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

26 May 2022

MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

<Trained Models >	<Model description >
Multivariate Linear Model	Simple Linear regression with multiple variables
Support Vector Machine Model	Use multiple learning algorithms (resampling and tree) to give us better results.
Decision Tree Model	Binary tree model have control statement.
Random Forest Model	Use multiple learning algorithms (resampling and tree) to give us better results.

Table 1: *Model Description of our Trained Models*

1 Modelling

1.1 What is the most important variable to explain the happiness score differences across different countries in different years?

In this part, we only keep the common variables (Economy, Health, Generousity and Freedom) for these years to conduct our analysis. To explore the factors that could be contributing to the happiness score differences between each year, we first test few common models listed in Table1:

After trained our model based on the all historical data (from 2015 to 2022). We can see from the Figure 1 that the best model to fit the data is Random Forest model while the Multivariate and SVM performed similarly. The decision tree performed the worst because it changed the structure of the data. The result is similar when we have different training split ratios (see).

Table 2: *Variable importance for Random Forest model*

	IncNodePurity
Economy	348.6690
Health	327.0846
Freedom	169.3299
Generosity	97.7974

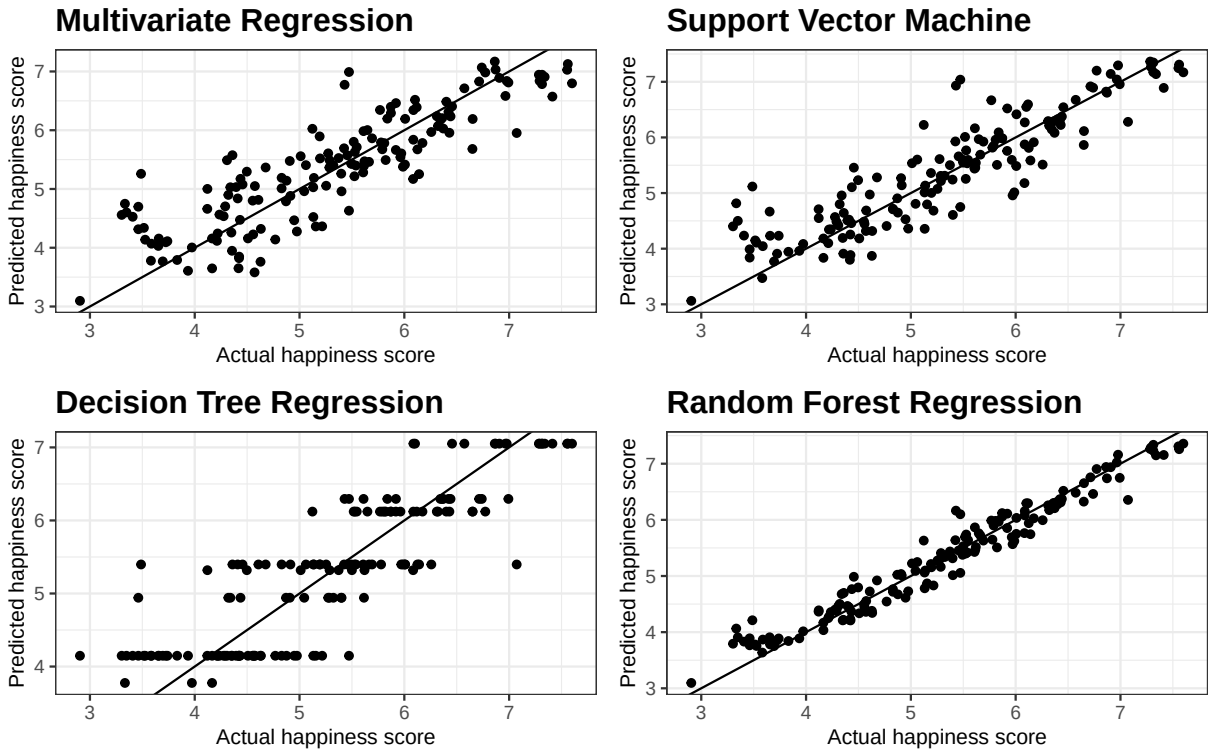

Figure 1: *Model Training Split Ratio is 0.8*

1.1.1 Random Forest regression

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 for various training set (see).

