The Economist INTELLIGENCE UNIT

Creation of a Composite Environmental Sustainability Index as an Indicator of a Country's Economic Performance

University Group Project by Team EcoSense

EcoSense

EXECUTIVE SUMMARY

This proof of concept proposes the creation of a new Environmental Sustainability Index (ESI) using machine learning to complement the Economic Intelligence Unit's (EIU) flagship product, its Country Report. The machine learning-driven approach aids in identifying patterns in data, providing EIU with a more quantitative and robust basis for its forecasting. Machine learning also does predictions quickly and the model makes all the predictions which reduces manpower needed. Furthermore, the accuracy of the model increases with an increased number of data and testing.

The ESI measures a country's environmental sustainability performance by considering key environmental variables of a country before assigning them with an Environmental Sustainability Score. The variables are weighted by using the absolute value of the coefficients of the seven environmental variables generated from the team's optimal linear regression model to reflect a certain percentage point in the calculations of the composite ESI. The ESI will be on a scale of 0 to 100, where the higher the score, the more environmentally sustainable the country is.

For the pilot study, the initial dataset consisted of data from 11 selected countries from the European Union (EU) over 10 years from 2010 - 2019. Upon realizing that there was not enough data for accurate predictions, all 27 countries in the EU were used instead. In this study, two forms of supervised machine learning - Classification and Regression Tree (CART) and Linear Regression were used. Based on the results, it was concluded that the CART model was not suitable for GDP per Capita. This is because realistically, the values may go beyond the range of the initial data input of the model. For linear regression, results revealed that the model with factored countries performed better than all other approaches, making it the best predictor of GDP per Capita due to its high degree of accuracy. The low error for the linear regression model also shows the effectiveness of machine learning in predicting GDP per Capita as it is both viable and feasible with lower costs required compared to expert opinion.

After linear regression, the study's final ESI consists of production-based CO_2 intensity, energy intensity and renewable energy supply, biomass, mean population exposure to PM2.5, environmentally related taxes, and population.

Analysis of the results between the relationship of ESI and GDP per Capita is performed from three perspectives. Firstly, the general trend indicates a positive correlation. Secondly, for each country, the relationship varies with some having negative correlation, some positive correlation and some no relationship. Lastly, the positive correlation remains largely similar across years.

The overall findings ascertain that machine learning can be used to create a new product component for EIU's Country Report. The creation of ESI would provide EIU's customers with a more comprehensive country analysis and a quantitative measure of sustainability, helping clients make more informed business decisions and improve their environmental sustainability. This increases EIU's value proposition and creates a new revenue stream for EIU.

Future improvements like the inclusion of more environmental variables or more past years' data may be included. With EIU's specialized domain knowledge and capabilities, EIU has the resources to further develop the machine learning model so as to create a more comprehensive Environmental Sustainability Index, bringing this machine learning pilot study to greater heights.

TABLE OF CONTENTS

BACKGROUND **1**

INTRODUCTION **1**

Revisions to Proposal 2

Data Preparation 2

CLASSIFICATION AND REGRESSION TREE APPROACH **3**

European Union as a Whole 3

Alternative Method: Remove "Country" as a Variable 4

Splitting Data into Subsets Based on GDP per Capita 5

CART Approach Conclusion 6

LINEAR REGRESSION APPROACH **7**

Evaluating Normality of Distribution 7

Evaluating Linearity 9

Evaluating Constant Error Variance 11

Evaluating for influential outliers 11

Evaluating Multicollinearity 12

Final evaluation of assumptions 12

Developing the Regression Model 13

COMPARISON BETWEEN LINEAR REGRESSION & CART **15**

Final Model Selection 15

CREATION OF ESI **16**

CONCLUSION **19**

FUTURE WORK **20**

REFERENCES **21**

APPENDIX A **22**

APPENDIX B **23**

APPENDIX C **25**

APPENDIX D **45**

APPENDIX E **50**

BACKGROUND

Currently, the Economist Intelligence Unit's (EIU) flagship country reports focus mainly on economic or political factors. Examples include the country's economic policies, economic indicators such as GDP and its political climate. The forecasts are primarily derived through non-machine learning methods like expert opinion and macro-indicators instead of machine learning. Specifically, EIU uses a global econometric model for 201 individual countries, applying statistical procedures in economic data to provide empirical analysis on economic relationships. This implies that EIU's focus is applying non-machine learning techniques to novel and large datasets in predictions with modifications according to the analysts' views.

The team would thus like to propose a complementary approach to EIU's statistical forecasting methodology - machine learning. Machine learning finds distinct and meaningful patterns in the data to make future predictions. The main advantage of machine learning is the ability to uncover generalizable patterns and discover new complex structures. It can fit complex and very flexible functional forms to the data, finding functions that are of good fit.

To provide a case study, machine learning has been successfully adopted by one of EIU's main competitors, Oxford Economics. The competitor uses the hybrid modelling method known as the Macroeconomic Error Correction that combines time series autoregression Arima machine learning model with purely statistical models such as Vector AutoRegression (VAR) and Dynamic-Stochastic General Equilibrium (DSGE) in their economic model. VAR can capture the linear interdependencies of time series data whereas DSGE replicates the behaviour of the economy. The hybrid model allows for predictive analysis such as for scenario planning, market sizing and forecasting to understand impacts of future economic or political events. Another competitor of EIU, the Organisation for Economic Co-operation and Development (OECD), also adopts machine learning. OECD uses nowcasting with Google trends and this approach allows them to predict GDP of the near future with higher accuracy based on big data gathered.

INTRODUCTION

Environmental sustainability has increasingly become an integral part of economic development for countries. According to the World Economic Forum's Global Risk Report 2020, the top five business risks were all linked to environmental challenges such as extreme weather events. With the effects of climate change impacting countries globally, a country's competitiveness has become inextricably linked with the effectiveness of their long-term sustainability strategies.

Considering the increasing importance of a country's environment, the creation of a new comprehensive Composite Environmental Sustainability Index is proposed, known as ESI. This report is a pilot study into the feasibility of this index which considers various quantitative environmental factors of a country, before assigning each country with an ESI score. This would provide countries with a benchmark of their level of environmental sustainability against other countries. The ESI will then be used in the prediction of a country's future GDP per Capita.

Revisions to Proposal

Scope

This report details the findings on the relationship between a country's environmental factors and its GDP per Capita from 2010 to 2019. Data across 21 variables were sourced from the websites of OECD as well as EIU and compiled into a .csv file.

After data exploration, it was found that constraints were faced due to the small amount of data points. As such, the scope of data collection and analysis has been extended beyond the initial proposal to include all 27 countries in the European Union (EU) Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, and Sweden. The list of variables and their coded data values can be found under **Appendix A**.

Methodology

In the proposal, the original intention was to use the Classification and Regression Tree (CART) to decide the variables to be used in the creation of the linear regression model. After which, the variables will then be plotted against GDP per Capita. This was a faulty decision as it is not appropriate to use different analytical techniques to determine the other's model. Furthermore, correlation is not causation. In a complex model, the variables that are said to be significant may not be true indicators but rather happen to have a high correlation with GDP per Capita.

For this report, a comparison was drawn between the prediction abilities of Classification And Regression Tree (CART) and Linear Regression. After weighing the advantages and disadvantages of each model, the superior method will then be selected to create the ESI. Based on a country's index, GDP per Capita was then predicted. Thereafter, the practicality of the index is then discussed based on the prediction results. Analysis was conducted using R Studio and the following packages were used: "data.table", "rpart" , "rplot", "e1071", "corrplot", "car", "ggplot2".

Data Preparation

Data cleaning improves data quality, providing for more accurate reporting, analysis and objective models. After compiling the data into a .csv file, the team performs cleaning in R Studio by importing the file as a data table using fread. All blanks will first be replaced with NA values. Secondly, columns were removed for those that have either more than one-quarter (at least 67) of the data points missing or those with countries that have missing data points for the whole period of 2010 to 2019. A total of 9 environmental indicators were dropped. Next, the team set a seed value before splitting the data into train-test sets using a 70%-30% division. For the train set, the team deliberately chose the period of 2010 to 2016 while the test set ranges from 2017 to 2019. This is to be in line with the objective of using machine learning to predict the future. Further unique data preparation steps will be mentioned under their respective sections.

