MONASH University

Report on Happiness up to 2022

Zhixiang Yang
EBS Honours Student

Yiqi Wang
Master of BA Student

Xintong You
Master of BA Student

Report for
Group 07 ETC5513

24 May 2022

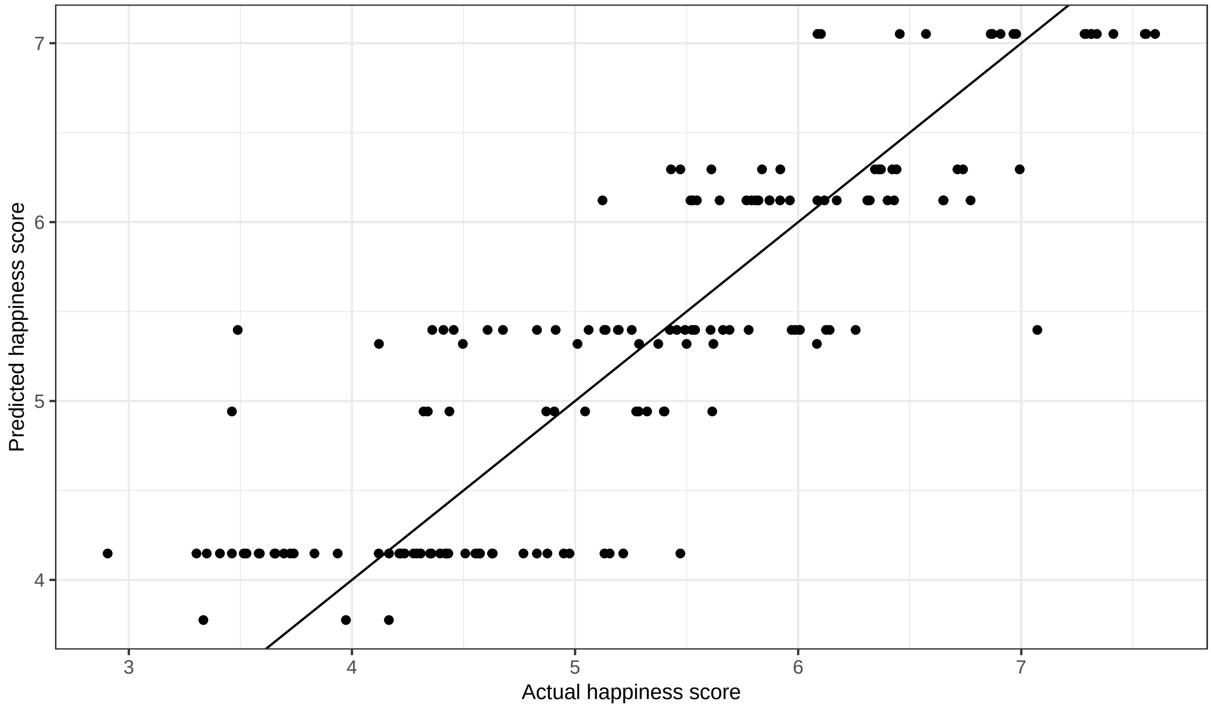
MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

