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

27 May 2022

MONASH
BUSINESS
SCHOOL

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

Contents

1 Modelling

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

In this part, since there are many different variables across each year in our data, we only keep the common variables (Economy, Health, Generosity and Freedom) for these years to conduct our analysis. Before exploring how each factor will contribute to the happiness score, we first need to select a model that can represents our data well by running several tests on a few common models, listed in Table1:

<Possible 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 Possible Models*

We have divided the historical data (from 2015 to 2022) into two separated sets, a training test and a test set with different split ratio (*see* Figure 1; Figure 7 to 10 *in appendix*). The test set will be used to examine which model has the best goodness-of-fit after building up the model with the training set.

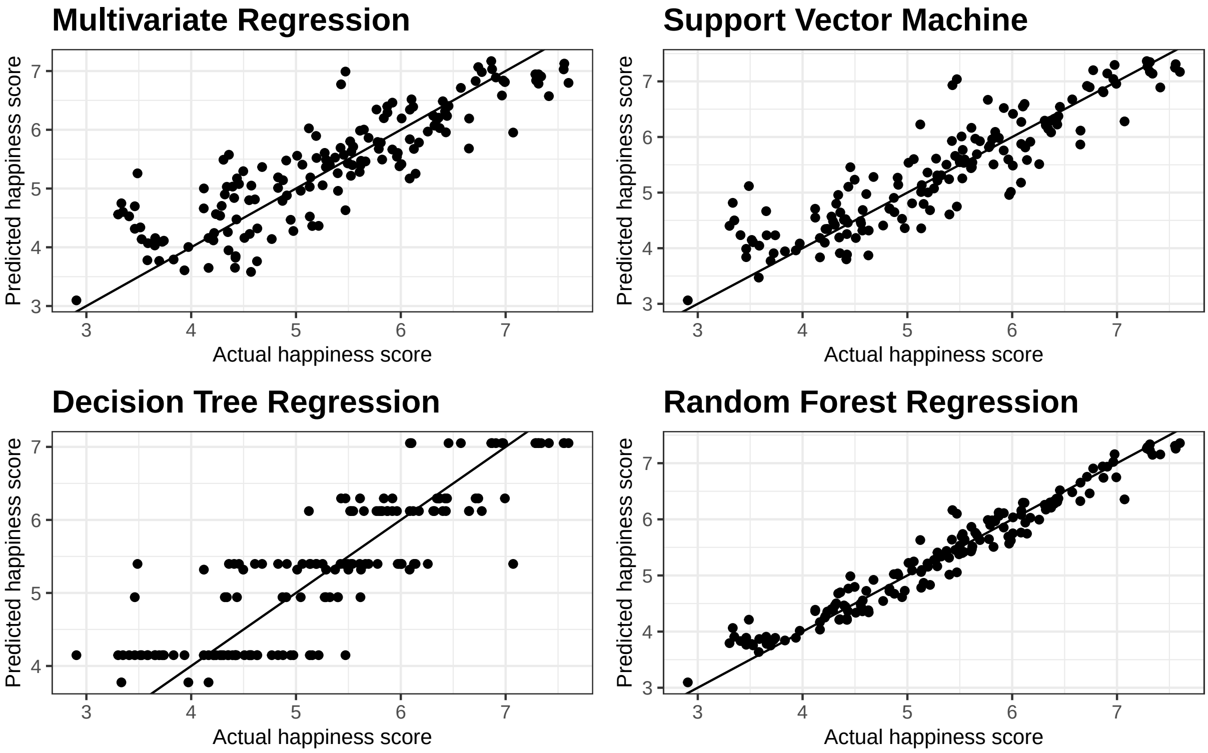

Figure 1: *Model Training Split Ratio is 0.8*

1.1.1 Random Forest regression

Random Forest Model has a variable selecting system (via bootstrapping) to decide the most significant tree and is able to reduce overwriting compared to the decision tree. With that said, the random forest is a strong modeling technique and more robust than other methods (Liberman, 2017). We can see from the plot that this model has captured the data very well in the past few years based on various training sets (see Figure 7 to 10).

