VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW

FINAL REPORT

Subject: PROGRAM PACKAGE IN FINANCE 2

LECTURER: PhD. Nguyen Thanh Liem

Student Name: Giáp Hoàng Long
ID: K194141728

Ho Chi Minh City, June 12th, 2022

Contents

1. Literature review

My topic belongs to leverage, so I found documents about the factors affecting capital structure, the ratio of financial leverage is influenced by 3 factors that I have chosen:

The Tangibility is determined by net fixed assets/ total assets, the Profitability of a business is determined through the ROA ratio and the Inventories to current assets. Here are all documents:

1. NGUYEN, C. D. T., DANG, H. T. T., PHAN, N. H., & NGUYEN, T. T. T. (2020). Factors affecting financial leverage: The case of Vietnam firms. *The Journal of Asian Finance, Economics and Business*, *7*(11), 801-808.

2. Šarlija, N., & Harc, M. (2012). The impact of liquidity on the capital structure: a case study of Croatian firms. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, *3*(1), 30-36.

3. Dũng, T. V., & Thanh, B. Đ. Các nhân tố ảnh hưởng đến cấu trúc vốn của các doanh nghiệp niêm yết trên Thị trường chứng khoán Việt Nam

4. Vijayakumaran, S., & Vijayakumaran, R. (2018). The determinants of capital structure decisions: Evidence from Chinese listed companies. Vijayakumaran, S., & Vijayakumaran, 63-81.

5. Alghusin, N. A. S. (2015). Do financial leverage, growth and size affect profitability of Jordanian industrial firms listed. International Journal of Academic Research in Business and Social Sciences, 5(4), 335-348.

The documents show that net fixed assets to total assets and inventories to current assets have a positive effect on the debt ratio. And profitability will have a negative effect on the ratio of financial leverage.

Variables	**Name**	**Measures**	**Expected Sign**
Leverage	**lev**	**total debt/ total assets**	
Profitability	**roa**	**return on assets**	**-**
Tangibility	**tang**	**net fixed assets/ total assets**	+
Inventories to current assets	**inv**	**inventory/ current assets**	+

2. Data collection and input

The data is collected from TCL| VietstockFinance, The firm chosen is Tan Cang Logistics & Stevedoring Joint Stock Company (HOSE: TCL) which engages in the port and container depot operation businesses in Vietnam. It also provides delivery service, loading and unloading service, packing service, fumigation service, etc. The company was founded in 2006 and is based in Ho Chi Minh City, Vietnam.

Raw data including 3 sheets: Balance Sheet, Income Statement, and Financial Ratio from quarter 3/ 2009 to quarter 1/ 2022. Here is the code for creating dataframe from excel and extracting the data necessary for the construction of variables (3 variables).

```
#Import Library

library(readxl)

library(pastecs)

library(tidyverse)

library(car)

library(zoo)

### Task 2

#Import dataset

df <- read_excel("K19141728.xlsx", sheet = "balance_sheet")

ratio <- read_excel("K19141728.xlsx", sheet = "ratio")

# Pick and transform needed variables

df["lev"] = df["Liabilities"]/ df["Total assets"]

df["tang"] = df["Fixed assets"]/ df["Total assets"]

df["inv"] = df["Inventories"]/ df["Current assets"]

df["roa"] = ratio["ROA"]/100

#List of independent and dependent variables

col = c("Quarter", "lev", "tang", "inv", "roa")

#Create dataframe

df = df[col]
```

RStudio Source Editor

df

Filter

	Quarter	lev	tang	inv	roa
1	3/2009	0.2930591	0.3290488	0.091346154	0.1184
2	4/2009	0.3144796	0.2986425	0.046025105	0.0779
3	1/2010	0.3590308	0.3105727	0.099137931	0.0509
4	2/2010	0.3875969	0.2616279	0.058181818	0.0554
5	3/2010	0.3780919	0.2420495	0.046439628	0.0400
6	4/2010	0.3872549	0.3218954	0.087096774	0.0420
7	1/2011	0.3972835	0.3446520	0.105633803	0.0302
8	2/2011	0.3704918	0.3196721	0.065359477	0.0435
9	3/2011	0.5269953	0.4061033	0.056122449	0.0285
10	4/2011	0.4850829	0.4397790	0.099173554	0.0318
11	1/2012	0.4977324	0.4342404	0.090909091	0.0192
12	2/2012	0.5085653	0.3929336	0.118203310	0.0198
13	3/2012	0.4742015	0.3980344	0.083582090	0.0256
14	4/2012	0.4969250	0.3936039	0.086153846	0.0161
15	1/2013	0.4584488	0.3905817	0.021276596	0.0305
16	2/2013	0.4139715	0.4036223	0.012121212	0.0278
17	3/2013	0.4025157	0.3761006	0.024000000	0.0310

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

```
### Task 3

#Create dataframe before and after pandemic

df_before = df[1:42,]

df_after = df[43:51,]

#Descriptive statistics

summary(df_before)

summary(df_after)
```

```
> #Descriptive statistics
> summary(df_before)
   Quarter              lev              tang              inv                 roa
 Length:42          Min.   :0.2471   Min.   :0.2420   Min.   :0.004310   Min.   :0.01610
 Class :character   1st Qu.:0.3150   1st Qu.:0.3222   1st Qu.:0.008728   1st Qu.:0.02580
 Mode  :character   Median :0.3757   Median :0.3744   Median :0.015484   Median :0.02930
                    Mean   :0.3727   Mean   :0.3644   Mean   :0.035006   Mean   :0.03307
                    3rd Qu.:0.4046   3rd Qu.:0.4047   3rd Qu.:0.057667   3rd Qu.:0.03252
                    Max.   :0.5270   Max.   :0.5045   Max.   :0.118203   Max.   :0.11840
> summary(df_after)
   Quarter              lev              tang              inv                 roa
 Length:9           Min.   :0.2974   Min.   :0.3224   Min.   :0.006276   Min.   :0.01750
 Class :character   1st Qu.:0.3288   1st Qu.:0.3432   1st Qu.:0.009153   1st Qu.:0.01980
 Mode  :character   Median :0.3313   Median :0.3456   Median :0.015075   Median :0.02790
                    Mean   :0.3651   Mean   :0.3473   Mean   :0.013148   Mean   :0.02651
                    3rd Qu.:0.4245   3rd Qu.:0.3555   3rd Qu.:0.017682   3rd Qu.:0.03090
                    Max.   :0.4792   Max.   :0.3706   Max.   :0.018182   Max.   :0.03830
```

