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

MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

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

ABN: 12 377 614 012

Report for
Group 07 ETC5513

27 May 2022

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

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 serval 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. 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 (see Figure 1; Figure 7 to 10 in appendix).

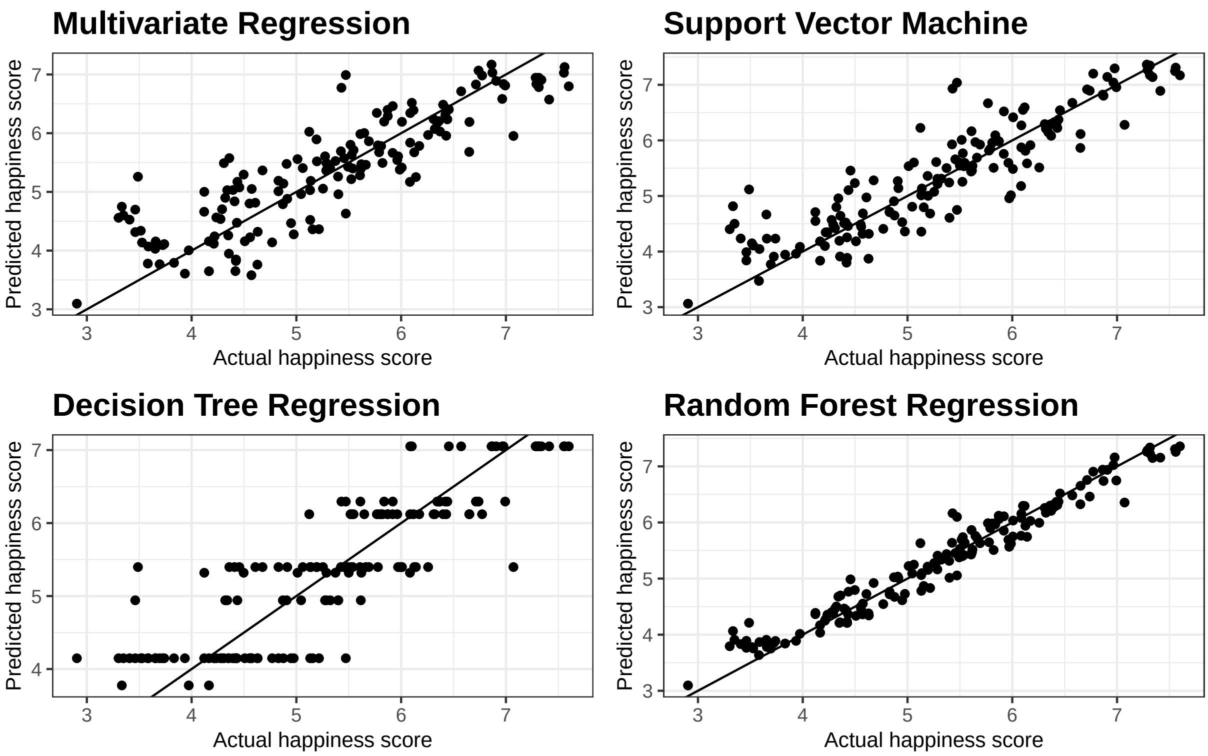

Figure 1: *Model Training Split Ratio is 0.8*

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 Figure 7 to 10). .

To get which are the most important variables, we then check their loadings in RF model (see 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

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.

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 Voxco (2022). For example, marginal effects and percentage changes.

In our classic linear model, which is

$$\textbf{Classic Multivariate Linear Model}:$$

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

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, we join another two new variables, which are CPI values and the population size for each country from The World Bank data (World Bank (2022)) between 2016 to 2022. After matching the data for each country and drop the NA values. We can use the new dataset to construct our new Linear model. Nevertheless, due to the limitation of the new dataset, we can only conduct our analysis based on the data up to 2020.

$$\textbf{Logarithm Multivariate Linear Model with added variables}:$$

$$log(score) = -4.6000 - 0.0008\ cpi + 0.2591\ log(economy)$$

$$-0.0026\ log(population) + 0.0114\ log(health) + 0.0032 log(year)$$

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 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.

Note that in econometric contexts, the intercept in log-linear model can be interpreted as the influence of one percentage change of a variable to dependent variable (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 lead to a 0.2591% rise in the happiness score. Similarly for the rest variables, a 1% percent increase in health will increase happiness for 0.0114%. On the contrary, we can see that a rise in population and CPI have a negative impact with -0.0026% and -0.0008 respectively.