By checking the loading for each variable in the RF model, we get the order of importance in Table 2.

Table 2: *Variable importance for Random Forest model*

	IncNodePurity
Economy	348.6690
Health	327.0846
Freedom	169.3299
Generosity	97.7974

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

Based on the result in Table 2, the most important variable is Economy, followed by Health, Freedom and Generosity being the last.

1.2 Multivariate Linear Model Analysis.

The RF model helps us to determine which variables to choose in the multivariate linear regression model. The advantage of using multivariate linear regression is that it can allow us to analyse the relationship between different variables in a statistical coherent way (Voxco, 2022),such as, marginal effects and percentage changes.

If using a classic linear model with all four variables,

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

They have similar impacts on explaining the happiness score and all of them are significant in Table 3. Thus, it is hard to tease the importance of each variable out with this model.

As a result, we will drop the least insignificant two and add another two new variables, Consumer Price Index and the population size of each country, gained from the World Bank data from 2016 to 2022 (World Bank, 2022). After matching the data for each country and drop the NA values, we can use the new dataset to construct a new Linear model. Nevertheless, due to the lack of data in 2021 and 2022, we can only use the data up to 2020.

New Multivariate Linear Model with a natural log of score :

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

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

We can see in Table 4 that 60.76% of the variation in $log(score)$ can be explained by the model. It is noticeable that the economy status(GDP per capita) has the largest influence on the happiness scores than other variables.

Besides, in econometric contexts, coefficients in the log-log model can be interpreted as the percentage change in dependent variable when there is a one percentage change increase in the regressor (Benoit, 2011).

For example:

$$\begin{aligned}\frac{\Delta\ \text{log(Happiness Score)}}{\Delta\ \text{log(Economy)}} &\approx \frac{\Delta\ \text{Happiness Score}}{\Delta\ \text{Economy}}\frac{\text{Economy}}{\text{Happiness Score}}\\ &= \frac{\%\Delta\ \text{Happiness Score}}{\%\Delta\ \text{Economy}}\\ &= \text{ME(Coefficient)}\\ &= 0.2591\end{aligned}$$

Here, we can interpret that a 1% increase in Economy(measured in GDP per capita) will increase the happiness score by 0.2591%, keeping all other regressors constant. Similarly for the rest variables, a 1% percent increase in health will increase happiness by 0.0114%. On the contrary, we can see that an increase in population and CPI have negative impacts, which are -0.0026% for Population and -0.08 for CPI respectively.

2 Endogenity and Sample Selection Bias

This model may have an endogenous problem, which is caused by omitting variables, as we only include 5 relevant variables in the model. There are some latent variables that affect the happiness score but are not included in this model such as education level and culture backgrounds.

Apparently, our model also has some bias in sample selection. In our model, data for 34 countries are not included, which are colored in red in Figure 2. These missing countries are either sparsely