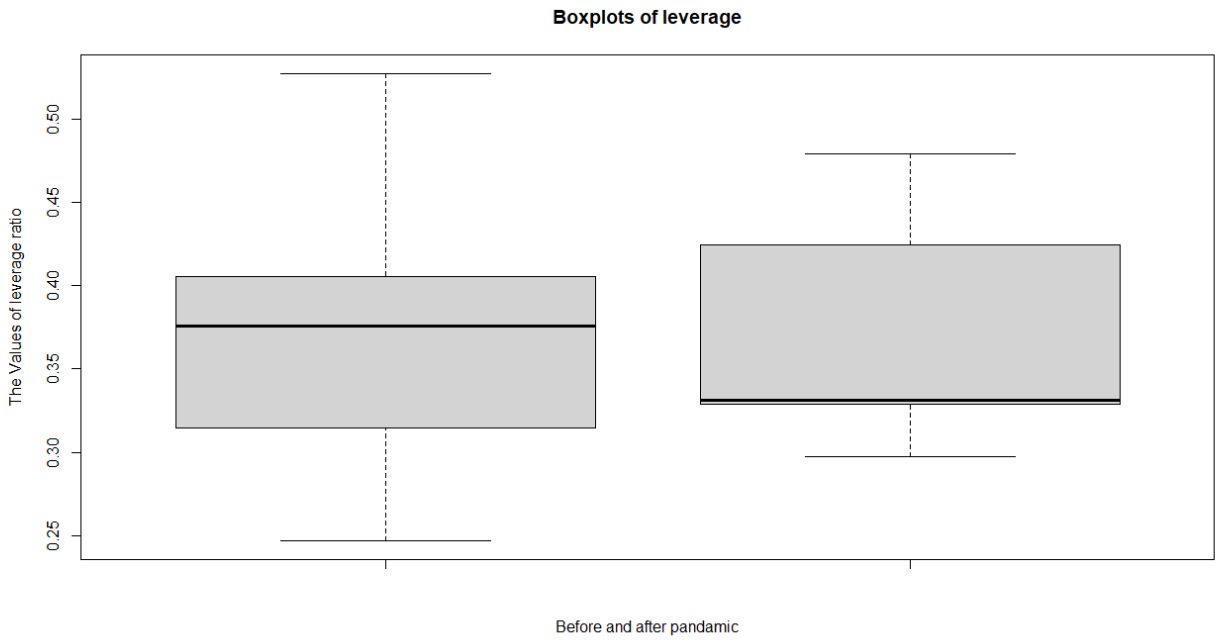
The first is that the leverage ratio of enterprises after the pandemic has increased slightly at the lowest level, but the good control after the pandemic makes the maximum level of leverage to be reduced. Therefore, the average also decreased slightly.

The fixed assets on total assets (tang) has the min increase but the max decrease. It can be seen that after the pandemic, enterprises always maintain fixed assets at the desired level and do not have too large fluctuations to ensure production activities as well as avoid unnecessary risks.

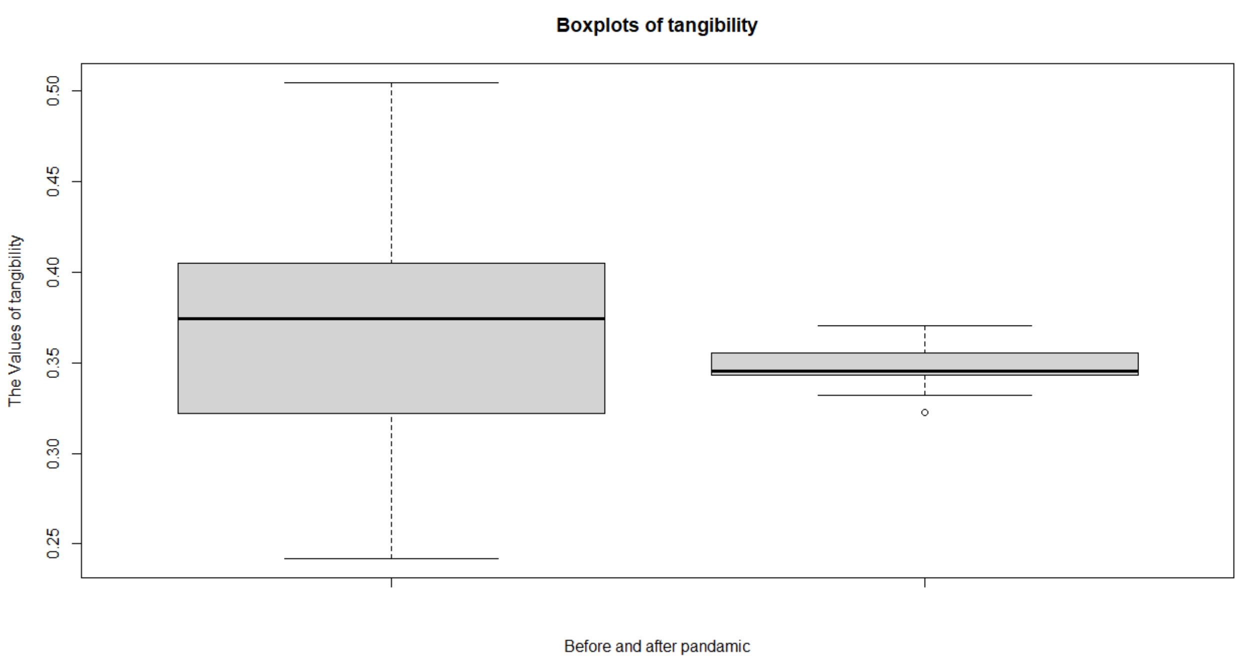
After the covid pandemic, the inventories to current assets has fallen sharply when the mean of the variable is more than halved. Inventory stock is always maintained at a low level. This is easy to explain because, during the closure of countries, the amount of goods arriving at ports is very limited, as well as businesses avoid the situation that logistics cannot import and export.

Roa has always been kept above its pre-pandemic lowest, but there are no more explosive quarters in revenue. The business still achieved return on assets as expected.

```
#Explore Data Analysis

boxplot(df_before$lev, df_after$lev, main = "Boxplots of leverage",

                    xlab = "Before and after pandamic",

                    ylab = "The Values of leverage ratio")
```

