Configuration Manual

MSc Research Project
Fintech

Tung Nguyen
Student ID: 21102759

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Tung Nguyen
Student ID:	21102759
Programme:	Msc in Fintech **Year:** 2021-2022
Module:	Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	15th August 2022
Project Title:	Dividend Pay-out Policy of Listed Companies on the Ho Chi Minh City Stock Exchange
Word Count:	1225 **Page Count:** 6

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……………………………………………………………

Date: 14th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Tung Nguyen
Student ID: 21102759

1 Introduction

This configuration manual describes all the technical details that were utilized to realize the research project titled "*Dividend Pay-out Policy of Listed Companies on the Ho Chi Minh City Stock Exchange*". Specifically, the hardware configuration of the machine being used, and the software used for data analysis will be presented in the next sections. Also, the necessary steps needed to replicate the research will be laid out to ensure that it can be easily reproduced, and the results presented in the thesis paper can be examined without any obstacle.

2 Hardware Configuration

The configuration of the device used for undertaking the research is presented in Table 1.

Table 1. Hardware Configuration

Processor	Intel® Core™ i3-8130U CPU @ 2.20GHz (4 Core)
Memory	12288MB RAM
System Configuration	64-bit operating system
Operating System	Windows 10 Home

3 Software Configuration

This research utilizes the R Studio software to process and analyse the data, with the configuration being summarized in Table 2.

Table 2. Software Configuration

Programming language	R version 4.1.1
Software Version	RStudio version 2021.09.0

4 Steps taken in the Research Project

Each subsection below illustrates the steps taken to realize this research project

4.1 Importing libraries

In R Studio, many functions used in performing data processing and analysis require packages to be installed and imported. The packages which were used for this research project are summarized in Table 2 below

Table 3. Imported Libraries

Package Name	Version
ggplot2	3.3.5
ggthemes	4.2.4
dplyr	1.0.7
corrplot	0.92
caret	6.0.90
pROC	1.18.0
ROSE	0.0.4
plm	2.6.1
lme4	1.1.29
zoo	1.8.10
Metrics	0.1.4

4.2 Loading data

The dataset is loaded into the software using the following code:

```
setwd("C:\\Users\\ASUS\\Documents\\NCI\\Thesis\\Data\\Final Dataset")
df <- read.csv("Final_Dataset.csv", header = TRUE)
```

4.3 Data Pre-Processing and Transformation

After loading, the dataset's structure is examined using the following code:

```
# Checking the dataset structure
str(df)
```

The results from this code show that the dependent as well as independent variables all have numeric type in accordance with the model's assumptions. Next, the correlation table of the dataset is drawn to check if there's multicollinearity between the independent variables, using the following code:

```
# Checking correlation between the variables
df3 <- df %>% select_if(is.numeric)
dfcor <- round(cor(df3),2)
corrplot(dfcor, type = "full", order = "hclust", tl.col = "black", tl.srt = 45,
addCoef.col = "black")
```

The correlation table proves that there's no multicollinearity in the dataset; therefore, all the independent variables can be used in the statistical learning model. Next, the boxplot distribution of the dependent variable is drawn by applying the following code:

```
# Box Plot distribution of the Dependent Variable
ggplot(data = df, aes(y=DPR)) +
  geom_boxplot() +
  ggtitle("Dividend Pay-out Ratio") +
  theme_few() +
  theme(axis.title.y = element_blank(), plot.title = element_text(hjust = 0.5))
```

The *Period* column is transformed into time-series quarterly type to reflect the financial period of each company, using the below code:

```
# Data Transformation
df$Period <- paste(substr(df$Period,4,8), substr(df$Period,1,2))
df$Period <- as.yearqtr(df$Period)
```

All the variables in the dataset are normalized using the min-max normalization technique as below:

```
# Normalization of Data
normal <- function(x) {
  (x-min(x))/(max(x)-min(x))
}
df$DPR <- normal(df$DPR)
df$Growth <- normal(df$Growth)
df$ROA <- normal(df$ROA)
df$Leverage <- normal(df$Leverage)
df$Liquidity <- normal(df$Liquidity)
df$Size <- normal(df$Size)
df$Risk <- normal(df$Risk)
df$GDP <- normal(df$GDP)
```

4.4 Choosing the statistical model for predicting dividend pay-out ratio

First, the dataset is split into a training and a test partition using the below code:

```
# Create Training and Test Partitions
data_train <- filter(df, Year < 2018)
data_test <- filter(df, Year > 2017)
```

Then the three types of panel regression are applied on the training dataset using the code below:

```
# Panel regression - Pooled OLS
plm_ols <- plm(formula = DPR ~ ROA + Growth + Leverage + Risk + Size +
Liquidity + GDP, data = data_train, index = c("Ticker"), model =
"pooling")
summary(plm_ols)

# Panel regression - Fixed Effects
plm_fix <- plm(formula = DPR ~ ROA + Growth + Leverage + Risk + Size +
Liquidity + GDP, data = data_train, effect = "individual", index =
c("Ticker"), model = "within")
within_intercept(plm_fix)
summary(plm_fix)

# Panel regression - Random Effects
```

```
plm_ran <- plm(formula = DPR ~ ROA + Growth + Leverage + Risk + Size +
Liquidity + GDP, data = data_train, effect = "individual",  index =
c("Ticker"), model = "random")
summary(plm_ran)
```