CLASSIFICATION AND REGRESSION TREE APPROACH

There were two main approaches in creating the CART model. The first approach was to use the EU dataset as a whole to generate one CART model. The other would be to split the dataset into three different groups based on GDP per Capita to construct three separate CART models to see if prediction ability will improve.

European Union as a Whole

Methodology

The “Country” column was set as a factor as it is categorical in nature. No additional data preparation was required as the model can handle null values based on surrogates. This means that the model is able to predict the missing value based on variables deemed suitable as replacements.

After setting seed value to ensure consistency in result reporting, a CART model was created based on the randomly generated train and was set with no complexity penalty to grow the tree to the maximum. In this model, the “Year” variable was left out as it does not provide value to the model, as it acts as a reference point. The resulting base CART model predicts the continuous variable “GDP per Capita” using the environmental indicators. However, it is difficult to understand the impact of the independent variables on the dependent variable due to the large tree size as seen in Figure 1.1 under **Appendix B**. Therefore, the next step is to find the optimal tree size.

When using the command printcp, R automatically tests 10 randomly generated samples at a specific tree size and uses the average error value of these tests. The penalty triggers that would change the tree size are also found in the same table. Then, the optimal prune trigger that would give the simplest tree with the lowest relative 10-fold Cross Validation error is then used to prune the tree. This optimal tree was later used to predict test set values. To determine the accuracy of the optimal tree, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for the prediction of the GDP per Capita were calculated.

Findings

The optimal CART model, with “Country” included as an indicator, is illustrated in Figure 1.2 in **Appendix B**. The top 5 variables listed by variable importance function are: Country, Energy intensity, Production-based CO2 intensity, Total primary energy supply, Mean population exposure to PM2.5.

It has included lists of countries as a splitting criterion in many split conditions. For example, if the data points belonged to countries like Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Greece, Hungary, Italy, Latvia, Lithuania, Malta, Poland, Portugal, Romania, Slovak Republic, Slovenia, and Spain, they generally had a lower mean GDP per Capita than their counterparts. However, the model is not able to explain exactly why the countries are split like so. It could suggest that the model is not comprehensive enough, and further research is required.

Based on the Environmental Performance Index report by Yale, countries like Austria, Denmark, Luxembourg, Austria, Finland, Sweden, Germany, Netherlands are the top 11 in the world. Likewise, the CART model predicts that these countries are to have higher GDP per Capita than their counterparts which were countries like Spain, Slovenia, Czech Republic and Italy. Their ranking ranged from 14th to 20th. At this point, it could suggest that having higher sustainability would result in better GDP per Capita. However, Belgium is ranked at 15, placing after Spain. Despite this, Belgium is found to have higher GDP per Capita. This implies that there are additional factors not within the CART model that affect the GDP per Capita as this phenomenon also applies to other countries like Ireland or Malta.

After comparing the predicted GDP per Capita with the actual GDP per Capita of the test set, MAE is about 8.6% of the actual GDP per Capita. On the other hand, the RMSE was up to about 11.7%. This difference between the two errors suggests the presence of prediction outliers which deviates significantly from the actual GDP per Capita and thus resulting in a larger squared error and thus root mean squared error.

The concern regarding this method is that with the inclusion of "Country" as a variable, there is a possibility that there are other factors not within the CART model that affect the GDP per Capita. As such, these additional factors are taken by the model to be under the variable. However, this is not in line with the objective of the project which is to build an environmental index. Therefore, another method proposed is to remove the "Country" variable in constructing the model.

Alternative Method: Remove "Country" as a Variable

Procedures are similar to the previous method, except that in constructing the base model, the "Year" and "Country" variables are excluded. Refer to Figure 1.3 for the optimal model. The top 5 variables listed by variable importance function have changed to: Production-based CO2 intensity, Energy intensity, Mean population exposure to PM2.5, Total primary energy supply, Renewable energy supply excluding solid biofuels.

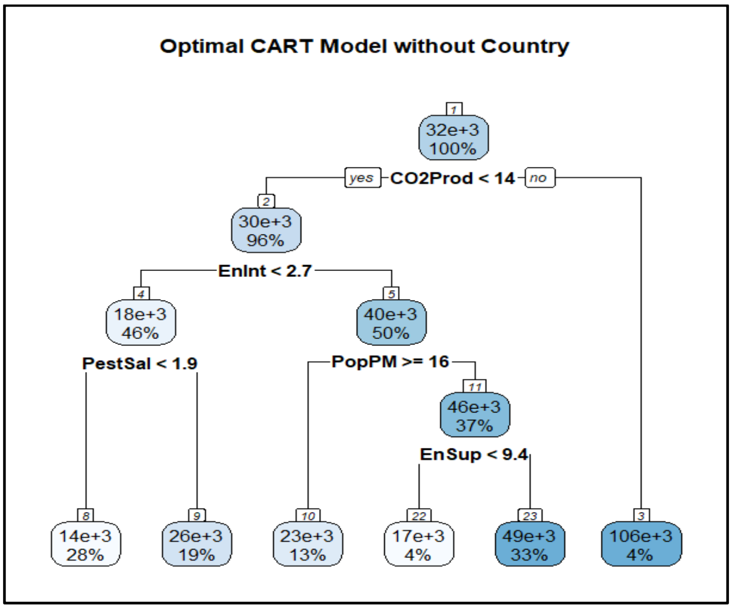


Figure 1.3: Optimal CART Model without Country as an indicator

The root node suggests that countries with a Production-based CO2 intensity of more than 14 would have a higher mean GDP per Capita. This suggests that countries with higher CO2 production than the benchmark are more likely to have higher GDP per Capita. Overall, it shows that countries with higher CO2 production, higher energy intensity, higher sales of pesticides, and higher primary energy supply with reference to the benchmarks are more likely to have higher GDP per Capita. To conclude, CART shows that higher pollution levels would cause a higher mean GDP per Capita. In reality, this is true as higher pollution levels are linked to increased productivity in a country and therefore a higher GDP per Capita.

The MAE for the prediction is about 17.58% of the actual GDP per Capita. On the other hand, the RMSE was up to about 25.84% of the actual value. This shows an increased amount of prediction points that deviates significantly from the actual GDP per Capita.

Removal of "Country" as a variable has caused errors to increase significantly. This shows that the "Country" variable itself is an important indicator for the model. It has indeed proven the original hypothesis that the inclusion of "Country" as a variable has caused the model to take other factors not within the CART model to be encapsulated within that variable. As the environmental factors do not fluctuate much within the 10 years, the only significant variable would be "Country", which has caused the predictions to change. It is believed that this is what has caused the high "Country" variable importance.

Splitting Data into Subsets Based on GDP per Capita

Given that these EU countries may have very different characteristics based on their market output, another approach is that the dataset can be further divided into smaller groups based on GDP per Capita.

Methodology

The dataset will be split into 3 different subsets based on GDP per Capita. The "low capita" group will have GDP per Capita of less than $20,000, "high capita" with more than $40,000 and "middle capita" with GDP per Capita in-between the two. With this, there should be a smaller margin of difference in the GDP per Capita within each segment. The aim is to see if the model's prediction ability has improved. Like the previous procedures, a CART model will be created based on the randomly generated train set for each of the datasets. Based on the analysis done in the previous section, these models will be generated without "Country" as a variable.

Findings

These findings will be specific to the "low capita" group as predictions for middle and high capita groups do not have any unique findings. With the separation of the countries, the prediction ability has improved slightly as compared to the CART basic model without "Country". For example, Latvia now has different predictions between the years 2017 to 2019. This result has never occurred before in the previous methods.

The top 5 variables listed by variable importance function have changed to: Energy intensity, Biomass, Production-based CO2 intensity, Renewable energy supply excluding solid biofuels, Population density.