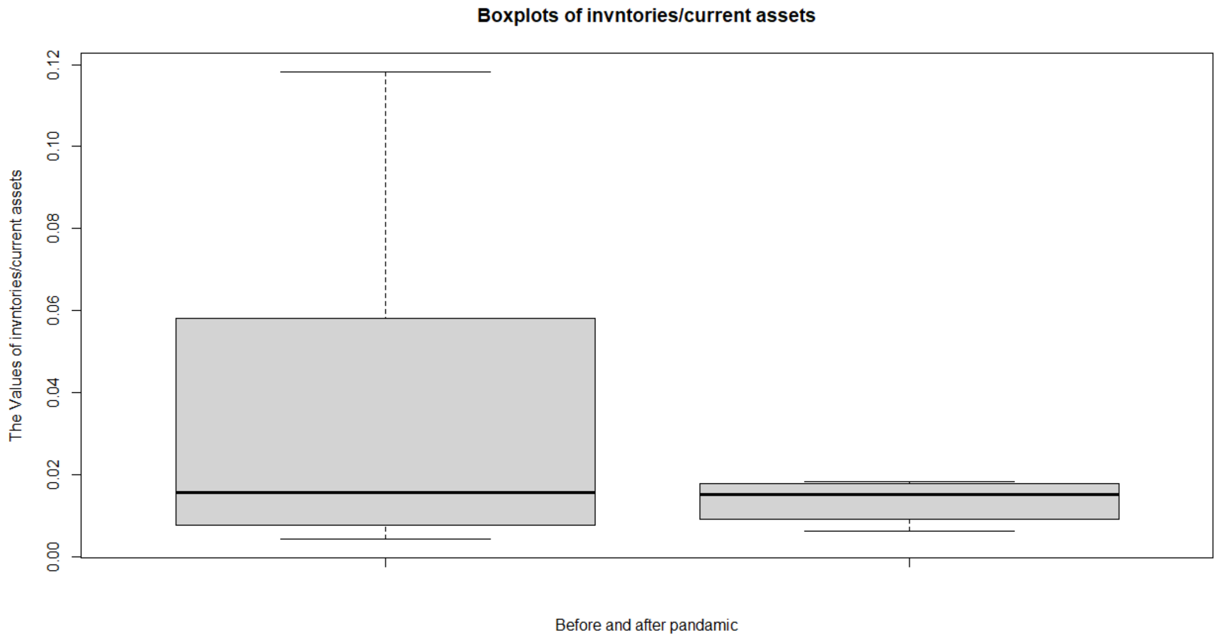



```
boxplot(df_before$tang, df_after$tang, main = "Boxplots of tangibility",
        xlab = "Before and after pandamic",
        ylab = "The Values of tangibility")
```

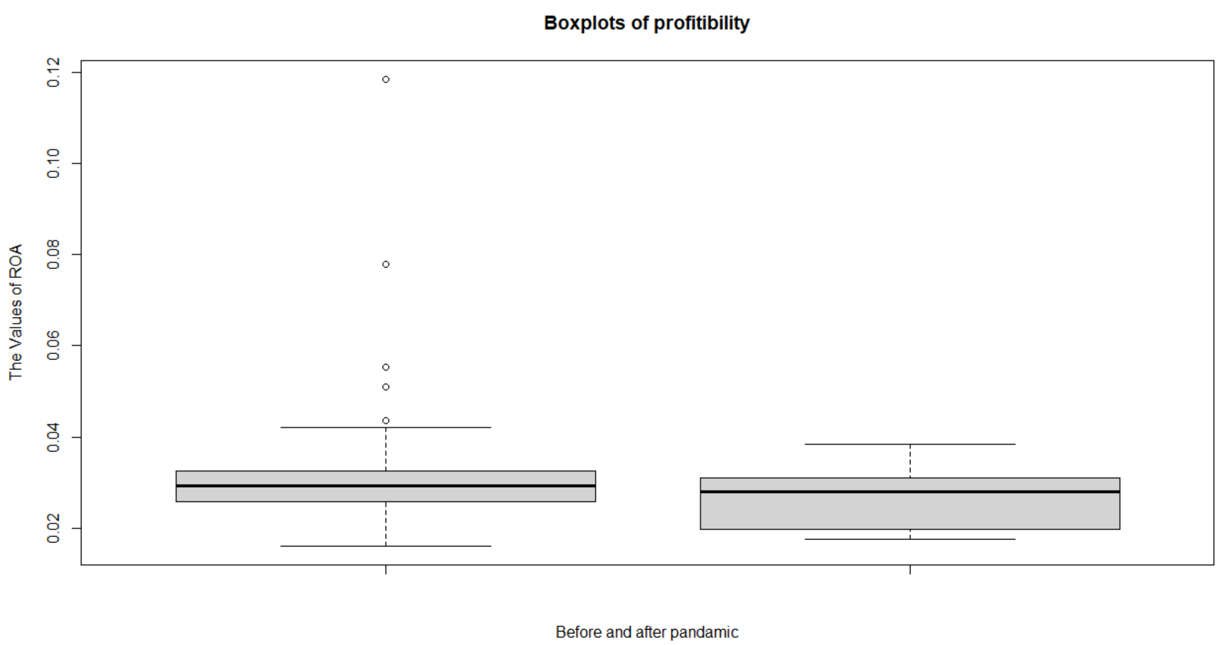



```
boxplot(df_before$inv, df_after$inv, main = "Boxplots of inventories/current assets",
```

```
    xlab = "Before and after pandamic",
    ylab = "The Values of invntories/current assets")
```




```
boxplot(df_before$roa, df_after$roa, main = "Boxplots of profitibility",
    xlab = "Before and after pandamic", ylab = "The Values of ROA")
```



#Standard dviation

stat.desc(df_before)[13,]

stat.desc(df_after)[13,]

```
> #Standard dviation
> stat.desc(df_before)[13,]
        Quarter        lev       tang        inv        roa
std.dev      NA 0.07180413 0.06483407 0.03561523 0.01725265
> stat.desc(df_after)[13,]
        Quarter        lev       tang         inv         roa
std.dev      NA 0.06238115 0.01470909 0.004615312 0.007705427
```

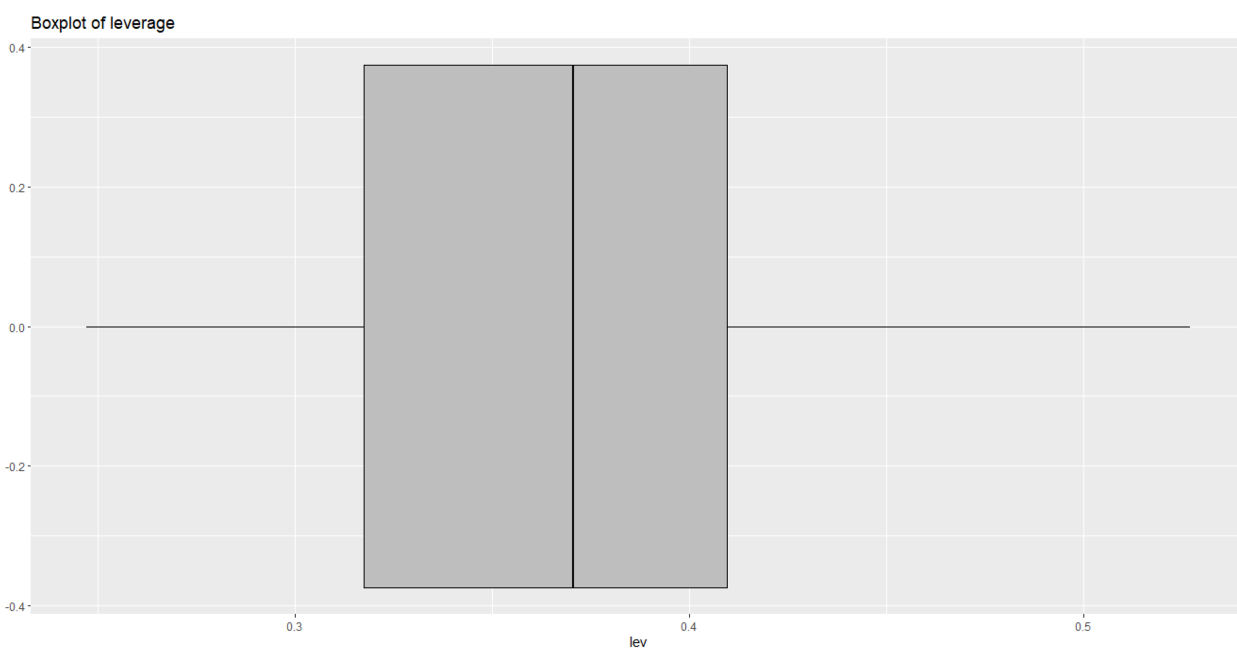
Except lev, the standard deviations of the other variables all decreased sharply after the pandemic