To get which are the most important variables we then check their loadings in RF model (see Table 2).

In the Table 2, we conclude that the most important variables on explaining the happiness scores will be the Happiness and Health due to the loading for them are significant higher than others.

Table 3: *Linear regression model for happiness scores without new data*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4374	0.0753	32.3704	0
Health	1.2232	0.1392	8.7871	0
Economy	1.4543	0.0842	17.2820	0
Freedom	1.3722	0.1096	12.5222	0
Generosity	1.1702	0.1397	8.3785	0

Table 4: *Multivariate Linear regression model for the log data with R-squared is 0.6076*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.5999	11.5632	-0.3978	0.6909
cpi	-0.0008	0.0002	-4.2994	0.0000
log_eco	0.2591	0.0094	27.5487	0.0000
log_population	-0.0026	0.0036	-0.7072	0.4797
log_health	0.0114	0.0042	2.7419	0.0063
year	0.0032	0.0057	0.5508	0.5820

1.2 Multivariate Linear Model Analysis.

One advantage of multivariate linear regression is that it can allow us to analyse the relationship between different variables in a statistical coherent way. For example, marginal effects and percentage changes.

In our classic linear model, which is

Classic Multivariate Linear Model :

$$Happiness\ score = Economy + Health + Generosity + Freedom$$

We can see from Table 3, in our classic linear model, they have similar loadings and all of them are significant, which we cannot tease out the important variables out of this model.

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 data up to 2020. Moreover, due to the endogenous bias, the t-test here is biased so we cannot reject variables based on the p-value in this case below.

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

We can see in Table 4 that the total proportion of variance explained by the model with these variables are 60.76%. For the 4 predictors, the economy status(GDP per capita) contribute most to the happiness scores than other variables.

In econometrics contexts, the intercept can be interpreted as

2 Endogenity and Sample Selection Bias .

We only included 5 variables and total R^2 is around 60.49% in our model. Then, considering the happiness score can be affected by many prospects. There should have some latent variables that do affect the happiness score for example the education level and culture backgrounds. In other words, our model have an endogeneity issue, which we can solve it by either adding more related variables or use the two-stage least square model.