Stistical tests are utilized to select the most suitable model for this dataset by applying the code as follows:

```
# Testing for the right model
pFtest(plm_fix, plm_ols)
phtest(plm_fix, plm_ran)
plmtest(plm_ols, type=c("bp"))
```

The test results show that the Pooled OLS Regression is the most suitable model. As such, this model is selected to predict the dividend pay-out ratio on the test partition

```
# Running Prediction on Pooled OLS model
data_test$ols <- predict(plm_ols, newdata = data_test)
```

4.5 Evaluation

After obtaining the prediction from running the Pooled OLS Regression model, error metrics are calculated to evaluate the prediction against the actual values.

```
# Evaluating the Forecast results
mae(data_test$DPR, data_test$ols)
smape(data_test$DPR, data_test$ols)
rmse(data_test$DPR, data_test$ols)
```

The residuals plot of the prediction is drawn to examine the distribution of the residuals

```
# Residuals plot
plm_ols_test <- plm(formula = DPR ~ GDP + Risk + ROA + Leverage + Growth + Size +
Liquidity, data = data_test, index = c("Ticker","Period"), model = "pooling")
data_test$residuals <- residuals(plm_ols_test)
ggplot(data = data_test, aes(x=ols, y=residuals)) +
  geom_point() +
  theme_few() +
  ggtitle("Residuals Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Prediction") +
  ylab("Residuals")
```

As the dataset is a time-series, the density plot of the DPR is drawn to examine the distribution of the dependent variable. The plot is shown in Figure 1, showing the outliers in the DPR.

```
# Density Plot of the Dependent Variable
plot(density(df$DPR))
plot(density(df$DPR^0.5))
```

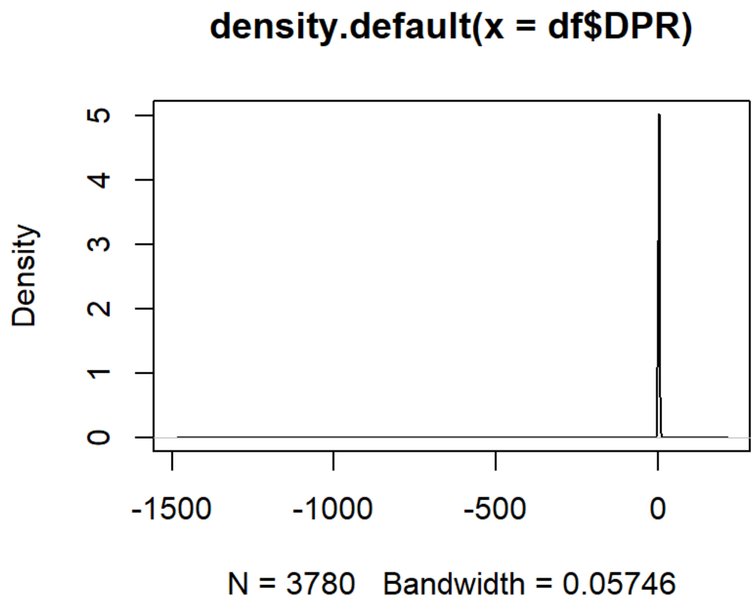

Figure 1. Density Plot

4.6 Re-run the model on a different dataset

Since the DPR in the dataset could be impacted by the existence of outliers, the author filtered data from the interquartile of the dependent variable to re-run the model, using the below codes:

```
#Check for the 1st and 3rd quartile of the DPR
quantile(df$DPR)

# Removing outliers from the dataset
df1 <- df[df$DPR >= 0.8745015 & df$DPR <= 0.8747626,]
summarise(count_DPR=n(), group_by(df1,Year))

#Rerun the model again with df1 - new training and test dataset
data_train1 <- filter(df1, Year < 2018)
data_test1 <- filter(df1, Year > 2017)

rerun <- plm(formula = DPR ~ GDP + Risk + ROA + Leverage + Growth + Size +
Liquidity, data = data_train1, effect = "individual" , index = c("Ticker"), model
= "random")
summary(rerun)
data_test1$rerun <- predict(rerun, newdata = data_test1)
```

The same error metrics as well as residuals plot were used to evaluate the prediction results, this time yielding much better results. However, results taken from such a small sample size (only 113 observations) cannot represent the dividend policy on the HOSE.

```
#Evaluate the rerun model's prediction
mae(data_test1$DPR, data_test1$rerun)
smape(data_test1$DPR, data_test1$rerun)
rmse(data_test1$DPR, data_test1$rerun)

#Residual plot of the new prediction
data_test1$res <- data_test1$DPR - data_test1$rerun
plot(data_test1$rerun, data_test1$res, xlab="Prediction", ylab="Residuals")
```

```
ggplot(data = data_test1, aes(x=rerun, y=res)) +
  geom_point() +
  theme_few() +
  ggtitle("Residuals Plot") +
  xlab("Prediction") +
  ylab("Residuals") +
  theme(plot.title = element_text(hjust = 0.5))
```

Overall, the results and plots obtained helped the author with giving conclusion to this research project. Specifically, machine learning did not predict well the dividend payment policy of listed firms on the Ho Chi Minh City Stock Exchange. Therefore, future research could identify a more effect statistical learning model and a more inclusive dataset. The complete code artefact and dataset can be found in the submitted zip file.