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.

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

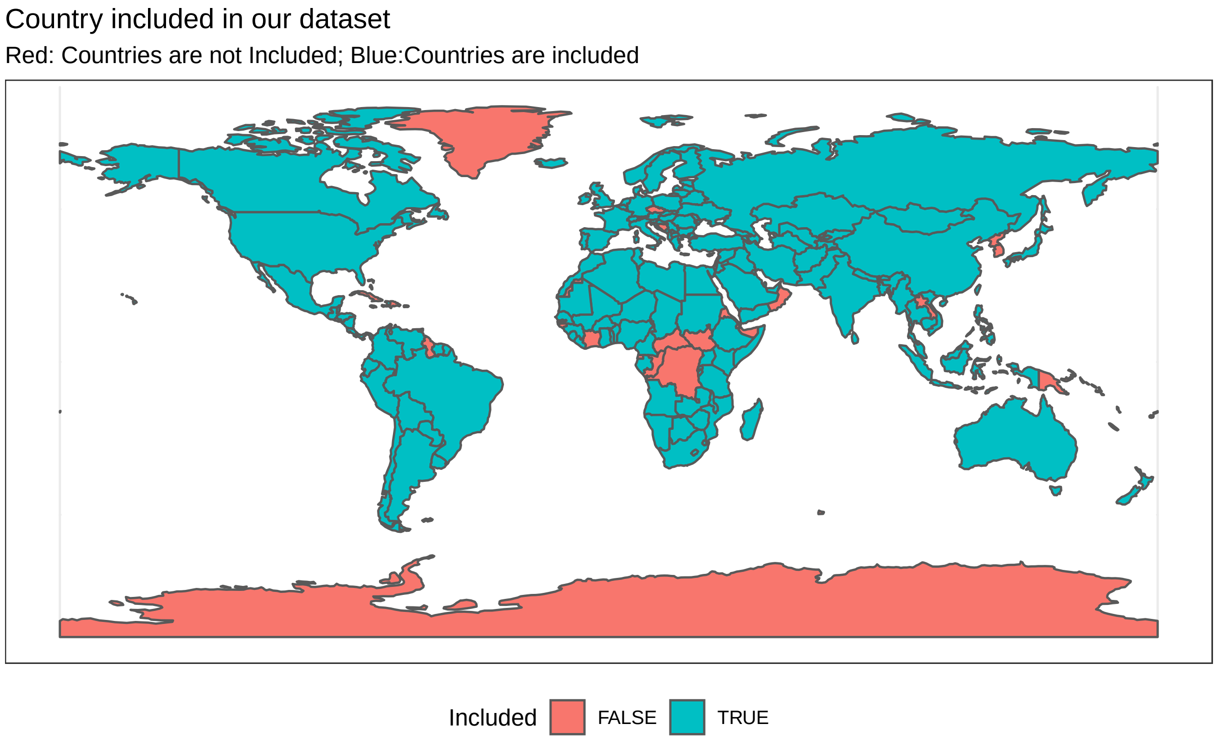

Figure 2: *Colour the country that have included in Worldmap*

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 Residual Diagonistic of Regression Model

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 both Figure 5 and Figure 6 also suggest the density of our model is close to a normal distribution but seems there are some influences by outliers.