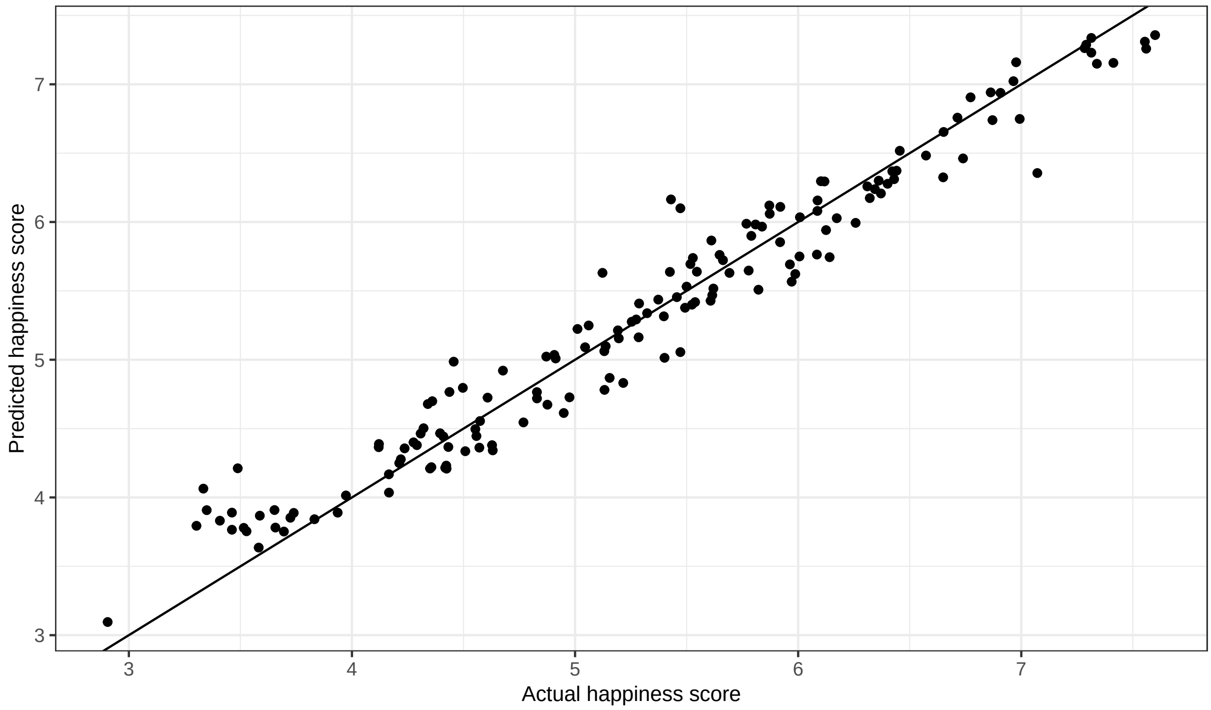
(03) 9905 2478
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