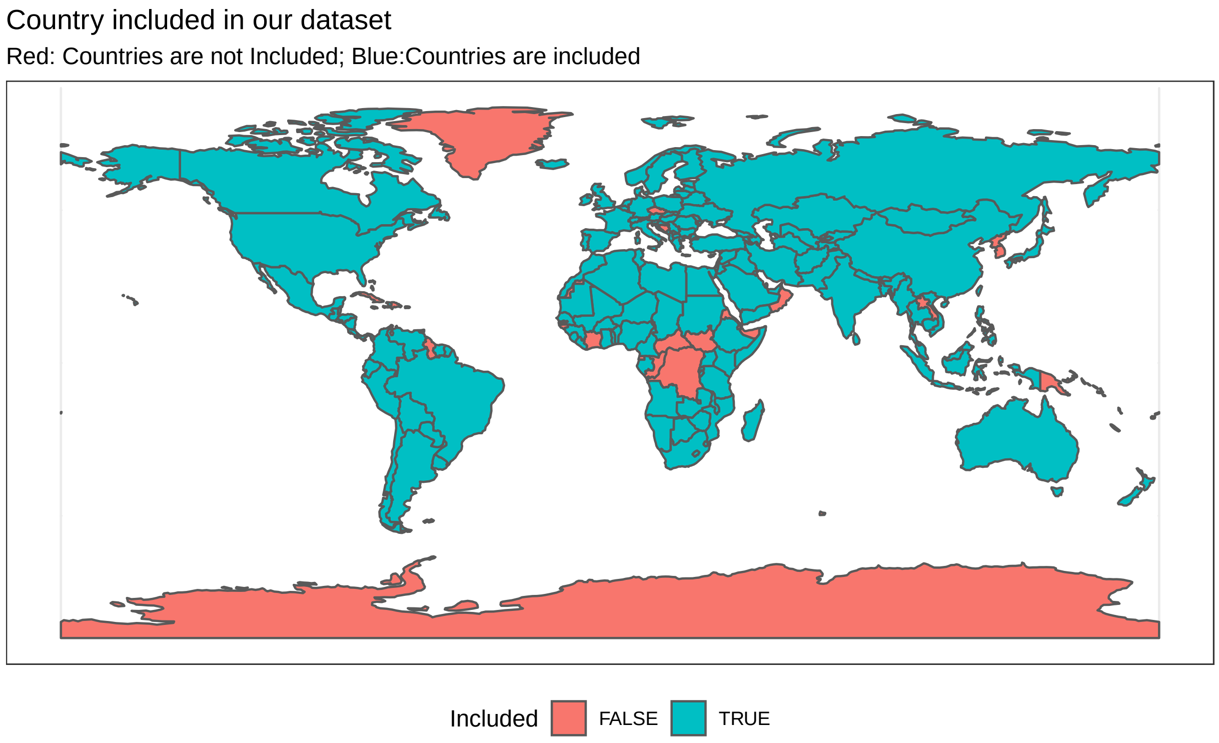
Country included in our dataset
Red: Countries are not Included; Blue:Countries are included
Included
FALSE
TRUE

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

populated or still in development, so the model is not revealing all information due to the sample selection bias. To solve this, sample selection models such as Heckman model or Tobit can be considered in Econometrics areas.

3 Residual Diagonistic for the Regression Model

From the Residual and Fitted plot in Figure 4, we can see it has a non-constant variance across the fitted value, which indicates the presence of Heteroskedasticity. Moreover, the error distribution (see Figure 3) and the Q-Q plot Figure (see both Figure 5 and Figure 6 also suggest the density of our model is somewhat to a normal distribution but influenced by outliers.