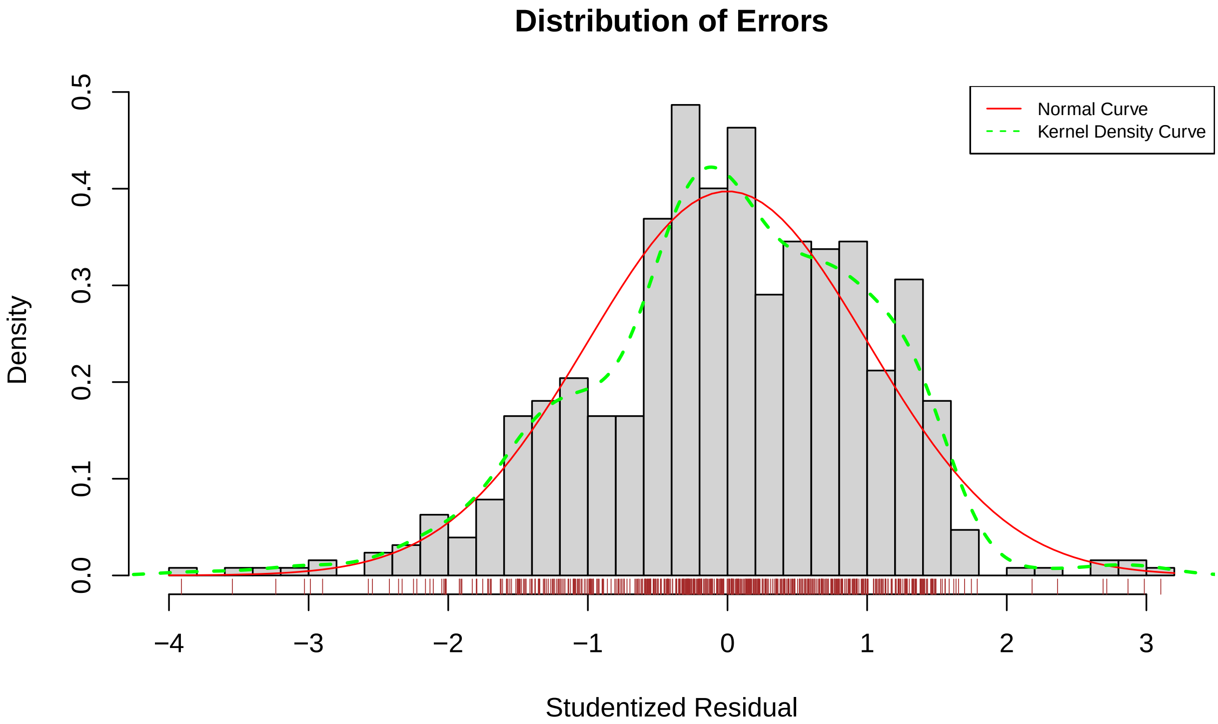

Figure 3: *Distribution of the residual term*

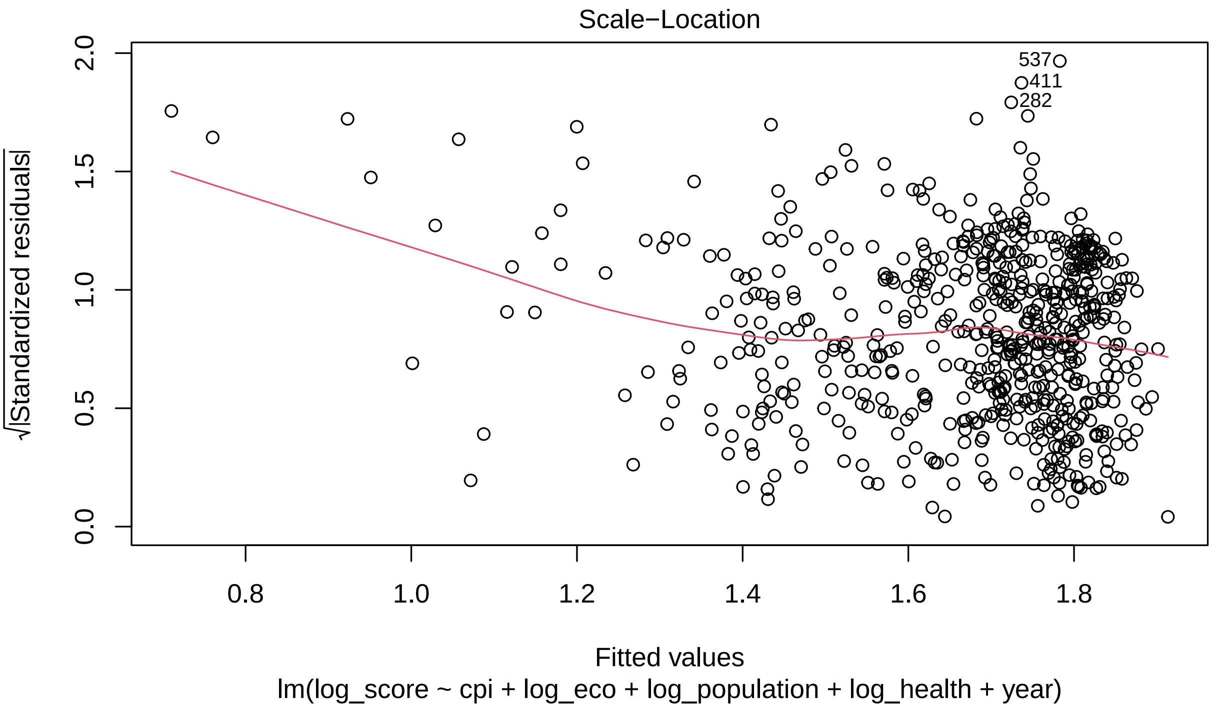

Figure 4: *Residual vs Fitted value