4. Provide box & whisker plot and histogram of the variable Leverage

###Task 4

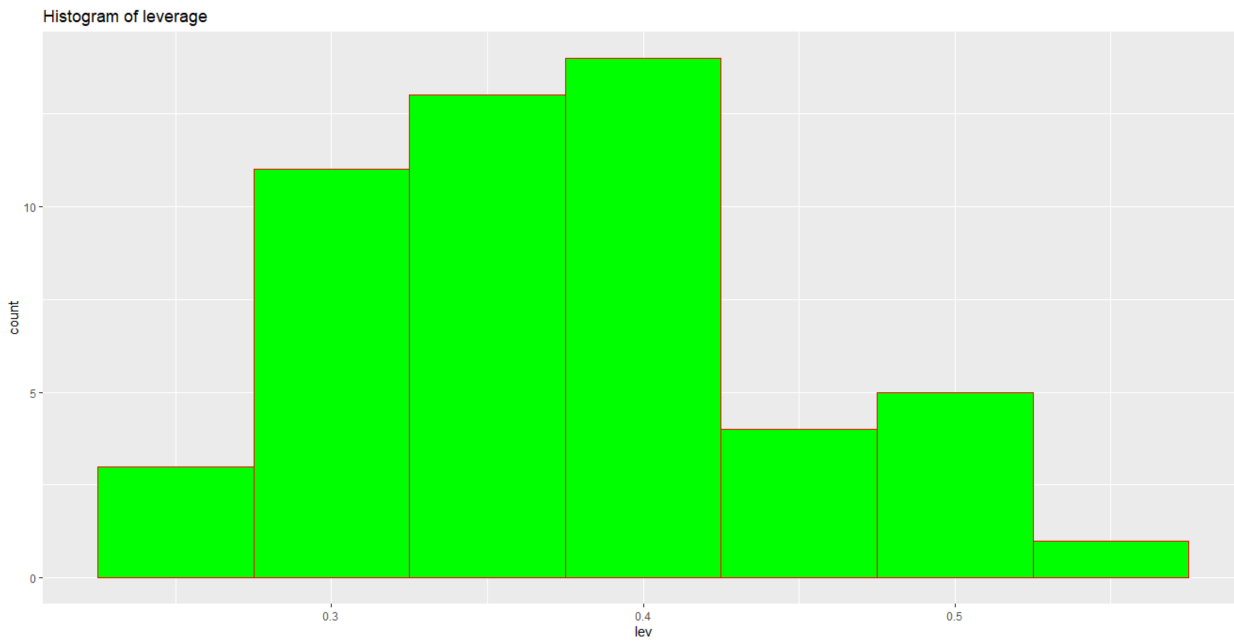
#box & whisker plot and histogram of the leverage

```
ggplot(df, aes(x= lev)) +
  geom_boxplot(col="black",
               fill="gray") +
  ggtitle("Boxplot of leverage")
```



##Histogram

```
ggplot(df, aes(x= lev)) +
  geom_histogram(bins =20,col="red",
                 fill="green",
                 binwidth = 0.05)+
  ggtitle("Histogram of leverage")
```



As the boxplot and histogram show, the debt-to-total asset is mainly concentrated from about 30% to over 40% of total assets. For an enterprise with many state shares, this is considered a reasonable and safe. Contacting boxplot before and after the pandemic, this is a business that does not lack money and does not want to use much debt in its capital structure.

5. Perform multiple regression to determine the significant determinants of the variable of assigned topic. The significance level is 10%.

```
### Task 5
##5.1
# Model multiple regression
model <-lm(lev ~ roa + tang + inv, data = df)
summary(model)
```

```
Call:
lm(formula = lev ~ roa + tang + inv, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.079992 -0.033555 -0.006464  0.022932  0.160137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20885    0.05039   4.145 0.000141 ***
roa         -1.17598    0.48070  -2.446 0.018226 *
tang         0.45231    0.12382   3.653 0.000651 ***
inv          1.17481    0.22605   5.197 4.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04988 on 47 degrees of freedom
Multiple R-squared:  0.5187,	Adjusted R-squared:  0.4879
F-statistic: 16.88 on 3 and 47 DF,  p-value: 1.405e-07
```

All variables and models are statistically significant at the significance level of 0.1, but R-Square is very low 51,87% the ability to explain the dependent variable. However, p-value of F-test is pretty small. The coefficients of variables like what we expect in the literature review. ROA has a negative relationship with the leverage ratio and the other variables have positive relationship. Therefore, model is still good.