MAE for the "low capita" group is about 17.36% of the actual GDP per Capita whereas the RMSE was up to about 20.67% of the actual value. The improved accuracy of this approach suggests that grouping countries into separate income levels will help to improve the predictions for GDP per Capita. However, the majority of the predictions of each country are still similar across the years. This is far from the ideal situation where GDP per Capita changes each year which is much more realistic.

Overall, while prediction ability has increased as the errors have decreased, this is still not a practical approach. The main reason is because it is very difficult to compare across various countries, as each grouping has different benchmarks. To illustrate, Figure 1.4 under **Appendix B** is the optimal CART model for "low capita" has a small optimal tree with only 2 splits with conditions of EnInt<2.8 and Biomass<25.

CART Approach Conclusion

Based on the findings above, CART is not an ideal approach to predict the time series data for this project as the data points generally do not vary significantly across the years. In addition, due to these minimal changes and the lack of yearly data points available, CART's predictions will always result in the same predictions for each country as illustrated by Figure 1.5. The key issue is that the country, as a whole, could be the most important indicator for GDP per Capita predictions as each country has its own unique range of values for each variable being considered. Therefore, the other variables are not as important as "Country" itself as it is the variable that seems to encapsulate the most change in GDP per Capita.

The predictive algorithm does not provide much insight into the future forecast of GDP per Capita as tree-based models are not good at extrapolation. This means that it is not suitable for predicting outside the ranges of the train set. It is highly unlikely in reality that a country would have the exact same GDP per Capita every year as implied by the piecewise approximations done by CART.

Nevertheless, some information from the CART model can still be gained, as most variables showing up in variable importance from the two machine learning approaches largely remain the same which will be discussed later on. Variables like Energy intensity, Production-based CO2 intensity and Mean population exposure to PM2.5 are always being listed with high importance. This suggests that these variables are crucial in predicting country GDP per Capita, which proves to be the same as the linear regression model findings.

Country	Year	GDPperCap	Prediction	SquaredErrors	AbsErrors
Bulgaria	2017	8353.97	9148.345	631031.641	794.375
Bulgaria	2018	8674.72	9148.345	224320.641	473.625
Bulgaria	2019	9058.74	9148.345	8029.056	89.605
Croatia	2017	15396.80	14017.810	1901612.530	1378.990
Croatia	2018	15971.15	14017.810	3815535.895	1953.340
Croatia	2019	16519.04	14017.810	6256149.899	2501.230
Estonia	2017	19153.39	17487.797	2774201.152	1665.593
Estonia	2018	19918.16	17487.797	5906665.932	2430.363
Hungary	2017	15912.01	14017.810	3587992.418	1894.200
Hungary	2018	16792.49	14017.810	7698847.312	2774.680
Hungary	2019	17579.84	14017.810	12688055.423	3562.030
Latvia	2017	15507.65	14017.810	2219622.264	1489.840
Latvia	2018	16257.91	9148.345	50545914.489	7109.565
Latvia	2019	16703.20	9148.345	57075834.071	7554.855

Figure 1.5: CART Prediction Results

LINEAR REGRESSION APPROACH

In the linear regression approach, multiple diagnostic measures must first be utilised to evaluate the suitability of linear regression. This is done by mainly evaluating if the following assumptions hold:

1. **Linearity assumption:** There is a linear relationship between the individual environmental variables, the independent variables, and GDP per capita, the dependent variable.
2. **Normal distribution assumption:** The errors follow a normal distribution.
3. **Constant variance assumption:** The errors have a constant variance.
4. **Little multicollinearity assumption:** Little multicollinearity between respective dependent variables
5. **No influential outliers assumption**: There must not be significant outliers that hugely influence the regression model.

Evaluating Normality of Distribution

Plotting of Histograms

Firstly, to observe the distribution of data, the team plotted histograms to evaluate the distribution of each individual variable. Based on the histograms, several environmental variables, as well as GDP per capita, had skewed data distributions, leading to a “non-normal” distribution. This would violate the normality of errors assumption.

Skewness

Afterwards, this was verified by calculating the skewness of each variable to see if skewness was greater than 1 or lesser than -1. This is because skewness measures if the distribution of a continuous variable is asymmetrical, where a value greater than 1 indicates a high degree of right skewness and a value greater than -1 indicates a high degree of left skewness.

CO2Prod	EnInt	EnSup	ReEng1	ReElect	ReEng2	Biomass	PestSal	PopPM	EnvTax	PopDens	GDP/Cap
1.7901	1.1589	2.2128	0.9215	0.6858	1.0657	0.4519	1.9231	0.1361	0.4117	3.7377	1.4520

Figure 2.1: Skewness of variables

Apart from GDP per capita, 6 of the variables, namely, Production-based CO2 intensity (CO2Prod), Energy intensity (EnInt), Total primary energy supply (EnSup), Renewable energy supply (excluding solid biofuels) (ReEng2), Sales of pesticides per unit of agricultural land (PestSal) and Population density (PopDens) also had high right skewness. This means that the median value of the data would be lesser than the mean and thus would violate the normality assumption of the data.

Hence, the team applied logarithmic transformation on GDP per Capita and the aforementioned 6 variables with a right high skewness to linearize the variables.

Histograms were then plotted for each of these logged variables to evaluate whether there were improvements in the distribution of data. It can be seen that after log was applied to GDP per capita and the 6 variables, the histograms had a more “normal” distribution.

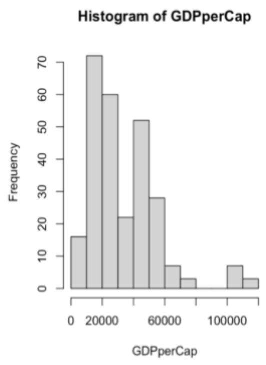
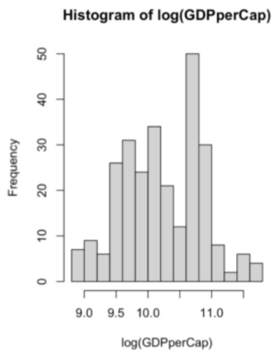
Histogram of GDP per Capita	Histogram of log(GDPperCap)
	

Figure 2.2: Histograms to illustrate effect of log on GDP per Capita

For example, the plotting of the above histogram for GDP per Capita (GDPperCap) reveals how its distribution is rightly skewed. Upon applying log on GDPperCap, the distribution becomes much more normalized.

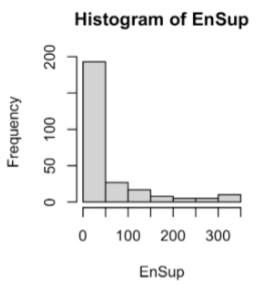
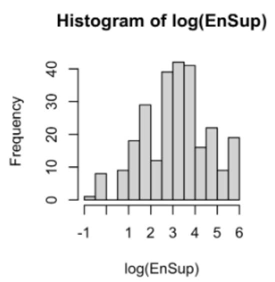
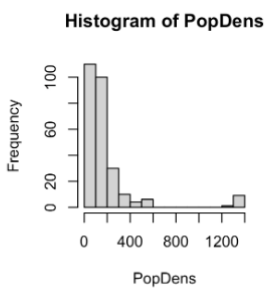
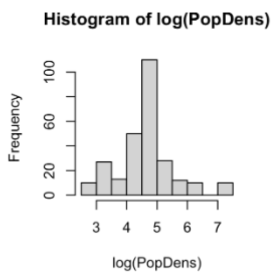
EnSup	log(EnSup)	PopDens	log(PopDens)
			
Skewness: 2.212798	Skewness: -0.2340014	Skewness: 3.7377154	Skewness: 0.4689561

Figure 2.3: Histograms to illustrate effect of log on environmental variables

After which, the skewness of the affected variables, after applying the logarithmic transformation to appropriate variables, were then checked again. This would check if the transformation was helpful in normalizing the errors. After conducting the check, we found that none of the variables had a high skewness value greater than 1 or lesser than -1.

log CO2Prod	log EnInt	log EnSup	ReEng1	ReElect	log ReEng2	Biomass	log PestSal	PopPM	EnvTax	log PopDens	log GDP/Cap
0.4974	0.3445	-0.2340	0.9215	0.6858	0.9214	0.4519	0.3493	0.1361	0.4117	0.4690	-0.0144

Figure 2.4: Skewness of variables

Therefore, with the newly transformed variables, it is evident that the variables are more likely to uphold the normal distribution assumptions. Refer to **Appendix C** for the detailed histograms and skewness calculations.