plot (with standardised residuals)*

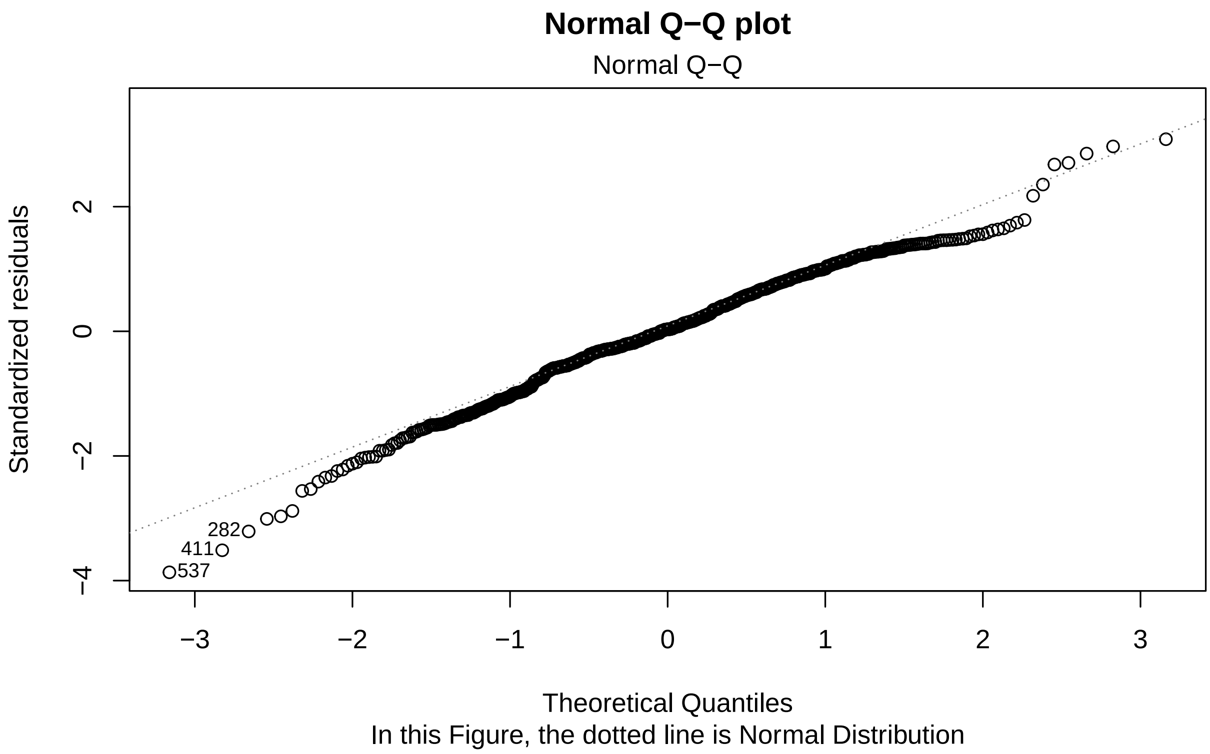

Figure 5: *Q-Q plot fitted model residual density vs normal distribution residual density*