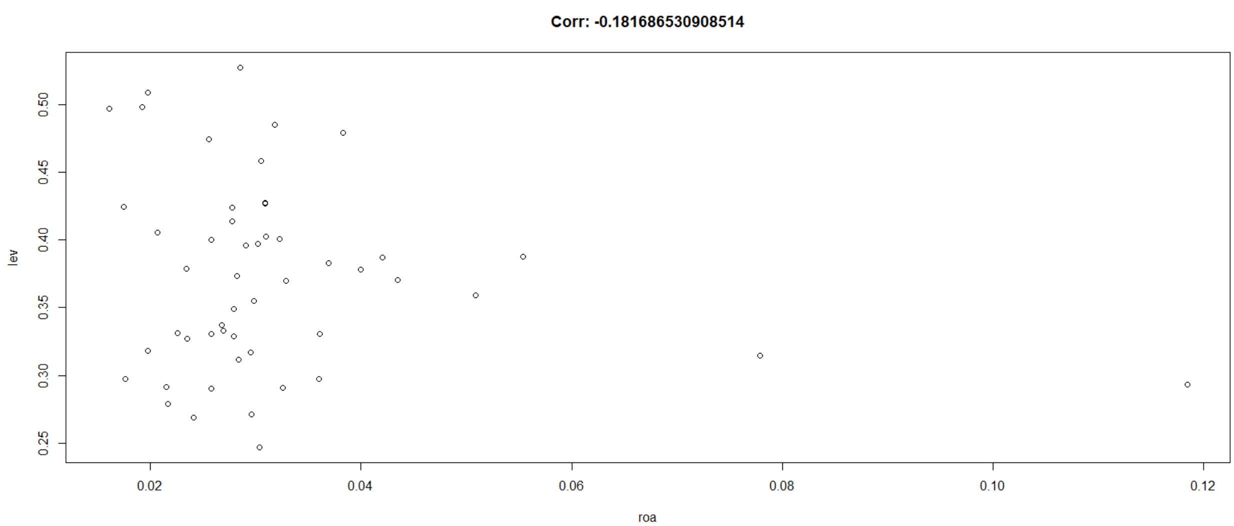
When ROA increases by 1 unit, leverage will decrease by 1.17598 times

When Tangibility is increased by 1 unit, the leverage ratio increases by 0.45231 times

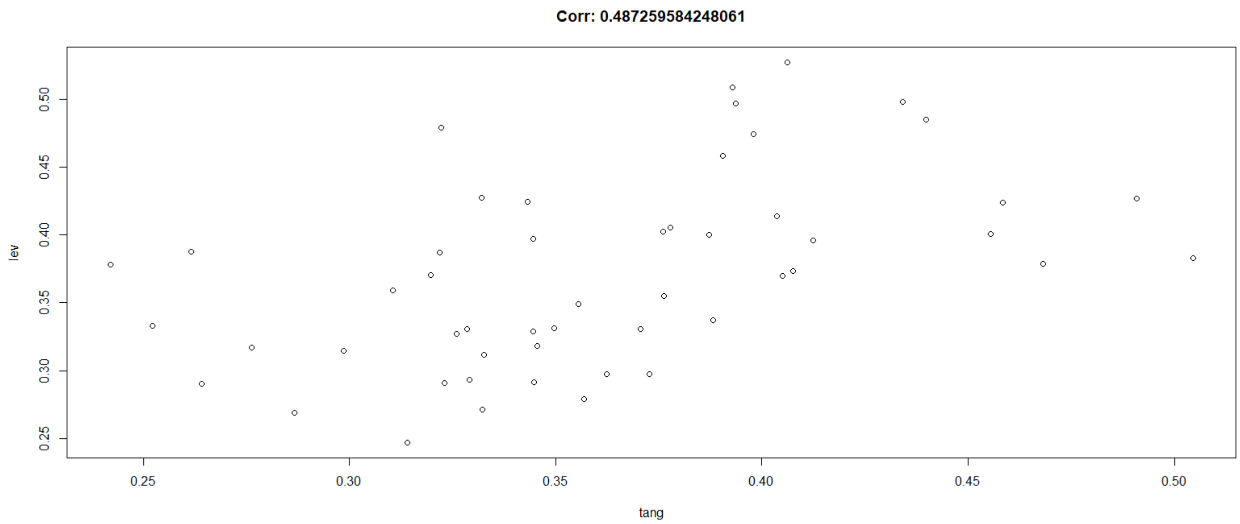
When inventory to current assets increases by 1 unit, the leverage ratio increases by 1.17481 times

Check Linearity

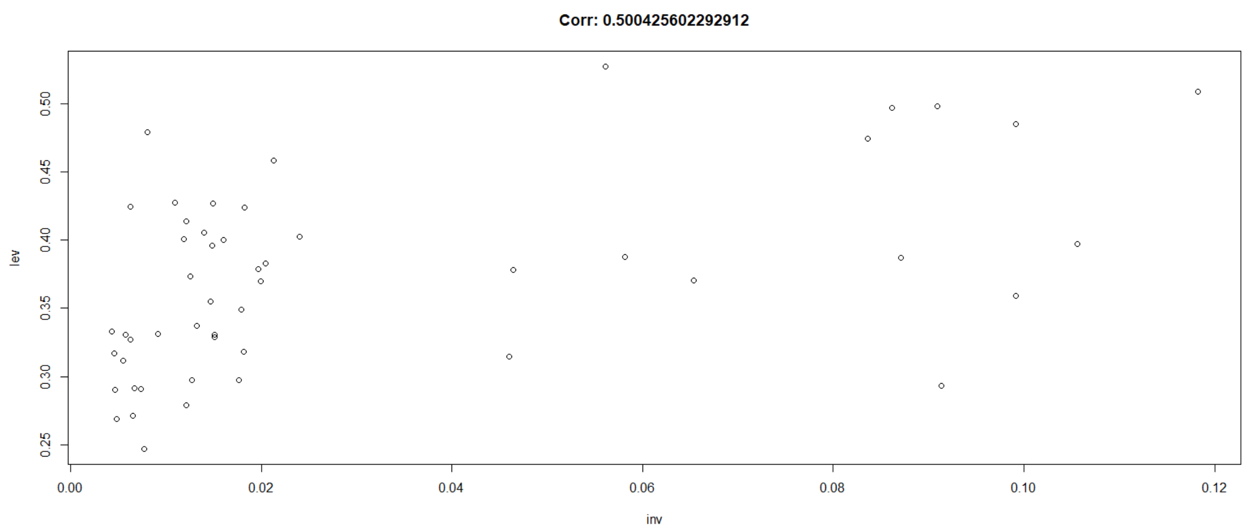
plot(lev ~ roa, data= df,main =paste("Corr:",cor(dflev, dfroa)))



plot(lev ~ tang, data= df,main =paste("Corr:",cor(dflev, dftang)))



plot(lev ~ inv, data= df,main =paste("Corr:",cor(dflev, dfinv)))



Check Multicollinearity

vif(lm(lev ~ roa + tang + inv, data = df))

```
> # Check Multicollinearity
> vif(lm(lev ~ roa + tang + inv, data = df))
     roa     tang      inv 
1.207109 1.086012 1.144252 
> 
```

* There is no multicollinearity

Check important assumptions

par(mfrow=c(2,2))

plot(model)

shapiro.test(resid(model)) # Null hypothesis is normality

#Shapiro-Wilk normality test

library(lmtest) #Null hypothesis is homoskedasticity

bptest(model)

```
> shapiro.test(resid(model)) # Null hypothesis is normality

        Shapiro-Wilk normality test

data:  resid(model)
W = 0.95439, p-value = 0.04817

> library(lmtest) #Null hypothesis is homoskedasticity
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 1.9173, df = 3, p-value = 0.5897
```