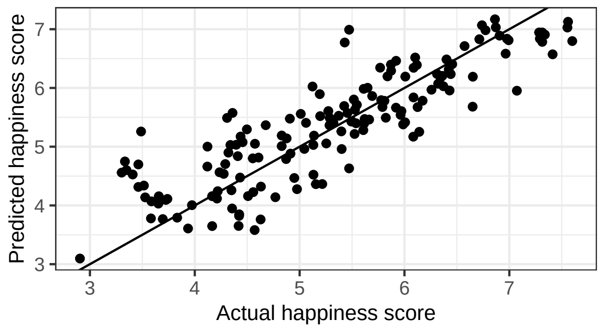
Decision Tree Regression



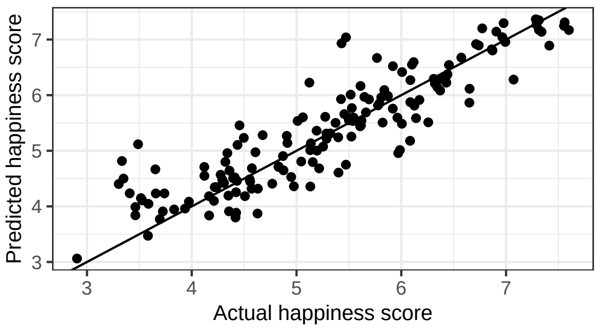
Random Forest Regression



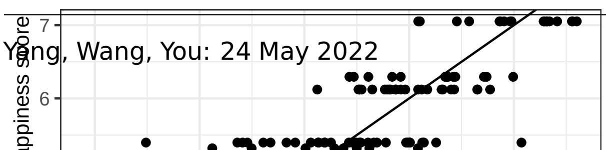
Multiple Linear Regression



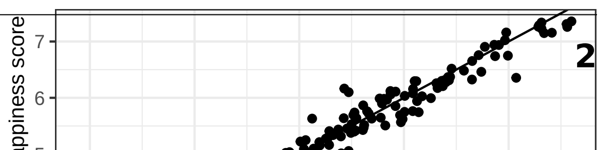
Support Vector Machine



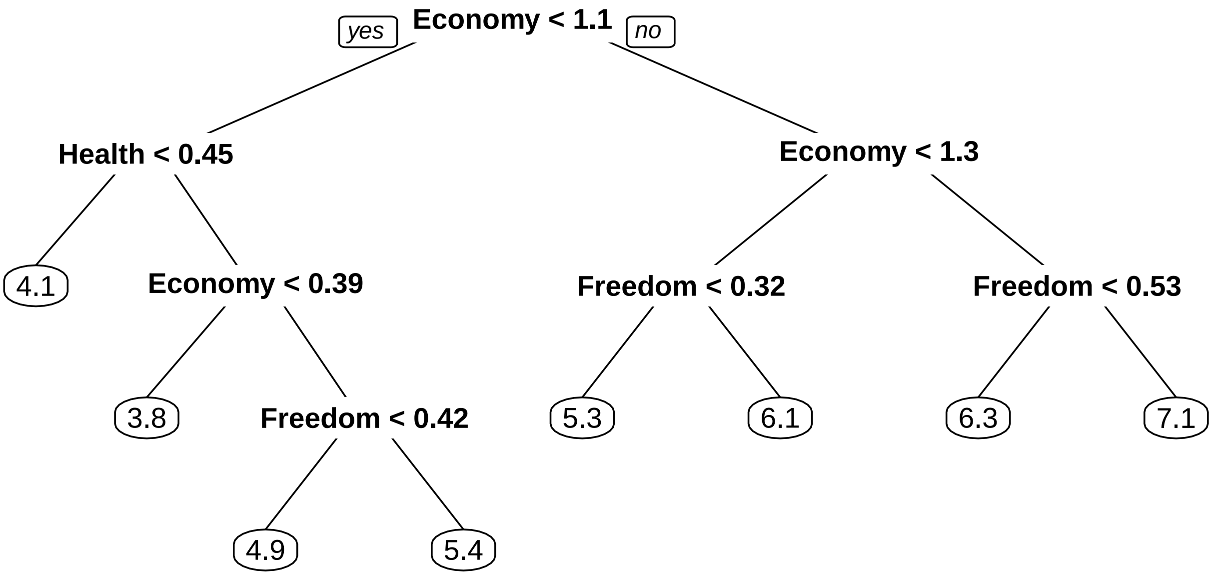
Decision Tree Regression



Random Forest Regression



1 Tree plot



2 Random Forest regression

RF model is used to predict causes for Random Forest is a practical way and regularly used in machine learning models. Random Forest Model have the variable selecting system (via bootstrapping) to decide the most significant tree and can reduce overwriting compared with decision tree.With that said, random forests are a strong modeling technique and much more robust comparing with many different methods. (Liberman, 2017). We can see from the plot that this model have captured the data well in the past few years.

From the report, we can see that our model is for 2020 data.

$$log(score) = -4.6000 - 0.0008\ cpi + 0.2591\ log(economy)$$
$$-0.0026\ log(population) + 0.0114\ log(health) + 0.0032 log(year)$$

We can see that the total proportion of variance explained by the model with these variables are 60.49%. For the 4 predictors, the economy status(GDP per capita) contribute most to the happiness scores than other variables.

3 Factor importance

Here, we concluded that the best method to fit the 2022 data is the Random Forest. This model will also allow us to tell the importance of variables via a factor loading summary table.

In this table, we can conclude that the most important variables on explaining the happiness scores will be the Happiness and Health.

Table 1: *Variable importance for Random Forest model*

	IncNodePurity
Economy	348.66900
Health	327.08463
Freedom	169.32991
Generosity	97.79738

Table 2: *Linear regression model for happiness scores without new data*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.437442	0.0752984	32.370444	0
Health	1.223204	0.1392052	8.787054	0
Economy	1.454316	0.0841520	17.282015	0
Freedom	1.372180	0.1095802	12.522156	0
Generosity	1.170207	0.1396671	8.378545	0