To further evaluate the normal distribution, skewness has to be combined with other measures such as developing model diagnostic Normal Q-Q plots for better evaluation.

Normal Q-Q Plot

The normal Q-Q plot was used to check the normal distributed errors assumption. Ideally, most points should fall close to the diagonal dotted line. The team below discussed the relationship between GDP per Capita (GDPperCap) and production-based CO_2 intensity (CO2Prod).

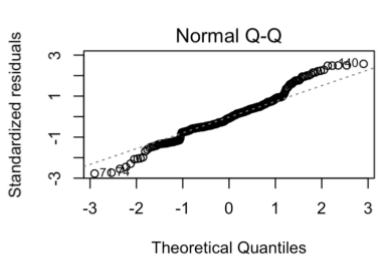
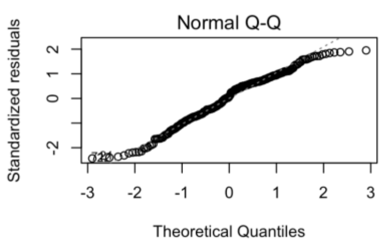
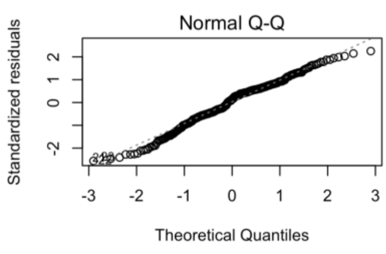
GDPperCap against CO2Prod	log(GDPperCap) against CO2Prod	log(GDPperCap) against log(CO2Prod)
		

Figure 2.5: Skewness of variables

In the original Normal Q-Q plot, the team noted a relatively significant deviation in the tail ends of the line from the diagonal dotted line, suggesting problems at the distribution tails. Upon applying log to GDPperCap, and subsequently the applying log to CO2Prod, the deviation at the tail ends of the line decreases, and the error distribution becomes more normalized. Hence, applying log to GDPperCap and the 6 environmental variables aids in normalization and satisfying the normal distributed errors assumption. Refer to **Appendix C** for the detailed Normal Q-Q Plots for each variable.

Evaluating Linearity

Correlation Coefficient

To obtain a preliminary sensing of the relationship between the variables, the team evaluated the correlation between each variable. The correlation value shows if the 2 variables have a linear relationship and how strongly related they are. This is not to imply that there is a causal relationship between the variables, as correlation does not imply causation. The team calculated the correlation between GDP per Capita and the respective environmental variables.

	CO2Prod	EnInt	EnSup	ReEng1	ReElect	ReEng2	Biomass	PestSal	PopPM	EnvTax	PopDens
Non-log	0.5797	0.7032	0.1243	-0.0198	0.2026	0.2344	0.2073	0.0199	-0.5193	-0.1714	0.0920
Log	0.4787	0.6434	0.1760	0.0120	0.1806	0.2996	0.2774	0.2552	-0.5875	-0.1116	0.2354

Figure 2.6: Correlation Coefficients of Environmental Variables against GDP per Capita

Upon running the correlation plot, it is evident that GDP per Capita has a relatively strong correlation with energy intensity (EnInt) and is moderately correlated with production-based CO_2 Intensity (CO2Prod) and mean population exposure to PM2.5 (PopPM). Yet, other variables may appear to have a weak correlation with GDP per Capita.

Nevertheless, upon adding the logarithmic transformation, the correlation coefficient of variables with weak correlation increases:

- EnSup's correlation coefficient increases from 0.12 to 0.18
- ReEng2's correlation coefficient increases from 0.2344 to 0.2996
- PestSal's correlation coefficient increases from 0.0199 to 0.2552
- PopDens' correlation coefficient increases from 0.09 to 0.24

However, applying log on variables with stronger coefficients, namely, CO2Prod and EnInt, led to a decrease in the correlation coefficient.

- CO2Prod's correlation coefficient decreases from 0.5797 to 0.4787
- EnInt's correlation coefficient decreases from 0.7032 to 0.6434

Given the spread of the data, correlation coefficient might not be a sufficient measure of linearity. Hence, the team made use of diagnostic plots to further evaluate the relationship between the variables and GDP per Capita in order to provide better evaluation on whether the variables can uphold the assumptions required for linear regression.

Residuals versus Fitted Plot

The residuals versus fitted plot was used to check the linear relationship assumption. Ideally, the plotted points should be randomly distributed about zero.

GDPperCap against CO2Prod	log(GDPperCap) against CO2Prod	log(GDPperCap) against log(CO2Prod)

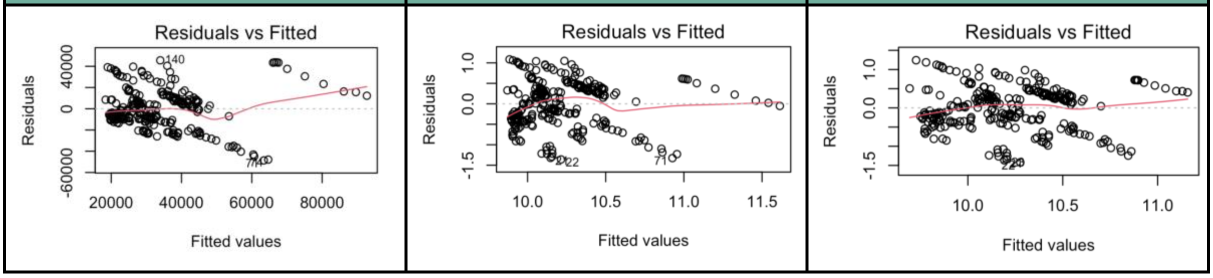


Figure 2.7: Residuals vs Fitted Plots

When a logarithmic transformation was applied to GDP per Capita (against CO2Prod), there was an improvement on the residuals versus fitted plot as the tail ends of the line became much closer to zero. This indicated that a log on GDP per Capita contributed to an improvement in the linear relationship between GDP per Capita and variables.

When a logarithmic transformation was applied to CO2Prod (with log(GDPperCap) as the X variable), there was a further improvement on the residuals versus fitted plot as the line became closer to zero. Hence, applying log to GDPperCap and CO2Prod aids in satisfying the linearity assumption, despite the original fact that there was a slight decrease in correlation coefficient.

Upon evaluating the Residual versus Fitted Plots of each variable (refer to **Appendix C**), it is evident that there is a linear relationship between each GDP per Capita and each of the transformed variables.

Evaluating Constant Error Variance

Scale-Location Plot

The scale-location plot checks the constant error variance assumption. Ideally, the plotted points should be randomly distributed with the same spread at each vertical slice.

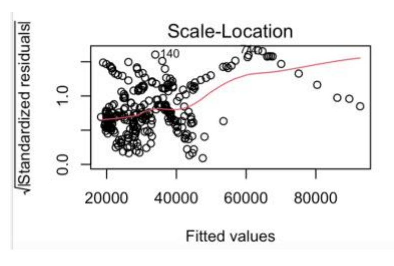
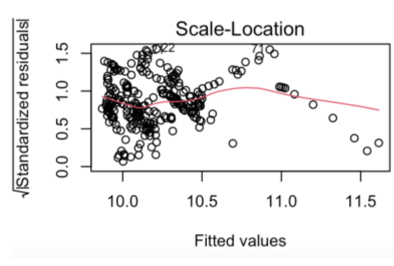
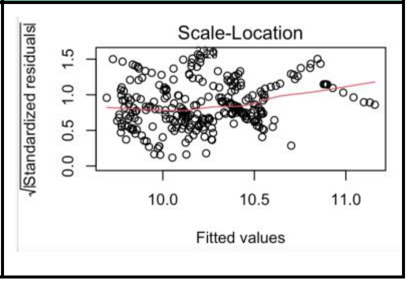
<table>
<tr><th>GDPperCap against CO2Prod</th><th>log(GDPperCap) against CO2Prod</th><th>log(GDPperCap) against log(CO2Prod)</th></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Figure 2.8: Scale-Location Plots

The team noted some minor issues at the rightmost slice due to insufficient data points. However, upon applying log to GDPperCap, and subsequently, CO2Prod, there was greater evenness in terms of spread. Hence, applying log aided in satisfying the constant error variance assumption.