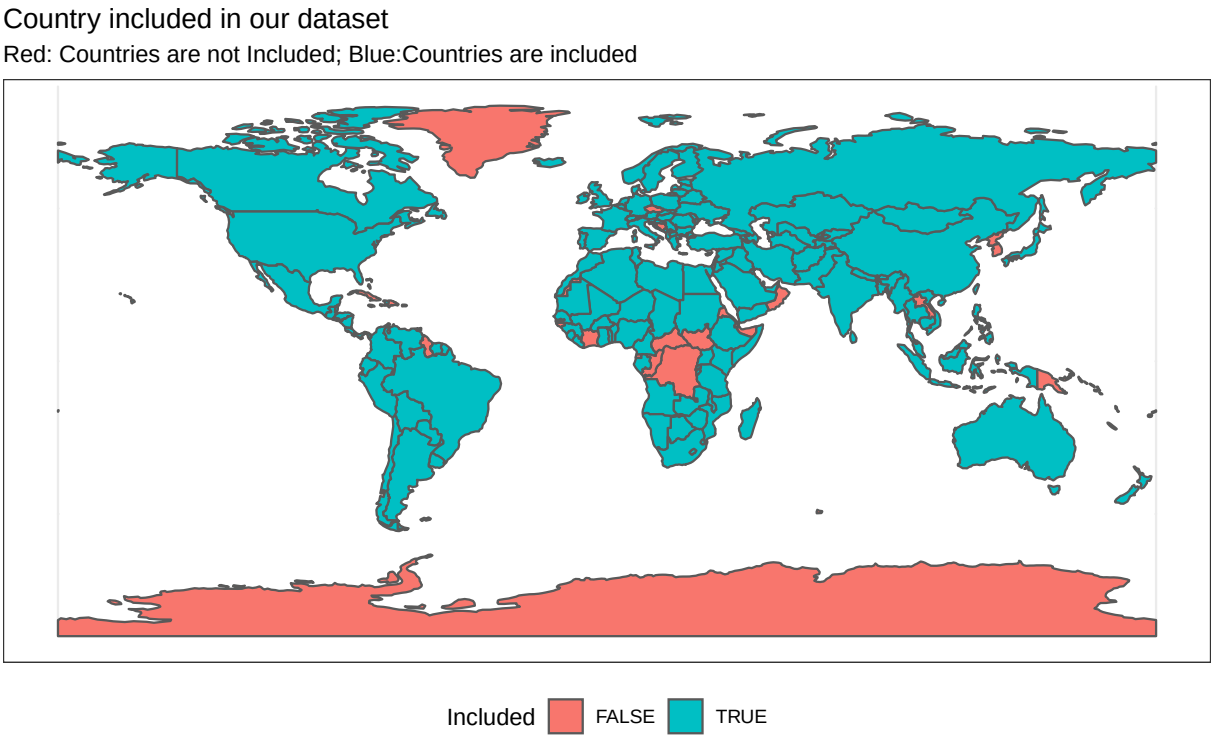

Figure 2: *Colour the country that have included in Worldmap*

Apparently, our model have some bias in selecting our sample. In our report, there are 34 are not included, which have been colored in red from Figure 2. In our sample, we can see countries are not included either from sparsely populated area or developing countries, which indicates that our sample is not random enough and exists a sample selection bias. To solve this, sample selection models such as Heckman model or Tobit can be considered in Econometrics areas.

3 Potential Problems

The diagnostic of our regression model is shown in Figure 4. From the Residual and Fitted plot, we can see it has non-constant variance across the fitted value, which means the presence of Hetroskedasticity. Moreover,both the error distribution (see Figure 3) and the Q-Q plot Figure (see Figure 5 also suggest the density of our model is close to a normal distribution but seems there are some influences by outliers. In the