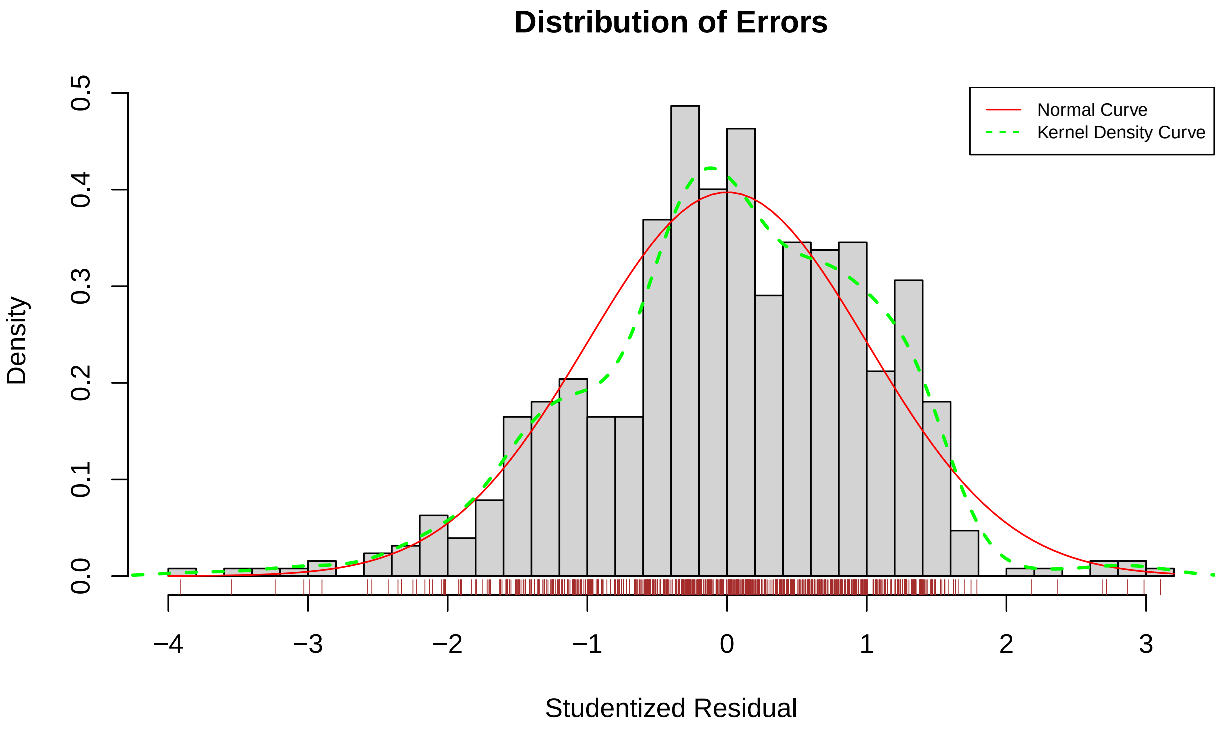

Figure 3: *Distribution of the residual term*

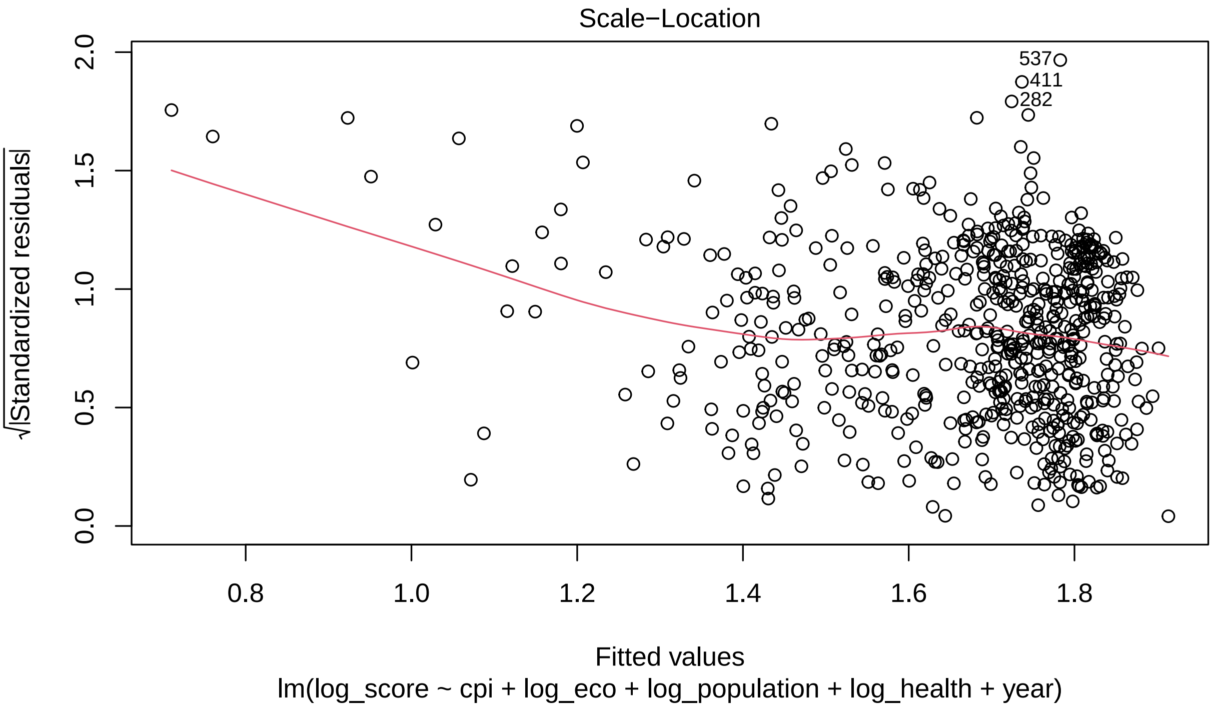

Figure 4: *Residual vs Fitted value