Evaluating for influential outliers

Residuals versus leverage plot

The residuals versus leverage plot checks for influential outliers. Any point that breached the red dotted "gates" at the top right or bottom right of the chart is an influential outlier.

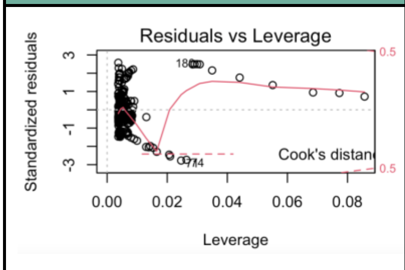
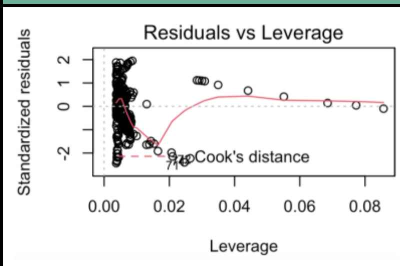
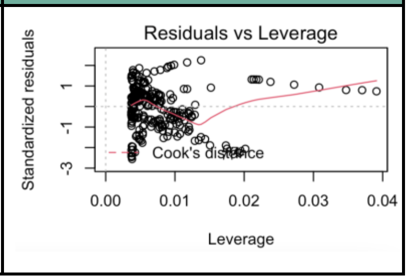
<table>
<tr><th>GDPperCap against CO2Prod</th><th>log(GDPperCap) against CO2Prod</th><th>log(GDPperCap) against log(CO2Prod)</th></tr>
<tr><td></td><td></td><td></td></tr>
</table>

Figure 2.9: Residuals versus Leverage Plots

From the above plots, it can be concluded that there are no influential outliers as there are no data points beyond any red dotted gates at the top right or bottom right of the graph. As such, there is no need to remove any outliers from the dataset as there are no data points that would influence the linear regression model and coefficients significantly. This implies that the data upholds the 5th assumption that there are no influential outliers that would require further evaluation of that data point.

Evaluating Multicollinearity

Variance inflation factor (VIF)

Another assumption that the team needed to address is to ensure that there are no multicollinear variables. Multicollinearity occurs when a variable can be expressed as a linear combination of other variables, meaning that that variable is dependent with intercorrelations. This contradicts the assumption that variables are independent which would make the regression equation less reliable as the regression needs to evaluate the individual impact of each variable.

Therefore, to identify multicollinearity, the Variance Inflation Factor (VIF) is the tool utilised to measure the multicollinearity of the variables.

log CO2Prod	log EnInt	log EnSup	ReEng1	ReElect	log ReEng2	Biomass	log PestSal	PopPM	EnvTax	log PopDens	log GDP/Cap
4.5959	3.97054	2.3777	8.0242	5.8321	3.7163	1.8966	2.6788	2.1882	1.7742	3.8410	-

Figure 2.10: Variance Inflation Factor (VIF) of environmental variables

As there is no consensus on the cut off of what VIF value determines multicollinearity, it was defined that multicollinearity occurs when VIF is greater than 10. Looking at each of the variables, none of the VIF coefficients are greater than 10, which implies that the variables are not multicollinear. As such, the assumption that there is little multicollinearity between variables is upheld.

Final evaluation of assumptions

Upon examining all the different tools and diagnostic checks used to evaluate the assumptions, it was concluded that all assumptions have been met and linear regression is a suitable model to evaluate the data.

Reflecting upon the 5 main assumptions stated earlier:

1. **Linearity assumption:** Upon evaluating the correlation coefficient and residual fitted plot, the team determines that there is a linear relationship between GDP per Capita, the dependent variable, and the individual environmental variables, with 6 of the variables being transformed.
2. **Normal distribution assumption:** Upon evaluating the histograms, skewness and Normal Q-Q Plots, GDP per Capita and the environmental variables approximately follow a normal distribution. Despite GDP per Capita and 6 of the environmental variables being skewed initially, the normality of the errors improved significantly upon applying a logarithmic transformation.
3. **Constant variance assumption:** The scale-location plot reveals that the errors of GDP per Capita against the different variables have a relatively constant variance.
4. **Little multicollinearity assumption:** There is little multicollinearity between respective dependent variables as shown when all variables have a VIF value less than 10.
5. **Little influential outliers assumption:** From the Residuals vs Leverage plot, there are no significant outliers that would drastically influence the linear regression model.

Developing the Regression Model

Methodology

The team ran linear regression for all variables excluding "Year" on the train set. "Year" is being removed as the objective of this regression is to identify environmental variables that are significant to GDP per Capita. The train set is being used as the team's goal is to predict GDP per Capita for the future. The team prepared 2 versions of the model, one with 'Country' and another without 'Country'. Including 'Country' as a factored categorical variable would consider the impacts of various countries in the prediction of GDP per Capita.

The regression equation for the model which included factored 'Country' is as shown below:

log(GDPperCap) = log(CO2Prod) + log(EnInt) + log(EnSup) + ReEng1 + ReElect + log(ReEng2) + Biomass + log(PestSal) + PopPM + EnvTax + log(PopDens) + Country

The regression equation for the model which excluded factored 'Country' is as shown below:

log(GDPperCap) = log(CO2Prod) + log(EnInt) + log(EnSup) + ReEng1 + ReElect + log(ReEng2) + Biomass + log(PestSal) + PopPM + EnvTax + log(PopDens)

In order to select the most suitable variables for the final regression model, the team decided to look at both the P-value and Akaike Information Criterion (AIC).

P-value

In order to evaluate the statistical significance of values, the team first looked at p-values. This is obtained by finding the values with a p-value lesser than a significant level of 0.05. The p-value in a linear regression represents the percentage risk of determining that a relationship exists when there is no real relationship. As such, a p-value lower than our team's chosen significant level of 0.05 means that there is enough evidence to reject the null hypothesis that there is no relationship between the variable and GDP per Capita.

	log CO2Prod	log EnInt	log EnSup	ReEng1	ReElect	log ReEng2	Biomass	log PestSal	PopPM	EnvTax	log PopDens
With countries	0.00898	0.20799	0.66699	0.43520	0.77519	0.00126	0.00567	0.85252	0.00012	4.585e-7	0.096007

Figure 3.1: P-Values of environmental variables

According to the linear regression model with factored countries, the following variables are statistically significant: log(CO2Prod), log(ReEng2), Biomass, PopPM, EnvTax and log(PopDens).

Detailed screenshots can be found in **Appendix D**.

Akaike Information Criterion (AIC)

In addition to p-values to select the best algorithm, the team also looked at the Akaike Information Criterion (AIC). AIC is a measure of goodness of fit that favours smaller residual error in the model and penalises for including further predictors which will help to avoid overfitting. This is because if one merely looks at individual p-values, the results might be less accurate due to the fact that variables that are collinear will naturally have higher p-values. Hence, backward elimination was run for each version to automatically select the best algorithm with the lowest AIC. Next, the team ran linear regression on the best algorithm and predicted GDP per Capita.