4 We abandon the current dataset to add more variables

```
##
## Call:
## lm(formula = `Happiness Score` ~ Health + Economy + Freedom +
##     Generosity, data = dataall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88683 -0.36864  0.04779  0.39916  1.76462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.43744    0.07530  32.370  < 2e-16 ***
## Health       1.22320    0.13921   8.787  < 2e-16 ***
## Economy      1.45432    0.08415  17.282  < 2e-16 ***
## Freedom      1.37218    0.10958  12.522  < 2e-16 ***
## Generosity   1.17021    0.13967   8.379 2.52e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.582 on 772 degrees of freedom
## Multiple R-squared:  0.732,  Adjusted R-squared:  0.7307
## F-statistic: 527.3 on 4 and 772 DF,  p-value: < 2.2e-16
```

5 What is the correlation between these variables in linear model.

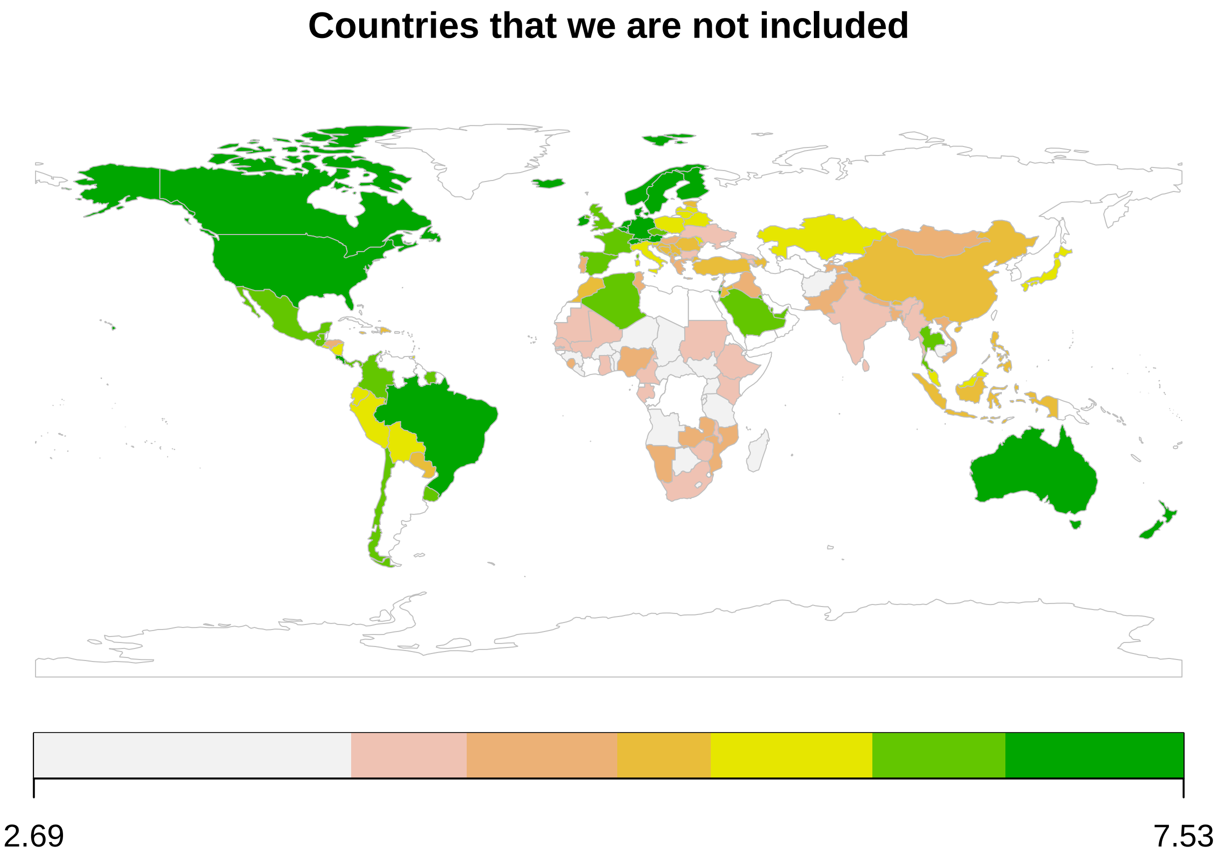
One advantage of multivariate linear regression is that it can allow us to analyse the relationship between different variables in a statistical coherent way. We can start with the correlation between each variables.

In order to better analyse the relationships, I add two new variables, which are CPI values and the population size for each country. However, due to the limitation of the new dataset, we can only conduct our analysis based on the 2020 data.