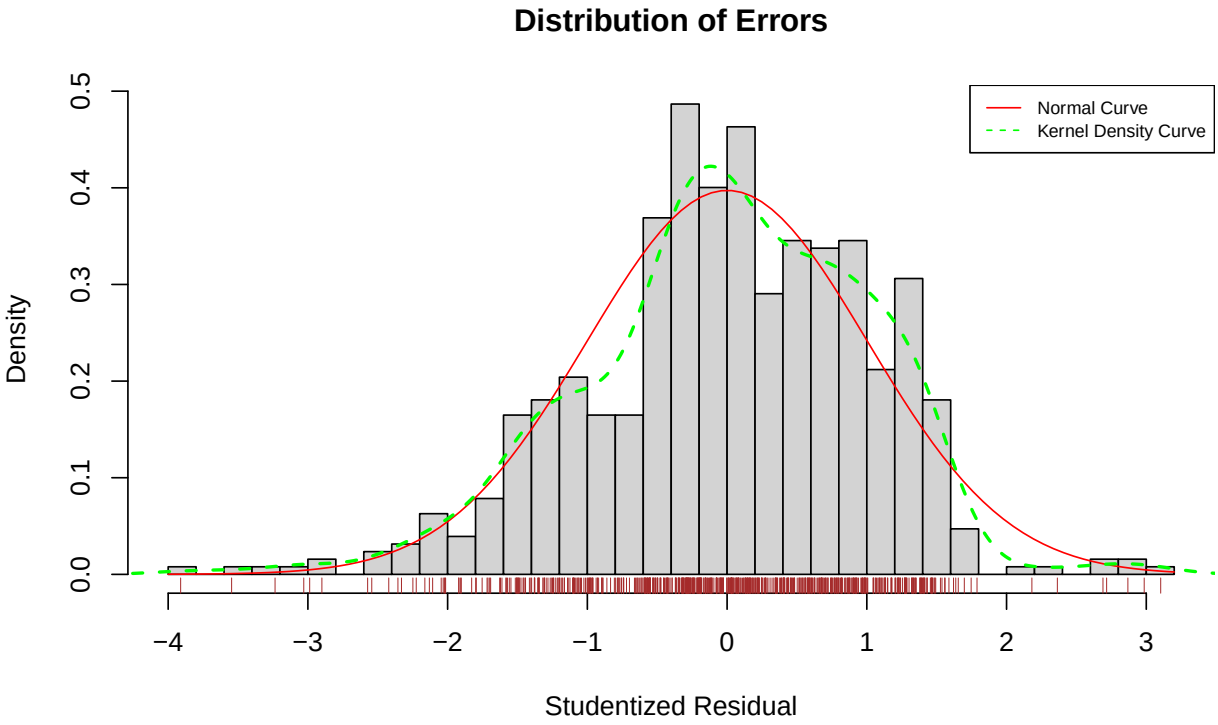

Figure 3: *Distribution of the residual term*

Figure 4: *Residual vs Fitted value plot (with standardised residuals)*

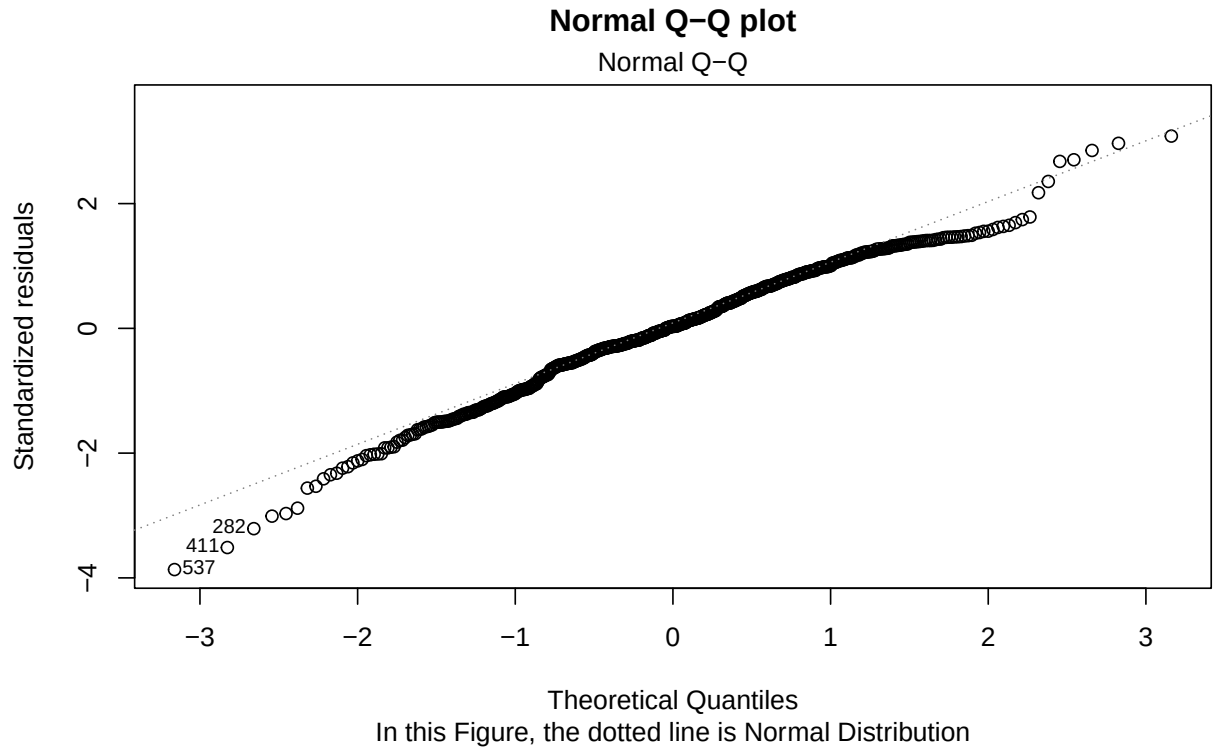

Figure 5: *Q-Q plot fitted model residual density