plot (with standardised residuals)*

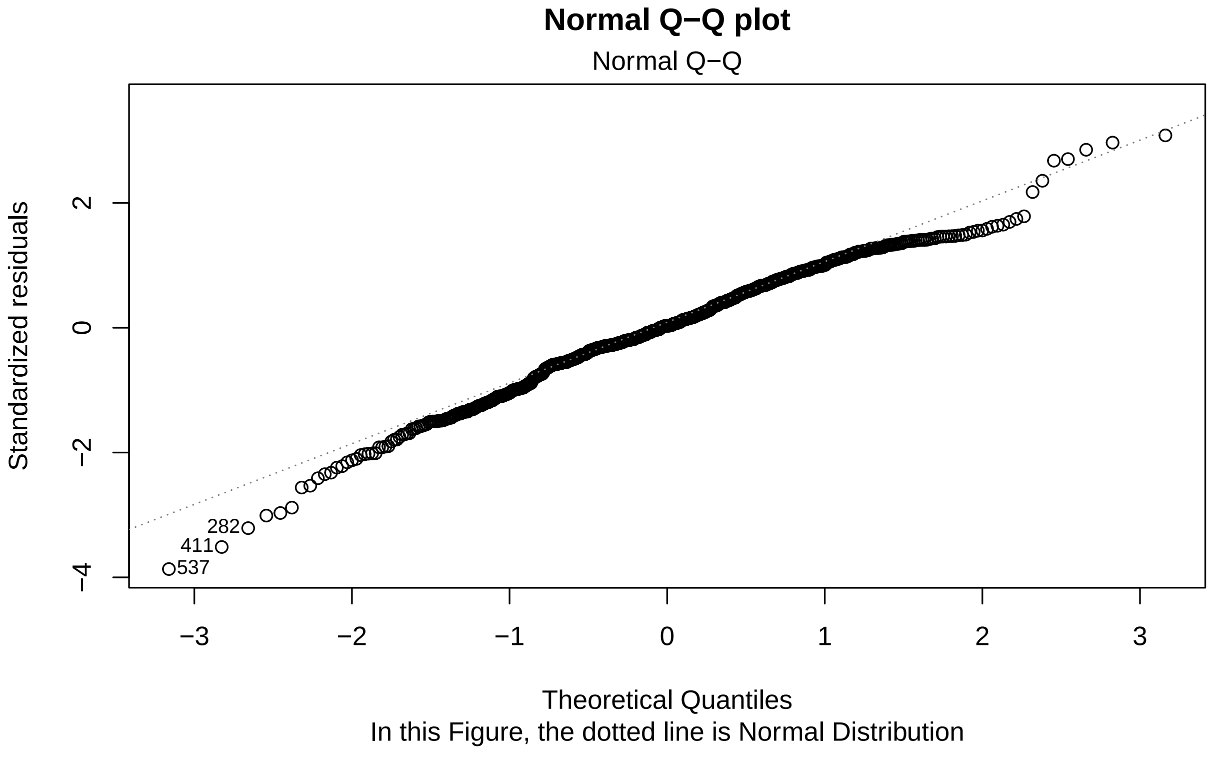

Figure 5: *Q-Q plot fitted model residual density vs normal distribution residual density*