For the model with factored 'Country', there are a total of 7 environmental variables chosen out of the 11 variables, with the regression equation for the model as shown below:

log(GDPperCap) = log(CO2Prod) + log(EnInt) + log(ReEng2) + Biomass + PopPM + EnvTax + log(PopDens) + Country

log(CO2Prod)	log(EnInt)	log(ReEng2)	Biomass	PopPM	EnvTax	log(PopDens)
0.002340	0.000098699	0.000229	0.006235	0.00001299	0.0000005417	0.00001864

Figure 3.2: P-Values of environmental variables

For the model with factored 'Country', there are a total of 10 environmental variables chosen out of the 11 variables, with the regression equation for the model as shown below:

log(GDPperCap) = log(CO2Prod) + log(EnInt) + log(EnSup) + ReElect + log(ReEng2) + Biomass + log(PestSal) + PopPM + EnvTax + log(PopDens)

log CO2Prod	log EnInt	log EnSup	ReElect	log ReEng2	Biomass	log PestSal	PopPM	EnvTax	log PopDens
8.58e-5	<2e-16	2.67e10	0.000906	<2e-16	1.11e-12	0.000129	2.77e-9	0.011787	6.16e-7

Figure 3.3: P-Values of environmental variables

In both final regression models, each environmental variable is statistically important to the model as seen from the low p-value. However, in the model without factored 'Country', it is evident that many of the p-values of environmental variables are extremely low compared to the model with factored 'Country'. This is probably due to the fact that the linear regression model without factored 'Country' is highly influenced by 'Country' in comparison to the other variables. Hence, it is necessary to use the model with factored 'Country' instead.

	With Country	Without Country
Adjusted R-Squared	0.9963	0.8806
Residual standard error	0.3881	0.2278
MAE	6.89%	22.39%
RMSE	11.50%	35.01%

Figure 3.4: Statistical comparison of the two models

Furthermore, the model with country has a higher multiple and adjusted R-square than the model without country, indicating a higher explanation power. The residual standard error is also lower for the model with country, indicating a better fit of the data.

Looking at the predictive power, MAE for the model with country is about 6.89% of the actual GDP per Capita whereas the RMSE was up to about 11.50% of the actual value. On the other hand, the MAE for the model without country is about 21.3% of the actual GDP per Capita whereas the RMSE was up to about 27.59% of the actual value. Based on the accuracy of the predictions for GDP per Capita shows that the model with Country is a better model to be used.

COMPARISON BETWEEN LINEAR REGRESSION & CART

In order to determine the ideal model used to predict GDP per Capita, a comparison will be made between the CART model and the Linear Regression model. The following is a top-level comparison at each model's characteristics.

	Linear Regression	CART
Output	Continuous	Categorical or continuous
Prediction Ability	Able to predict a point of x-value which is not within the range of the train set	Looks for certain consistency in the data and makes estimates to fit the train set
Data Preparation	More data preparation needed; have to manually handle missing values from raw data such as by taking the mean	Little data preparation needed; uses surrogate values to handle missing values from raw data
Assessing Significance	p-value	Variable Importance

Figure 3.5: Comparison of the two models

Final Model Selection

Based on the aforementioned characteristics, linear regression is the most ideal model to adopt. This is because of its key advantage of being able to predict a point of x-value which is not within the range of the train set, allowing for extrapolation. It is the most feasible approach to predicting GDP per Capita as realistically, GDP per Capita can continue to grow from previous years. The disadvantages of linear regression can be negated through certain command procedures and ensuring data quality.

Using the CART model would always constrict GDP per Capita to be within the original train set's boundaries, resulting in inaccurate predictions in the future. Furthermore, its predictions will always result in a "fixed value" based on the splitting conditions determined, and countries with similar indicators are likely to have the exact same predicted GDP per Capita. By combining linear regression findings with domain knowledge, the result can also be explained more contextually, thus providing better usability in helping countries to find ways to improve their environmental sustainability.

CREATION OF ESI

Linear regression will be used to create the ESI as it is the model with the higher predictive power. Weightage is created by using the coefficients of the seven environmental variables in the model and absolute values will be applied. The ESI is created based on the weightage assigned to the seven environmental variables as shown below:

	log(CO2Prod)	log(EnInt)	log(ReEng2)	Biomass	PopPM	EnvTax	log(PopDens)
Coefficients	-0.269	0.542	0.088	-0.002	-0.014	-0.096	-1.096
Weightage	0.128	0.257	0.042	0.001	0.007	0.046	0.520

Figure 4.1: Weightage assigned to the seven variables

In order to make the ESI more understandable to the general users, the team scales the seven environmental variables from 0 to 100 based on the maximum value for each variable before assigning the weightage. Thus, making the ESI on a scale of 0 to 100 as well. For the distribution of the ESI, the minimum value is 45.24 and the maximum is 80.57. After checking the skewness of the ESI, there is no need to apply logarithmic transformation. To visualize the distribution, refer to **Appendix E**.

Testing is done by running linear regression with the ESI, Year and Country. As expected, the adjusted R-square has decreased as compared to the original model with the environmental variables separately. However, the value still remains high at 0.9934. In addition, the ESI also has a very low p-value that is less than 0.001 which shows a high statistical importance in the prediction of GDP per Capita (Refer to **Appendix E**).

General Trend Between ESI & GDP per Capita

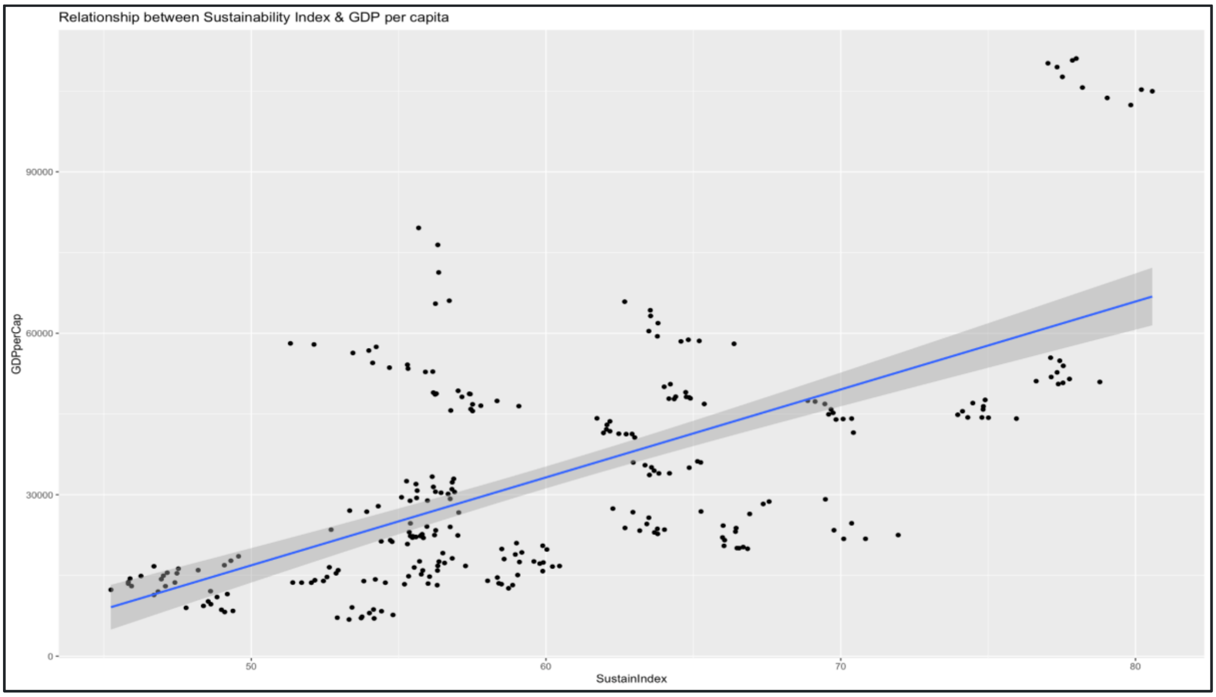


Figure 4.2: Relationship between ESI and GDP per Capita

Referring to the diagram above, it can be inferred that GDP per Capita and the ESI are positively correlated. The shaded band is a pointwise 95% confidence interval on the fitted values which means that there is 95% confidence that the true regression line lies within the shaded region. It shows us the uncertainty inherent in the estimate of the true relationship between GDP per Capita and the ESI. To understand the uncertainty of the relationship, each country's performance will be analyzed on a case-by-case basis.