vs normal distribution residual density*

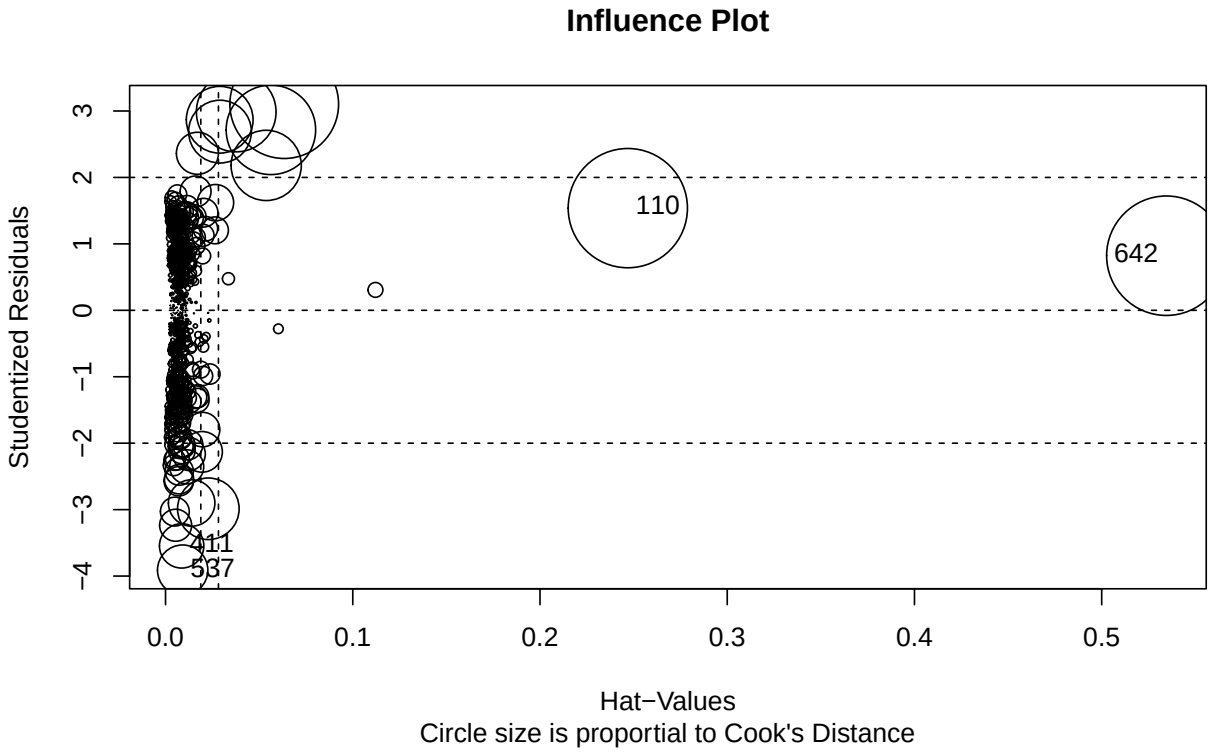

Figure 6: *Influence Plot, the bigger circle means outliers*

4 Concclusion

Appendix

Liberman, N (2017). Decision trees and random forests. *Towards Data Science*.