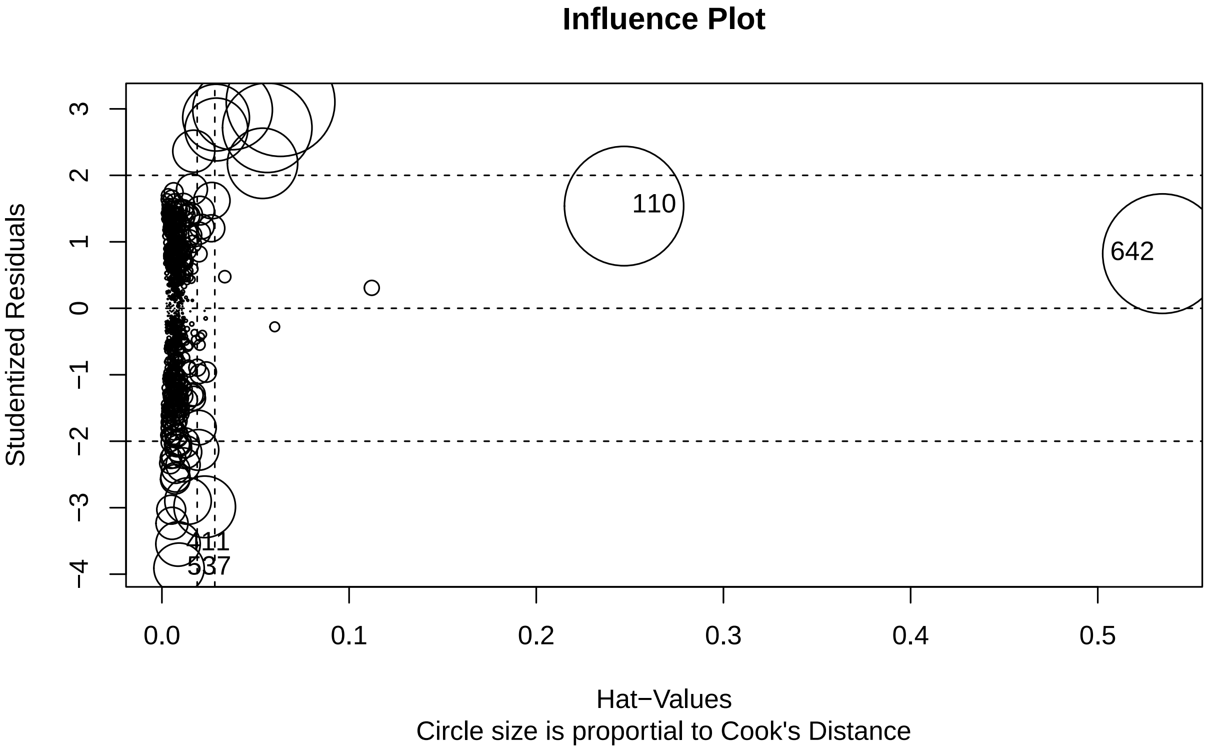

Figure 6: *Influence Plot, the bigger circle means outliers*

4 Conclusion

In conclusion, our model did a good job in explaining the relationships between happiness score and our selected aggressors even though there are still some limitations.

5 Appendix

We here take the Economy Situation (measured in GDP per capita) as an example:

$$\frac{\partial \log(\text{Happiness Score})}{\partial \log(\text{Economy})} = \text{ME(Coefficient)}$$

From (1) we can say, when other variables remain constant

$$\Delta \ \log(\text{Happiness Score}) = \text{ME(Coefficient)} \times \Delta \ \log(\text{Economy})$$

By using the infinite approaching approximation of log function, we can know

$$\begin{aligned} log(x_0 + \Delta x) - log(x_0) &\approx log(x_0) + log'(x_0)\Delta x - log(x_0) \\ &= \frac{\Delta x}{x_0} = \text{percentage change in x} \end{aligned}$$