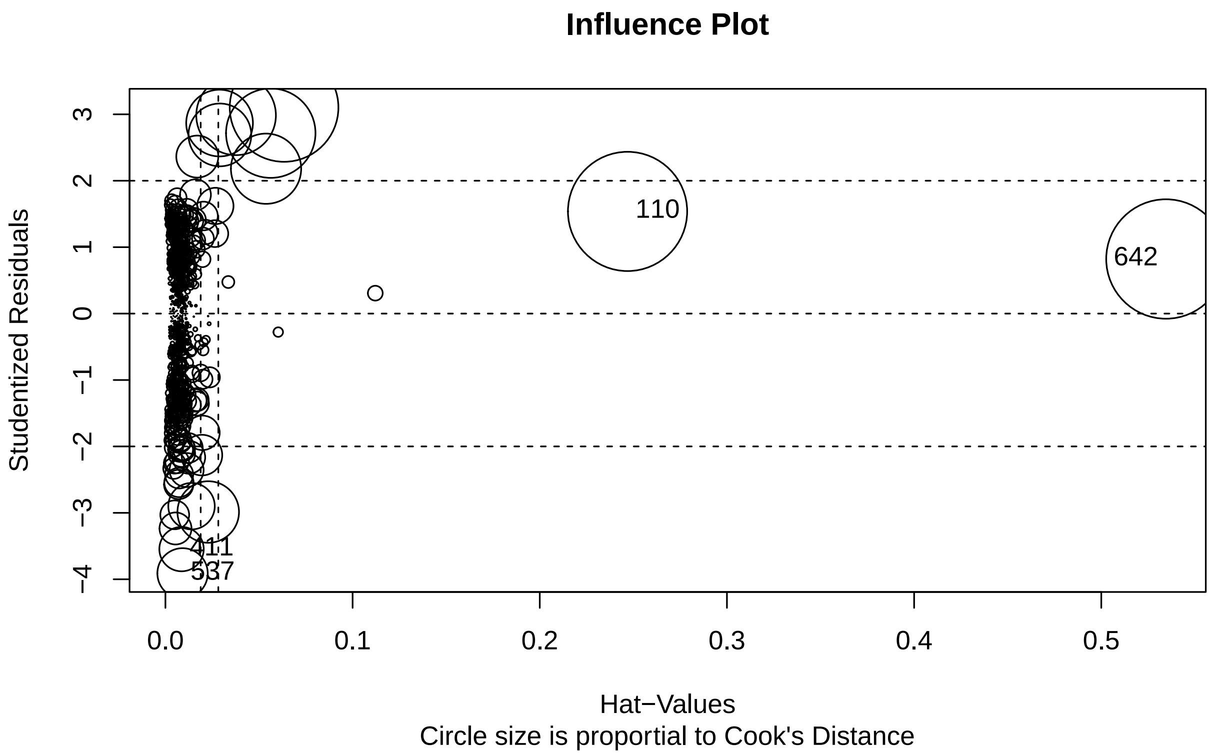

Figure 6: *Influence Plot, the bigger circle means outliers*

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

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

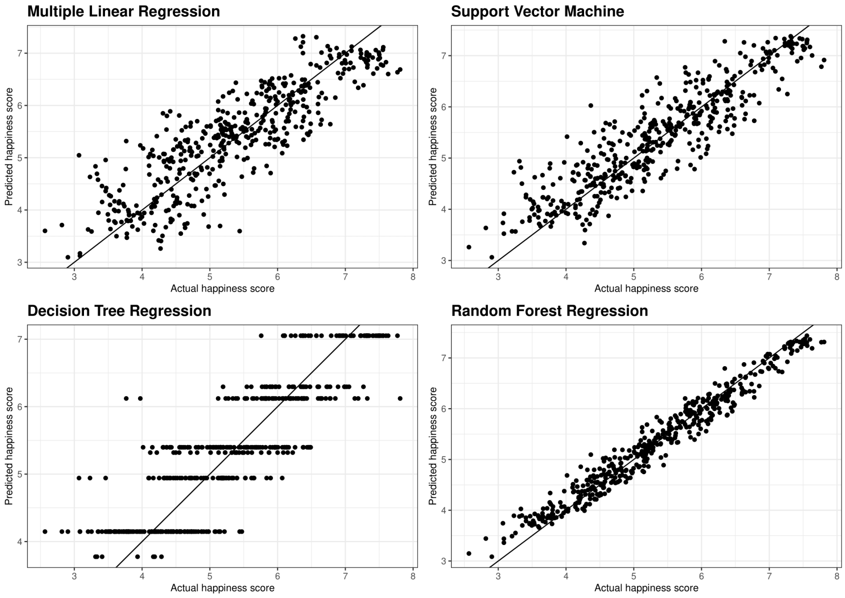

Figure 7: *Models' performance under the training test split ratio = 0.4*

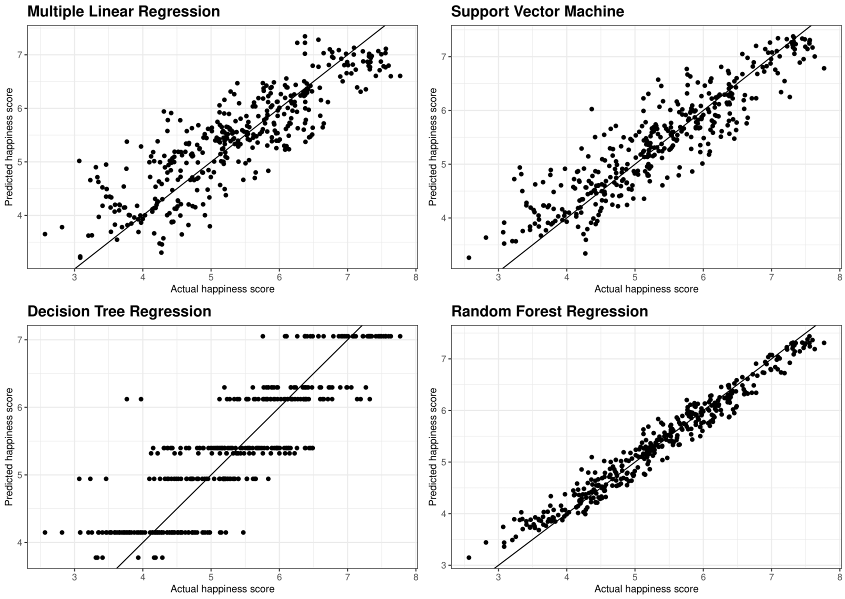