Relationship Between ESI & GDP per Capita by Country

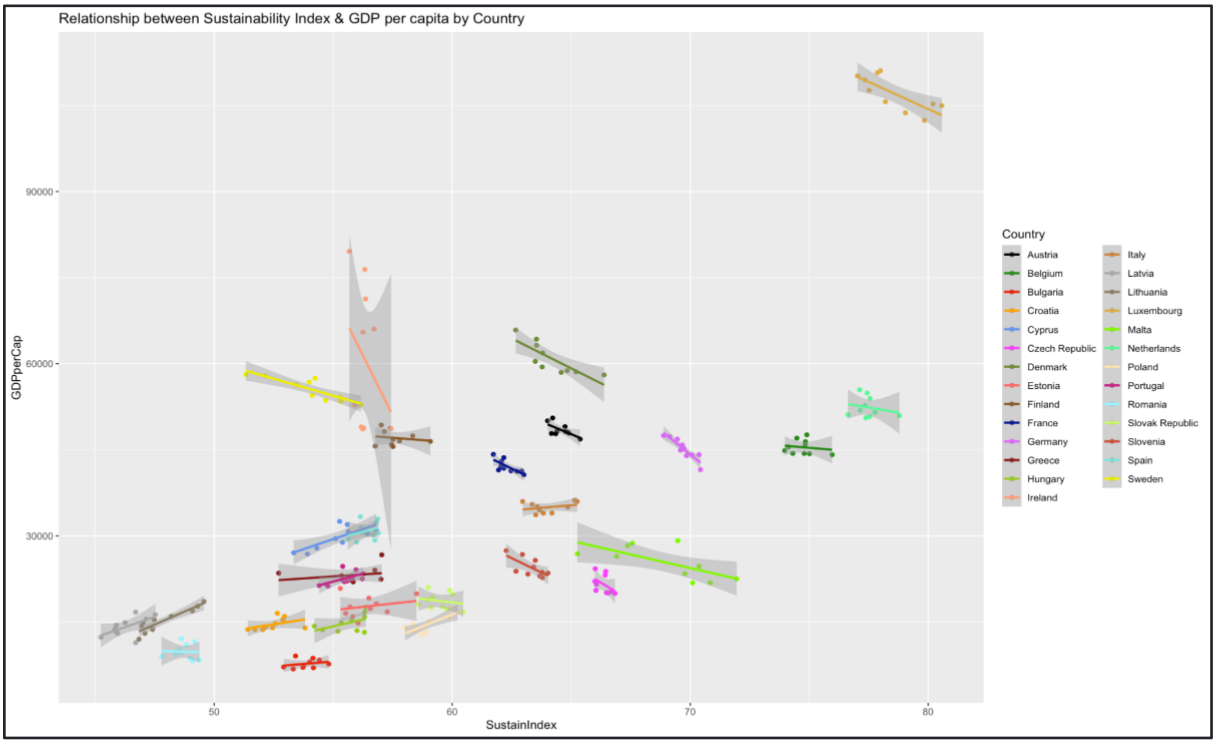


Figure 4.3: Relationship between ESI and GDP per capita by Country

This diagram shows that GDP per Capita and the ESI is not positively correlated for all countries. For example, in Romania, the horizontal-seeming line seems to indicate there is not much relation between GDP per Capita and the ESI. For Sweden, the downward slope indicates a very strong negative correlation between GDP per Capita and the ESI. On the flipped side, in Cyprus, the upward slope indicates a very strong positive correlation between GDP per Capita and the ESI. In-depth analysis has been conducted on the following countries: Romania, Sweden and Cyprus. This is where differentiation occurs and allows EIU to form an independent country view for the flagship country reports, helping clients make better business decisions by providing a thorough understanding of the countries.

For Romania, it is ranked number 39 in the Sustainable Development Goals (SDGs) Report. It has just developed its own National Sustainable Development Strategy in November 2008 and is currently in the process of mapping the SDGs to determine the best ways to achieve their goals and targets. This may be the reason why there is not much relationship between GDP per Capita and the ESI. The effect of environmental sustainability on GDP per Capita depends on the environmental effort placed and may need more time to be visible. By looking at the shaded band, the "real" relationship can either be positive or negative in the long run.

For Sweden, it is ranked second in the SDGs Report and is well known as a pioneer in environment sustainability. It was one of the first countries to introduce environmental tax since 1995. Yet, being the top five wealthiest countries in the EU as of 2020 according to Bloomberg. This indicates a high economic activity happening and is inevitable that Sweden will generate a higher environmental footprint in the process. This trade-off that exists between environmental sustainability and economic activity may be the reason for the negative correlation. For instance, when citizens are more environmentally conscious, the environmental tax that contributes to the GDP per Capita will be lower. Especially when the environmental tax is one of the environmental variables in the creation of ESI, this further illustrates why when ESI increases, GDP per Capita decreases. However, this does not mean that it is bad to be more environmentally conscious. Looking from another perspective, Sweden has been investing significantly in environmental effort yet still manages to sustain a high GDP per Capita demonstrates that environment and development can go hand in hand.

For Cyprus, it is ranked number 40 in the SDGs Report. Even though Cyprus's environment performance is below EU average, Cyprus has given attention to certain priority areas such as SDG 8 which is about the promotion of green development and the creation of green jobs. For instance, the government has prepared an investment plan of more than €3.6 billion to work on the three main pillars of environment, economic and social development. This may be the reason why when ESI increases, GDP per Capita increases.

Relationship between ESI & GDP per Capita by Year

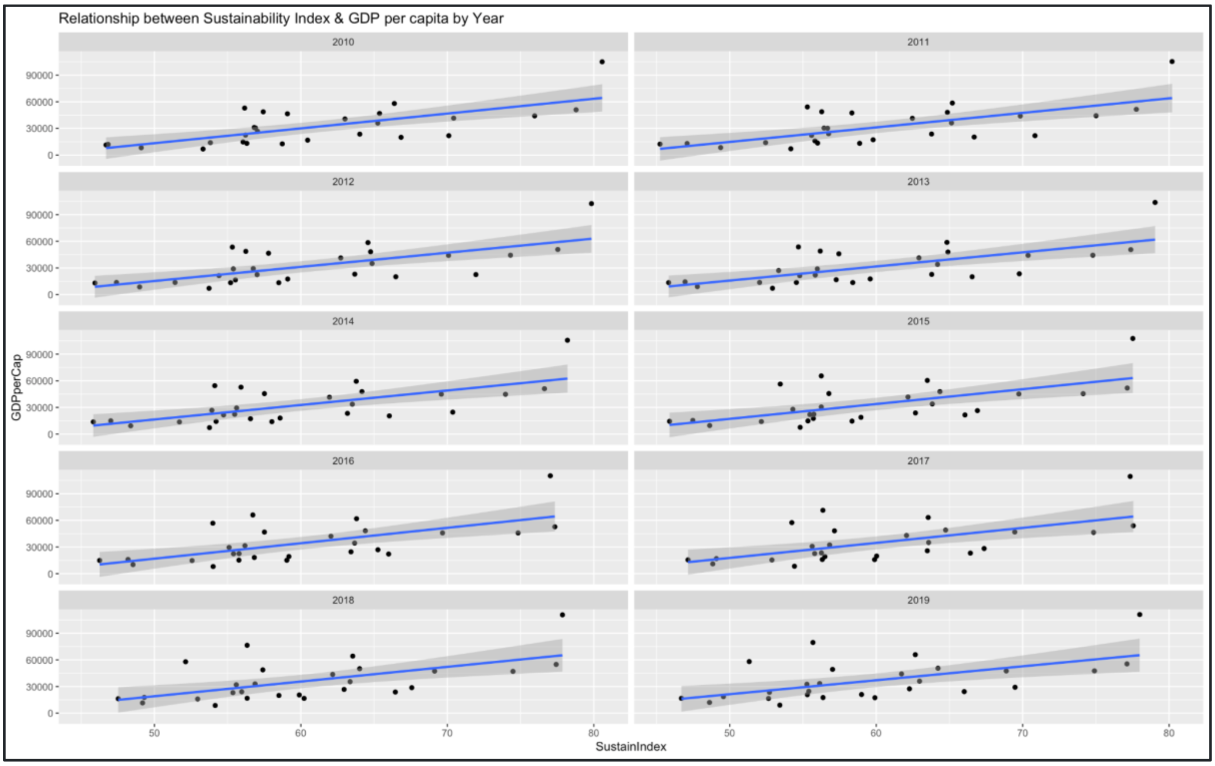


Figure 4.4: Relationship between ESI and GDP per Capita by Year

Our findings reveal that there is generally a positive correlation between the ESI and GDP per Capita. However, there is little variation for the relationship between ESI and GDP per Capita across the years for each country. The graph indicates that a short time period of one year is insufficient to establish a relationship between environmental factors and economic performance. Therefore, the team recommends EIU to analyse country data with an even longer time horizon e.g. a 10-years horizon in order to better establish a changing relationship between environmental factors and economic performance across time and increase the business value of the ESI.