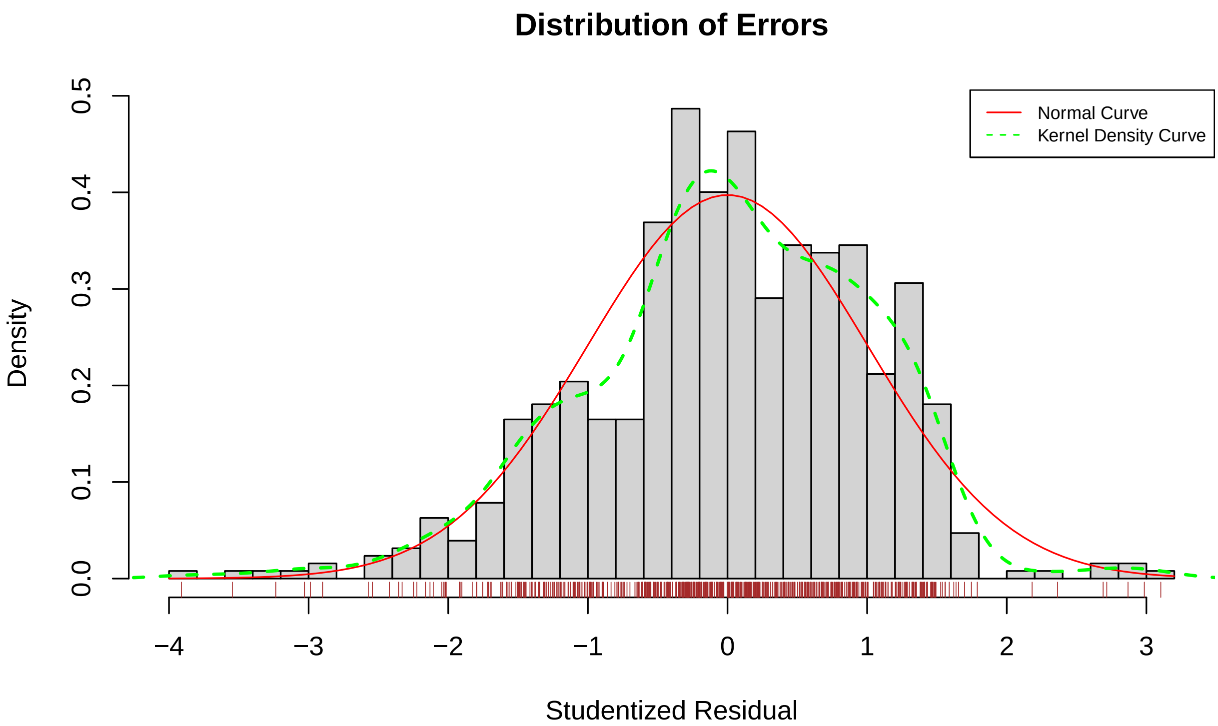
```
##
## Call:
## lm(formula = log_score ~ cpi + log_eco + log_population + log_health +
##     year, data = lognarm_hapall)
##
## Coefficients:
##   (Intercept)            cpi        log_eco  log_population      log_health
##    -4.5998547     -0.0007945      0.2590779      -0.0025806       0.0113946
##          year
##     0.0031566
```