Therefore

$$\begin{aligned} \frac{\Delta \ \log(\text{Happiness Score})}{\Delta \ \log(\text{Economy})} &\approx \frac{\Delta \text{ Happiness Score}}{\Delta \text{ Economy}} \frac{\text{Economy}}{\text{Happiness Score}} \\ &= \frac{\%\Delta \text{ Happiness Score}}{\%\Delta \text{ Economy}} \\ &= \text{ME(Coefficient)} \end{aligned}$$

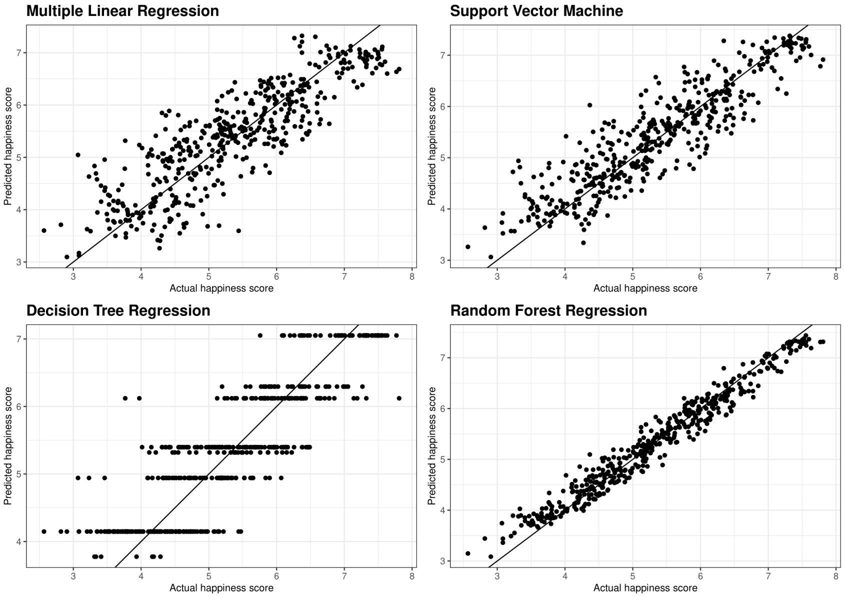

Figure 7: *Models' performance under the training test split ratio = 0.4*

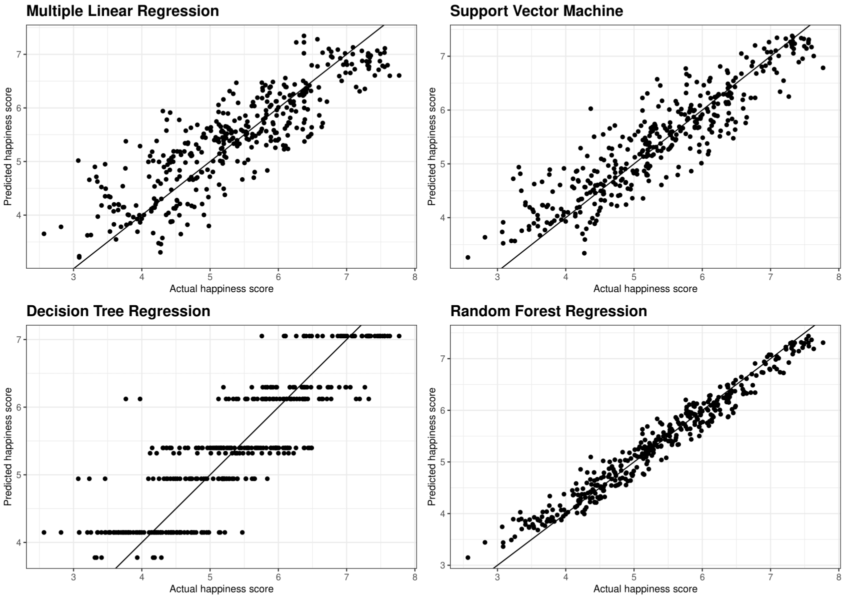