Figure 8: *Models' performance under the training test split ratio = 0.5*

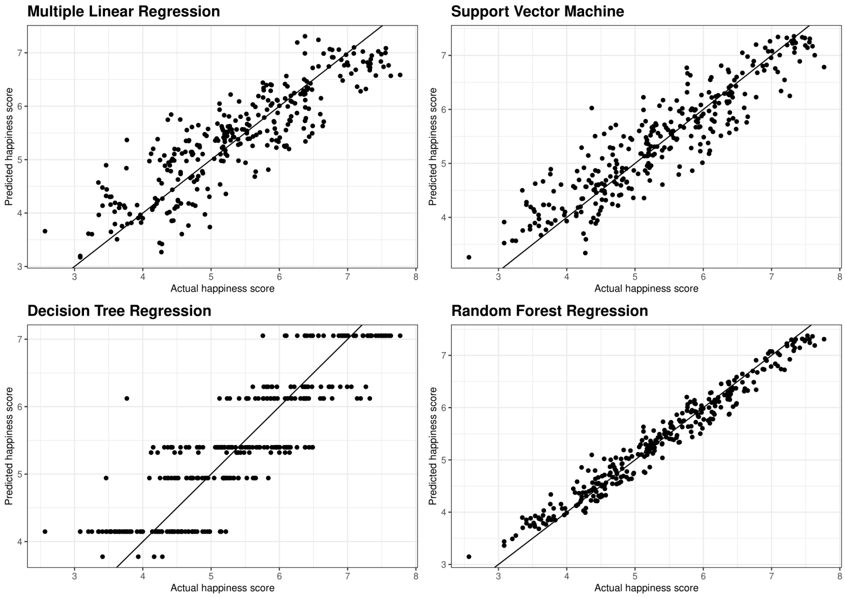

Figure 9: *Models' performance under the training test split ratio = 0.6*

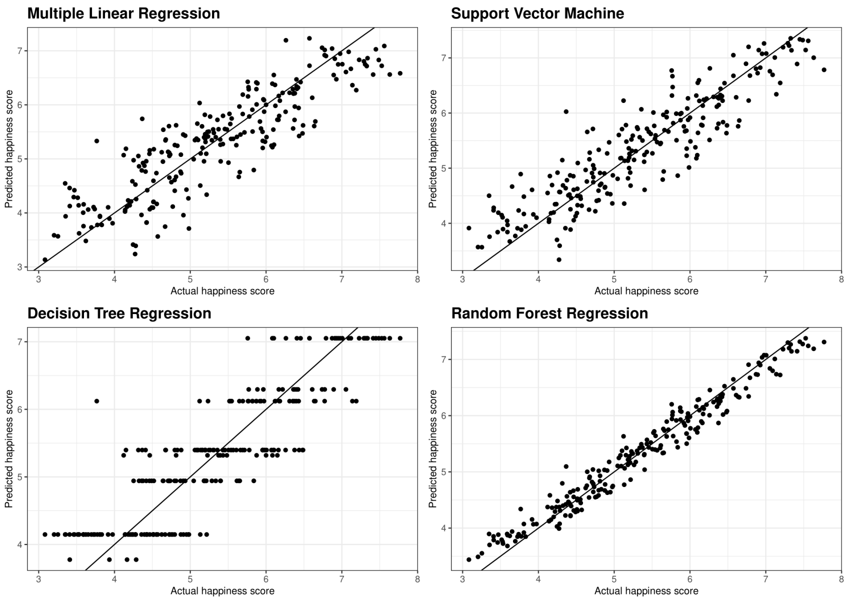

Figure 10: *Models' performance under the training test split ratio = 0.7*

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*.