* There is no Heteroscedasticity

##5.2

Create covid dummy variable (before covid: 0, after covid: 1)

df["covid"] = 0

df[43:51,"covid"] = 1

Model with the interaction between Covid-19 dummy variable and the independent variables

model_dummy <-lm(lev ~ roa + tang + inv + roa*covid +tang*covid +inv*covid, data = df)

summary(model_dummy)

```
Call:
lm(formula = lev ~ roa + tang + inv + roa * covid + tang * covid +
    inv * covid, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.075615 -0.026831 -0.000354  0.020565  0.100237

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19144    0.04273   4.480 5.44e-05 ***
roa         -1.33663    0.40447  -3.305  0.00192 **
tang         0.49457    0.10260   4.821 1.82e-05 ***
inv          1.29308    0.18928   6.831 2.26e-08 ***
covid        0.89461    0.42076   2.126  0.03927 *
roa:covid    3.52927    1.97462   1.787  0.08094 .
tang:covid  -2.50983    1.24563  -2.015  0.05019 .
inv:covid   -7.31273    3.87165  -1.889  0.06568 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04079 on 43 degrees of freedom
Multiple R-squared:  0.7056,    Adjusted R-squared:  0.6576
F-statistic: 14.72 on 7 and 43 DF,  p-value: 1.314e-09
```

The model's R-Square improved with an explanatory rate of 70.56% when adding dummy variables that interact with the independent variables.

The explanatory variables and the model both have very small p-values and are completely statistically significant at the alpha level of 10%.

Check important assumptions

par(mfrow=c(2,2))

plot(model_dummy)

Null hypothesis is normality

#Shapiro-Wilk normality test

shapiro.test(resid(model_dummy))

#Null hypothesis is homoskedasticity

library(lmtest)

bptest(model_dummy)

```
> # Null hypothesis is normality
> #Shapiro-Wilk normality test
> shapiro.test(resid(model_dummy))

        Shapiro-Wilk normality test

data:  resid(model_dummy)
W = 0.98291, p-value = 0.6676

> #Null hypothesis is homoskedasticity
> library(lmtest)
> bptest(model_dummy)

        studentized Breusch-Pagan test

data:  model_dummy
BP = 3.3323, df = 7, p-value = 0.8527
```

There is no multicollinearity

##5.3

#define new observation

new = df[,3:5]

#use the fitted model to predict the value for the new observation

pred = predict(model, newdata = new)

print(pred)

```
> pred = predict(model, newdata = new)
> print(pred)
        1         2         3         4         5         6         7         8         9 
0.3257557 0.3063864 0.4059316 0.3303858 0.3258456 0.4073732 0.4533203 0.3790664 0.4249482 
       10        11        12        13        14        15        16        17        18 
0.4868762 0.4894793 0.5021556 0.4569688 0.4691580 0.3746385 0.3729562 0.3707000 0.3767852 
       19        20        21        22        23        24        25        26        27 
0.3725315 0.3907515 0.4120414 0.4176887 0.4161691 0.4049456 0.3786855 0.3747635 0.3718413 
       28        29        30        31        32        33        34        35        36 
0.3613025 0.3684099 0.3500318 0.3590338 0.3323224 0.3252170 0.3242116 0.3158150 0.3045423 
       37        38        39        40        41        42        43        44        45 
0.2963761 0.3035434 0.3474002 0.3360203 0.3338544 0.3319979 0.3728627 0.3632500 0.3508779 
       46        47        48        49        50        51 
0.3517944 0.3511919 0.3190496 0.3354856 0.3578446 0.3495932 
> 
```

#RMSE

library(Metrics)

rmse(df$lev, pred)

```
> rmse(df$lev, pred)
[1] 0.04788507
>
```

Using model 1 to predict the dependent variable for all quarterly data, the RSME is quite small at 0.047. This means that the model gives a low and reliable prediction result

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

###6

#import lib

library(forecast) #forecast, accuracy

library(tseries) #adf.test

library(lmtest) #coeftest

library(stats) #Box.test

par(mfrow=c(1,1))

Transform the variable into time series

#(the beginning quarter starts from q3/2019 and the frequency is 4 quarters each year)

ts = ts(df$lev,start = c(2009,3),frequency = 4)

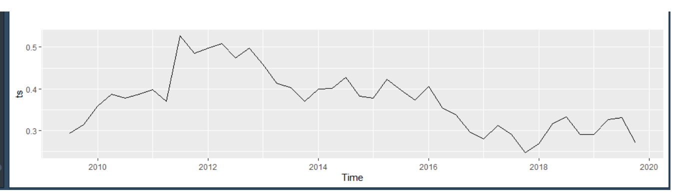
Visualize time series data

autoplot(ts)

#Check stationary

adf.test(ts, k=4) #Since p is greater than significance level, the Series is NON Stationary

```
> # Visualize time series data
> autoplot(ts)
> #Check stationary
> adf.test(ts, k=4) #Since p is greater than significance level, the Series is
 NON Stationary

        Augmented Dickey-Fuller Test

data:  ts
Dickey-Fuller = -2.6808, Lag order = 4, p-value = 0.3053
alternative hypothesis: stationary
```



Decompose time series and check it again

ts_d1 = diff(ts, differences = 1)

adf.test(ts_d1, k=4) # p-value is still greater than significance level, the Series is NON Stationary

```
> # Decompose time series and check it again
> ts_d1 = diff(ts, differences = 1)
> adf.test(ts_d1, k=4) # p-value is still greater than significance level, the Series is NON Stationary

        Augmented Dickey-Fuller Test

data:  ts_d1
Dickey-Fuller = -2.1852, Lag order = 4, p-value = 0.5007
alternative hypothesis: stationary
```

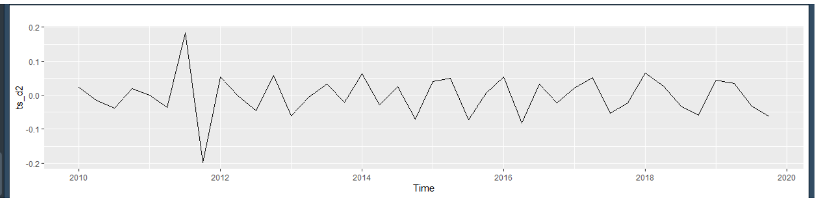
#Second Difference

ts_d2 = diff(ts, differences = 2)

adf.test(ts_d2, k=4)