Figure 8: *Models' performance under the training test split ratio = 0.5*

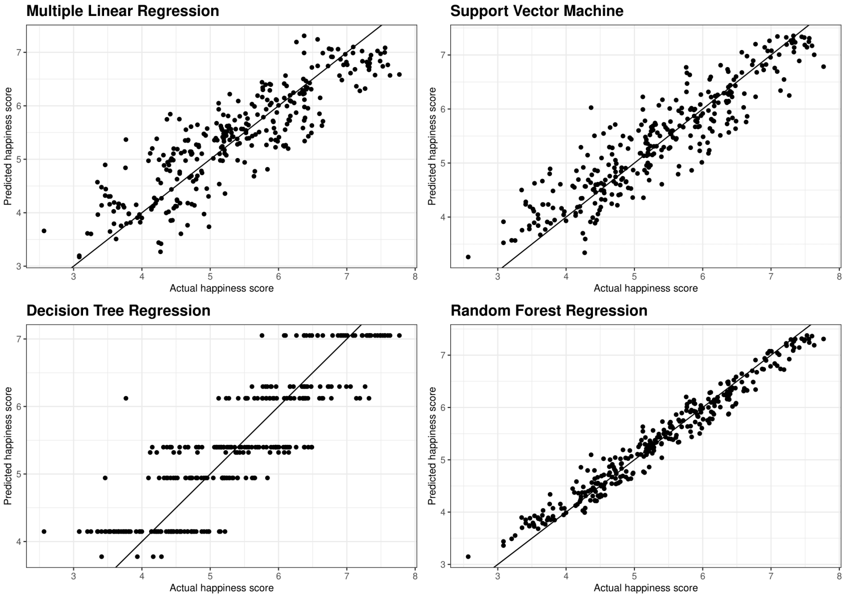

Figure 9: *Models' performance under the training test split ratio = 0.6*

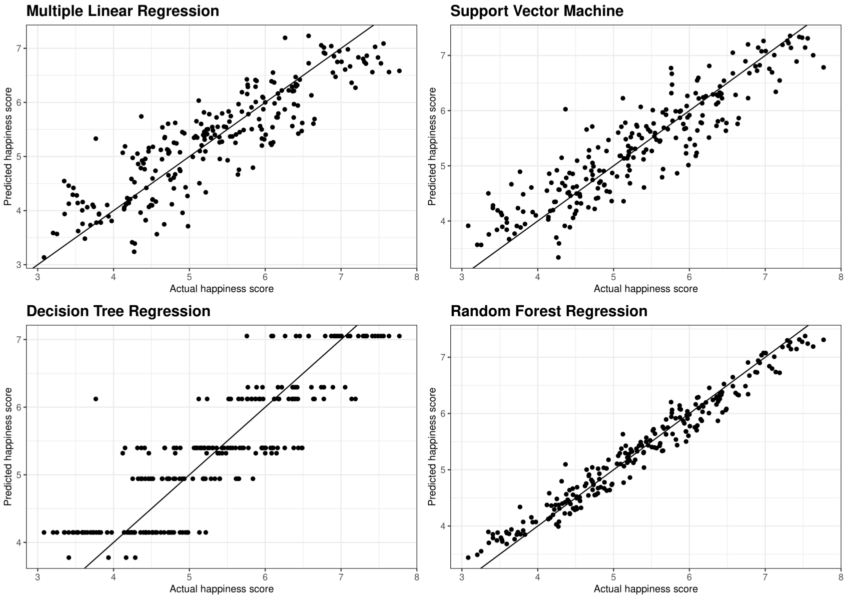

Figure 10: *Models' performance under the training test split ratio = 0.7*

References

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London*, **22**(1), 23–36.

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

Voxco. (2022). *Multivariate regression: Definition, example and steps*.

World Bank. (2022). *World bank data*.