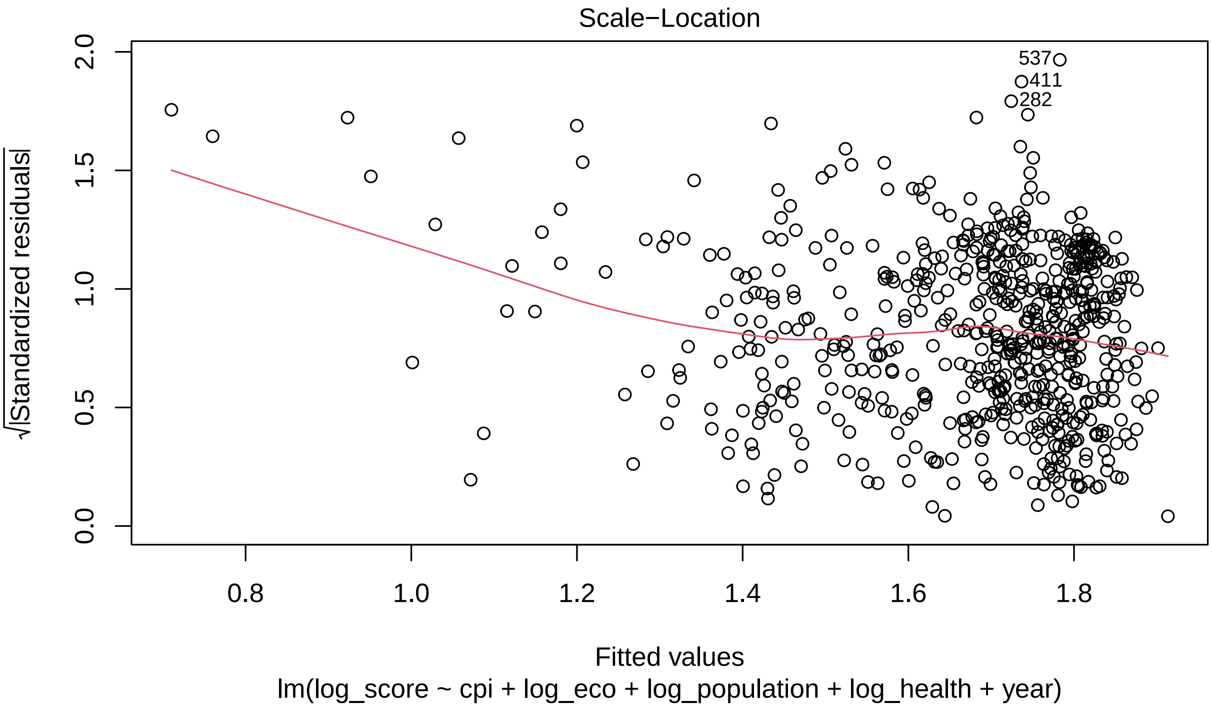
6 The countries that we missed in the dataset.

```
## 641 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
## 109 codes from the map weren't represented in your data
```

