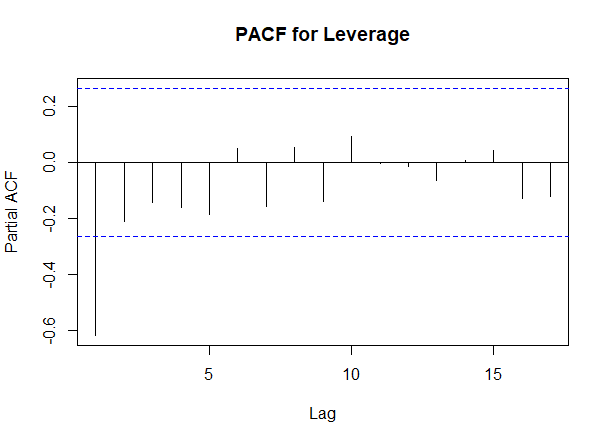
# Prediction

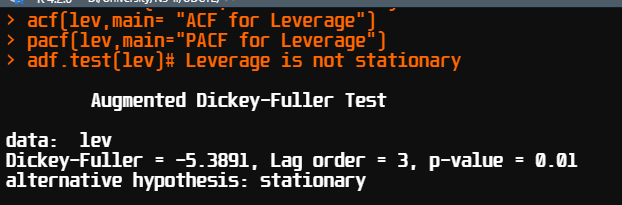
forecast(auto,h=4)

plot(forecast(auto,h=4))

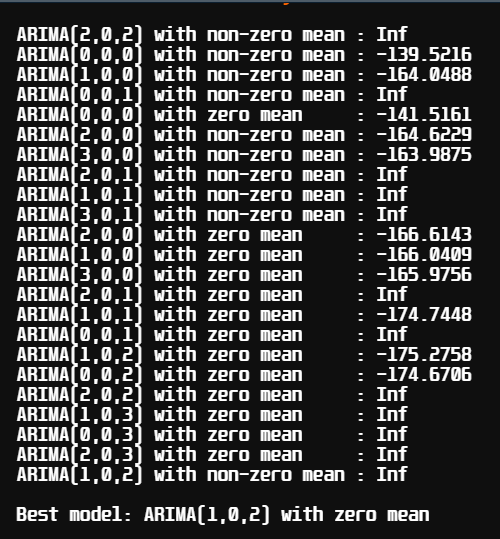
* I import the library. Because my data is not stationary so I use the difference to make the data stationary. At first my data pass all the test for ARIMA: adf test, box test. But the model can’t forecast so I make the differenc up to 2. Here are the results:



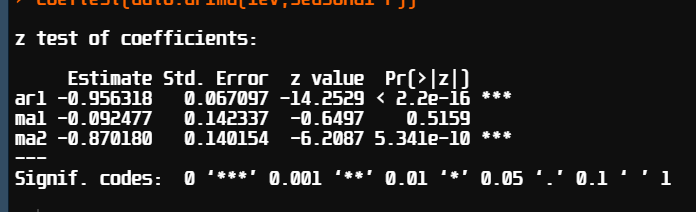


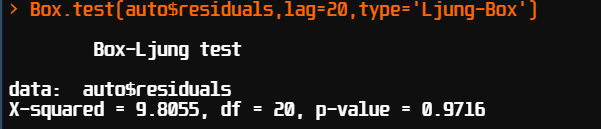


* The p-value of the adf was smaller than 0.05. The data is stationary.

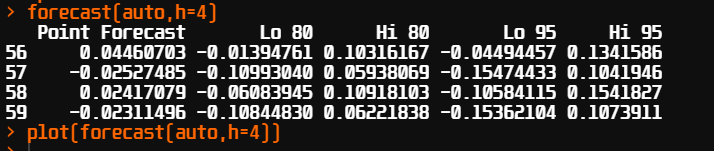


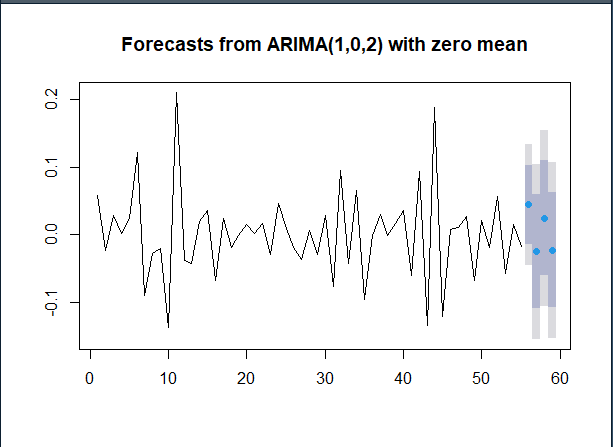
* I use auto ARIMA for the model.





* The box test p-value is higher than 0.05 so the model is acceptable. Here are the forecast result:





1. **Explain in fewer than 150 words how Random forest can be used in this case to predict the variable of interest for the 4 quarters in 2022.**

* Random forest is built from many Decision trees. Because the data is too little and not great for Machine learning algorithm. First, we have to see if the variables from before and after Covid-19 have the differences in their distribution by using many plots. If the data is distinguish from each class, we could use oversampling in SMOTE for balancing the data distribution between before and after Covid-19. The data may be duplicate but it still good for Machine Learning algorithm.
* We have to divide the data into train and test set( if necessary for evaluating the final performance of the model).
* Then we input the data to the Random forest model, the model will automatically calculate different Decision trees and then arregate the results of the trees.
* We could evaluate the result by using metrics. Then we can use the baseline model for predicting the 4 quarter in 2022.

**References:**

1. Al‐Sakran, S. A. (2001). Leverage determinants in the absence of corporate tax system: the case of non‐financial publicly traded corporations in Saudi Arabia. Managerial Finance.
2. A Kadim, K., & Nardi, S. (2018). Eviews Analysis: Determinant Of Leverage And Company’s Performance. Global and Stochastic Analysis (GSA), 5(7), 249-260.
3. Jermias, J., & Yigit, F. (2019). Factors affecting leverage during a financial crisis: Evidence from Turkey. Borsa Istanbul Review, 19(2), 171-185.
4. Ali, L. (2011). The determinants of leverage of the listed-textile companies in India. European Journal of Business and Management, 3(12), 54-59.
5. Onofrei, M., Tudose, M. B., Durdureanu, C., & Anton, S. G. (2015). Determinant factors of firm leverage: An empirical analysis at Iasi county level. Procedia Economics and Finance, 20, 460-466.
6. Karim, S., Rabbani, M. R., & Khan, M. A. (2021). Determining the key factors of corporate leverage in Malaysian service sector firms using dynamic modeling. Journal of Economic Cooperation and Development, 42(3), 1-20.