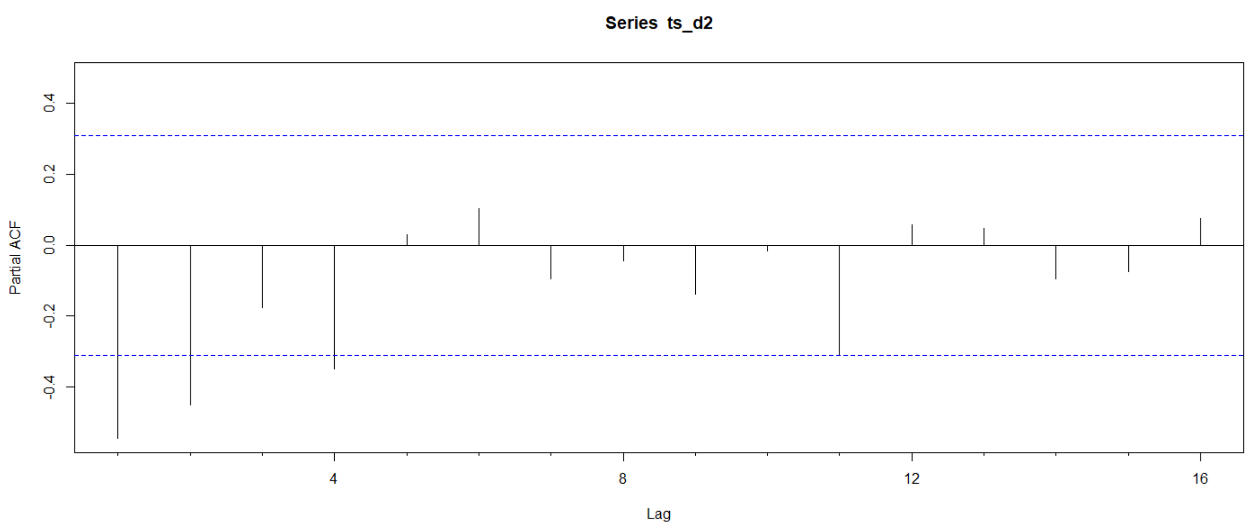
autoplot(ts_d2)

```
> #Second Difference
> ts_d2 = diff(ts, differences = 2)
> adf.test(ts_d2, k=4)

        Augmented Dickey-Fuller Test

data:  ts_d2
Dickey-Fuller = -4.1148, Lag order = 4, p-value = 0.01531
alternative hypothesis: stationary

> autoplot(ts_d2)
>
```

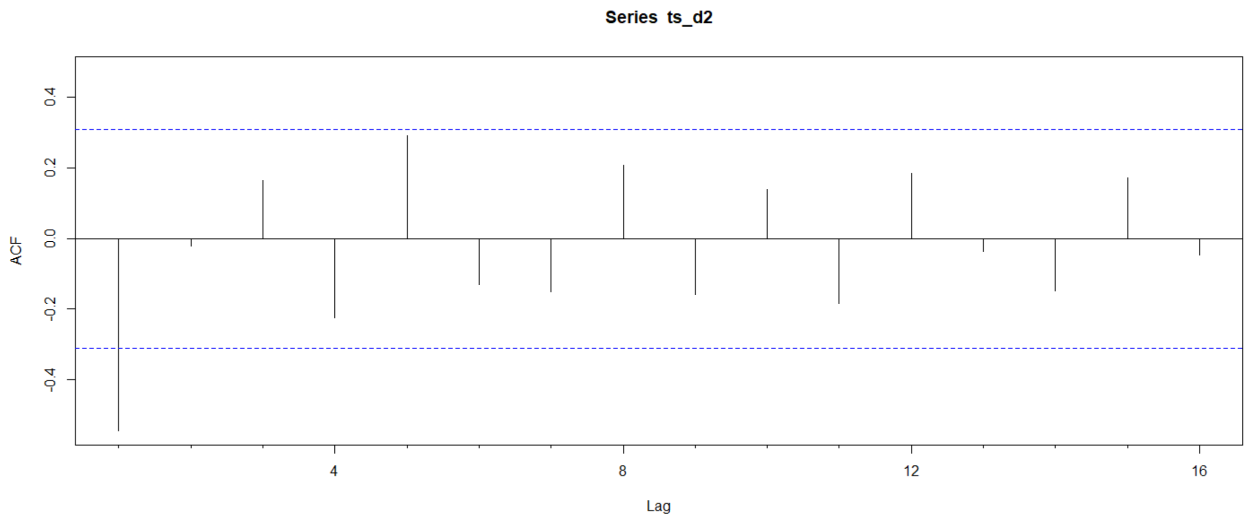


==> d =2 , the d in term ARIMA(p,2,q)

Pacf(ts_d2) # => p =4



Acf(ts_d2) # => q = 0



Finally, I have appropriate p,d,q for the ARIMA model. That is: ARIMA(4,2,0)

Model

mod = Arima(ts, order= c(4,2,0))

Summary model

summary(mod)

```
> summary(mod)
Series: ts
ARIMA(4,2,0)

Coefficients:
          ar1      ar2      ar3      ar4
      -1.0354  -0.9624  -0.6177  -0.3834
s.e.   0.1485   0.2029   0.2024   0.1460

sigma^2 = 0.00169:  log likelihood = 72.14
AIC=-134.28   AICc=-132.51   BIC=-125.83

Training set error measures:
                       ME       RMSE        MAE       MPE    MAPE      MASE       ACF1
Training set -0.003409008 0.03806332 0.02929776 -1.098784 7.78139 0.5738371 -0.038947
```