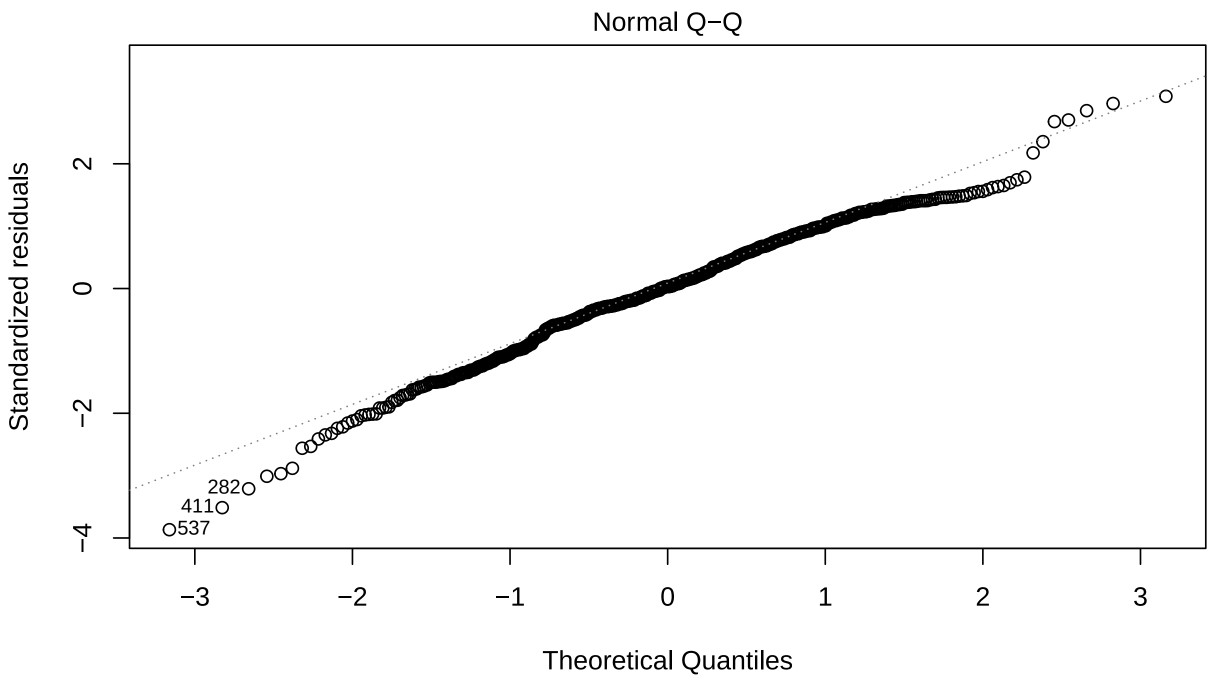


7 Residual Diagonistic

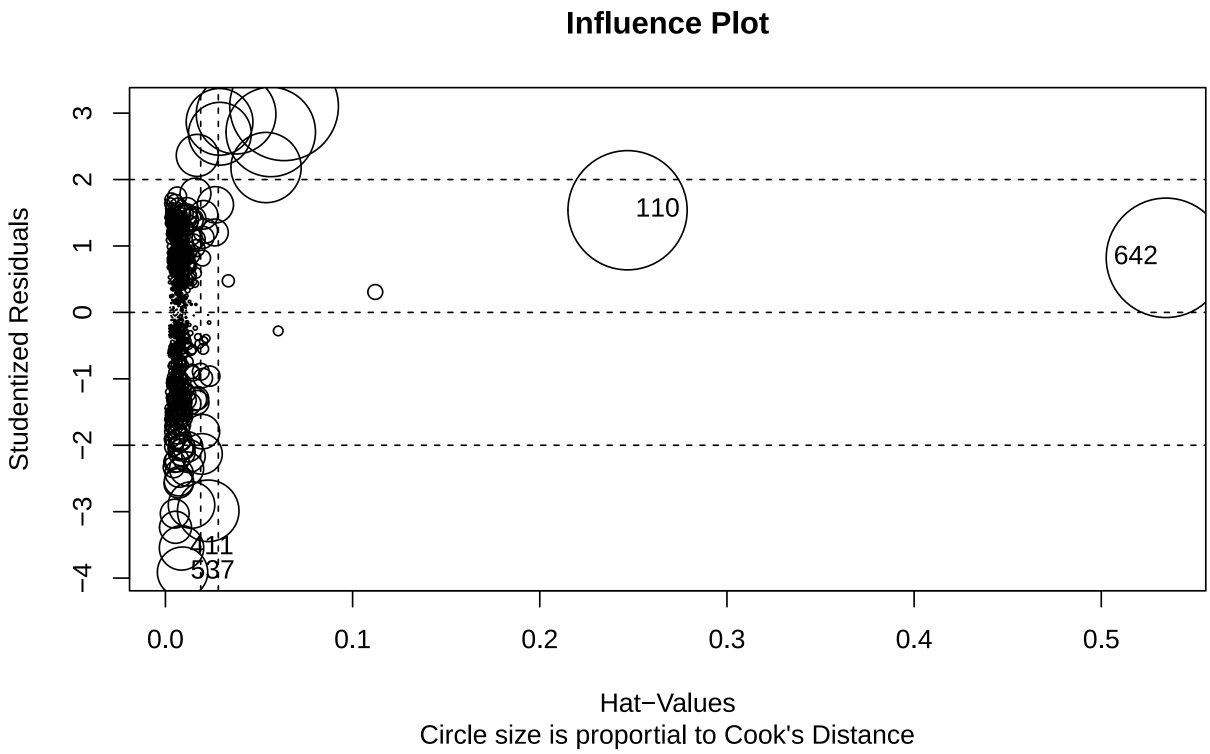




Normal Q−Q plot



In this Figure, the dotted line is Normal Distribution




```
##        StudRes         Hat      CookD
## 110  1.5379757 0.246899853 0.12896655
## 411 -3.5457416 0.008605488 0.01786073
## 537 -3.9107652 0.009146505 0.02300857
## 642  0.8219777 0.534532896 0.12938311
```