Check p-value

coeftest(mod) # The model is completely statistically significant

```
> # Check p-value
> coeftest(mod) # The model is completely statistically significant

z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1 -1.03538    0.14851 -6.9716 3.134e-12 ***
ar2 -0.96242    0.20293 -4.7426 2.110e-06 ***
ar3 -0.61774    0.20244 -3.0516  0.002277 **
ar4 -0.38338    0.14596 -2.6266  0.008623 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

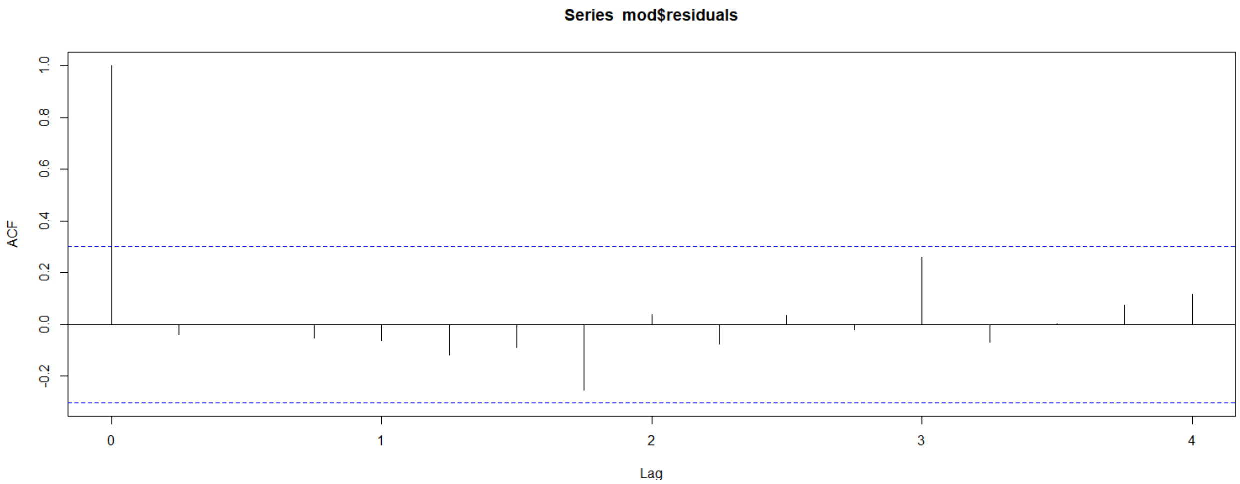
AutoCorrelation of Residuals test

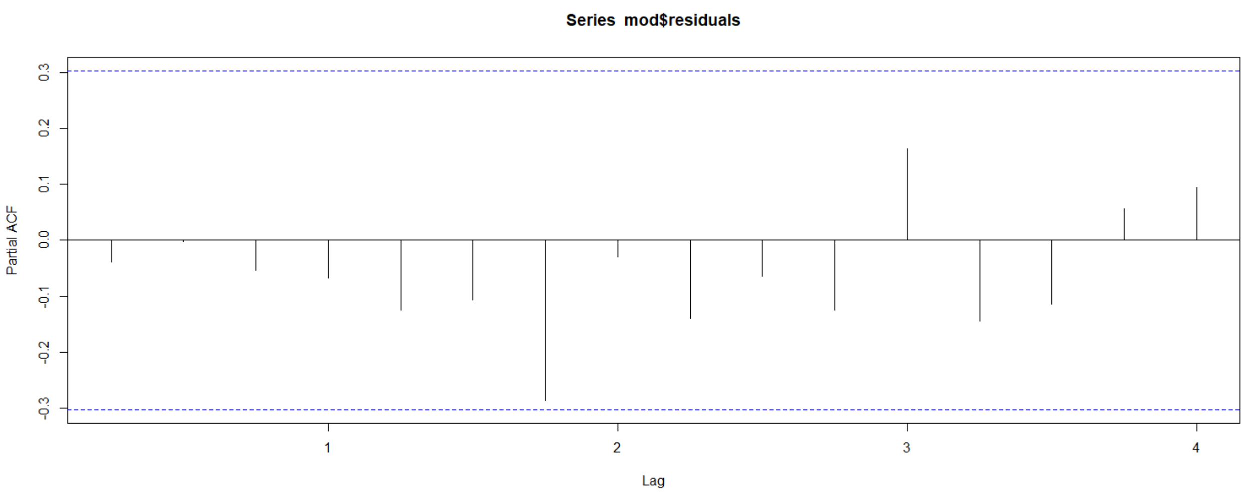
acf(mod$residuals)

pacf(mod$residuals)

Box.test(mod$residuals,lag=12,type='Ljung-Box')

```
> Box.test(mod$residuals,lag=12,type='Ljung-Box')

        Box-Ljung test

data:  mod$residuals
X-squared = 9.5375, df = 12, p-value = 0.6565
```





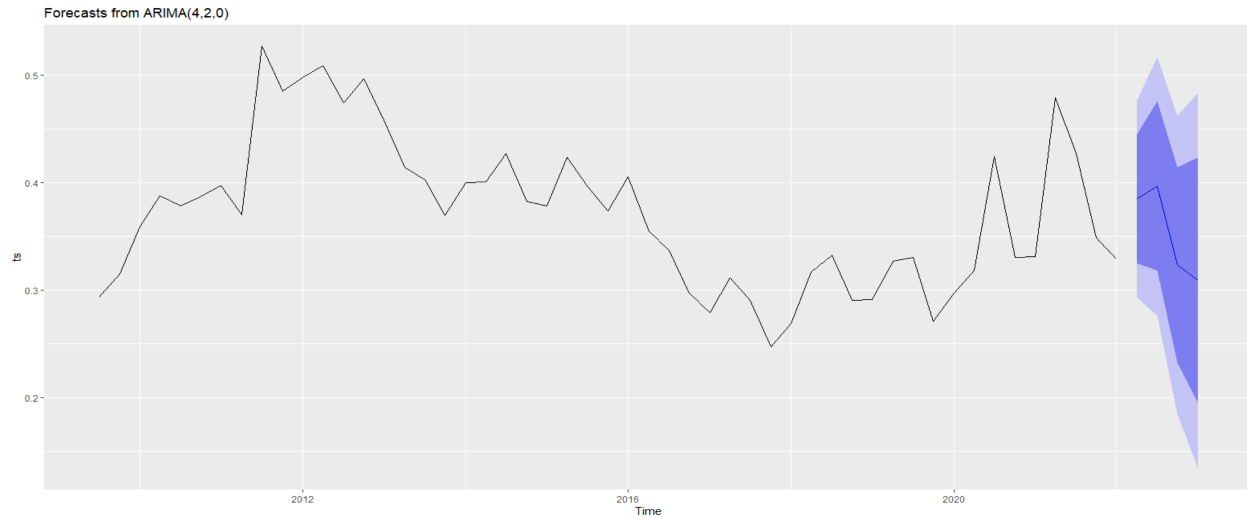
Forecast next 4 quarters in 2022 and 2023

forecast(mod, h =4)

```
> # Forecast next 4 quarters in 2022 and 2023
> forecast(mod, h =4)
        Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
2022 Q2      0.3852010 0.3251388 0.4452631 0.2933439 0.4770580
2022 Q3      0.3967219 0.3179673 0.4754765 0.2762771 0.5171667
2022 Q4      0.3233734 0.2324029 0.4143438 0.1842461 0.4625006
2023 Q1      0.3088470 0.1943096 0.4233844 0.1336772 0.4840169
```

Visualization the forecast

autoplot(forecast(mod,h=4))



The model ARIMA(4,2,0) is selected and gives a model that is completely statistically significant in the variables.

AIC level = -134.28

The indicators measuring the model's prediction error on the training set give acceptable results.

Autocorrelation of residuals test shows that the model does not violate the assumption of autocorrelation between residuals

7. Explain how Random forest can be used in this case to predict the variable of interest for the 4 quarters in 2022.

When using the Random Forest Regression to predict the Leverage Ratio, it is important to first identify the most important variables that directly affect it. If the variable is a random walk then it will be complicated to predict. Another thing is that when the input variable has new data and is outside the recognition level of the algorithm, the possibility of the resulting output is not good.

The classification problem of the Random Forest algorithm can be used to predict the probability that the dependent variable will increase or decrease in that quarter by encoding the dependent variable. Let the algorithm learn and make recommendations based on a tree diagram showing which elements have the best classification ability. Make a conclusion about how the increase or decrease in the features will affect the output.