CONCLUSION

The pilot study has created an ESI comprising 7 environmental variables that can predict GDP per Capita. With the breakdown of the various variables, EIU is able to provide countries with a more quantitative measure to improve their sustainability without having a significant negative impact on GDP per Capita. Referring to Figure 4.1, the renewable energy supply (ReEng2) has a positive correlation with the GDP per Capita, while CO2 Production (logCO2) has a negative correlation with GDP per Capita. Within the EU, the renewables sector accounted for over 1.5 million jobs in 2018 proving that investing in renewable energy supply will potentially increase GDP per Capita. Furthermore, the EU has pledged to cut CO2 emissions by at least 55% by 2030 compared with 1990 levels. With the EU's push towards sustainability, economically advanced countries have greater financial capacity to adopt environmentally friendly methodologies to reduce CO2 emissions.

The overall findings ascertain the feasibility of using machine learning in creating a new product component - the ESI, for EIU's Country Report. The creation of ESI would provide EIU's customers with a more comprehensive country analysis and a quantitative measure of sustainability, helping clients to make more informed business decisions and improve their business' environmental sustainability. The ESI is in line with EIU's business activity, which is to provide forecasting and advisory services to its clientele base who are primarily business leaders. These business leaders rely on these services to make informed business decisions like company expansions or to understand the impact of macro-environmental factors on their business. Incorporating the ESI will expand EIU's current product line and thus improve situation visibility as well as providing a more comprehensive country analysis. This could also possibly provide an additional stream of revenue to its current business model. With its success, the company would be able to provide its users with a benchmark on how a country is performing in terms of environmental sustainability on a global scale.

Another benefit would be an increased value proposition. Due to the growing trend of customers being inclined to companies who have good sustainability practices, more companies are increasing their sustainability efforts as a form of Corporate Social Responsibility. The ESI would be useful for the business leaders as it would allow them to understand a country's future economic performance as well as its green practices or regulations. Thus, enabling them to make better investment decisions like expanding to countries with greater support for companies with more environmentally-friendly practices.

With a quantitative measure of sustainability and a more accurate prediction of countries' economic performance potential, countries would also benefit by being better able to refine their environmental policy agenda by learning from other countries, and make more informed environmental sustainability investments to advance towards a sustainable future. Moreover, with the increasing emphasis on environmental sustainability, this ESI would become a key consideration for business leaders in the years to come to analyse a country's performance holistically.

FUTURE WORK

For EIU to adapt this pilot study on a global scale, there are additional steps that need to be taken to ensure the success of this product line. This section details the four key improvements that can be made.

Firstly, it might be worth noting that a country moving towards sustainability may not see an immediate positive effect on its economy. With this in mind, it would be preferred that more past years' data be included with a longer time horizon. This not only serves to build a more comprehensive model, but also allows EIU to explore if there are possible connections between past data and future GDP per Capita. This means that it might be possible that predicted GDP per Capita in future years could be more greatly correlated to a country's environmental factors from several years ago.

Secondly, it would be better to include more environmental variables in the ESI. These variables need not be continuous data like the amount of air pollution. Categorical data like whether a country has signed a new environmental treaty may be included in its calculations. This would make the index more meaningful and provide greater insight into the environmental status of a country.

Thirdly, EIU may want to consider more frequent data collections. Gathering data more frequently, such as on a monthly basis, would allow EIU to generate a more comprehensive model. This would improve prediction accuracy as machine learning works best with a larger dataset, assuming that data quality is assured. Having a larger dataset will also minimize the impact on the prediction incurred by outliers and reduce the prediction error.

Lastly, for the missing data points, it will be preferable to find the actual value rather than using the mean value or dropping them. Machine learning is data-driven, which means that it uses the input data to make predictions. Due to this, the data provided will inevitably affect its predictions. By providing the actual values rather than an estimate, this improves data quality and thus increases the accuracy of predictions by using the best information available.

A separate pilot study can also be conducted to evaluate the effectiveness of the CART model in predicting a different type of data using the same dataset. While it is not ideal to predict continuous data like the GDP per Capita, CART works well in predicting categorical variables like in predicting the response of "Yes" or "No" to a question. It may be possible to use CART to make short-term predictions on whether GDP can increase by a predetermined percentage.

Overall, with EIU's long history and wide-reaching operations, the division has the means to further develop the machine learning models to produce an extensive Environmental Sustainability Index for its clients. With the incorporation of machine learning into its forecasting methodology, the possibilities are endless.

REFERENCES

About us. (n.d.). Economist Intelligence Unit. Retrieved 24 October 2021, from https://www.eiu.com.remotexs.ntu.edu.sg/n/about/

Arakelian, S. K. (n.d.). *Overall Coordination: Directorate General for European Programmes, Coordination and Development Supervision: Georghia Solomonidou-Christofidou Assistance: Evita Michaelides, Andri Christoforou Publisher: Press and Information Office*. 148.

Chaturvedi, M. (2021, June 8). Machine Learning Through The Lens of Econometrics. *Analytics India Magazine*. https://analyticsindiamag.com/machine-learning-through-the-lens-of-econometrics/

Cyprus .:. Sustainable Development Knowledge Platform. (n.d.). Retrieved 21 October 2021, from https://sustainabledevelopment.un.org/memberstates/cyprus

Cyprus country profile—SDGs and the environment—European Environment Agency. (n.d.). [Briefing]. Retrieved 15 October 2021, from https://www.eea.europa.eu/themes/sustainability-transitions/sustainable-development-goals-and-the/country-profiles/cyprus-country-profile-sdgs-and

Data tool content guide. (n.d.). Retrieved 23 October 2021, from http://graphics.eiu.com/data_services/contentguide/methodol.htm

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

NIEUWELING, C. (2020, July 6). *Employment in the energy sector: Trends and impact of the green energy transition* [Text]. EU Science Hub - European Commission. https://ec.europa.eu/jrc/en/science-update/employment-energy-sector

Oxford_model.pdf. (n.d.). Retrieved 20 October 2021, from https://www.oxfordeconomics.com/Media/Default/landing-pages/presentations/Oxford_model.pdf

PART 2: Machine Learning and Data Analysis. (n.d.). DATA DRIVEN SCIENCE & ENGINEERING. Retrieved 23 October 2021, from http://www.databookuw.com/page/databookuw.com/page/index.html

Romania country profile—SDGs and the environment—European Environment Agency. (n.d.). [Briefing]. Retrieved 14 October 2021, from https://www.eea.europa.eu/themes/sustainability-transitions/sustainable-development-goals-and-the/country-profiles/romania-country-profile-sdgs-and

Sustainable Development Report 2021. (n.d.). Retrieved 30 October 2021, from https://dashboards.sdgindex.org/rankings

Sweden and sustainability—Sweden.se. (n.d.). Retrieved 18 October 2021, from https://sweden.se/climate/sustainability/sweden-and-sustainability

These Are the EU's Richest and Poorest Countries—Bloomberg. (n.d.). Retrieved 23 October 2021, from https://www.bloomberg.com/news/articles/2021-06-30/europe-s-richest-live-where-pension-saving-starts-early-denmark

APPENDIX A

List of Variables and their Acronyms:

CO2

CO2Prod: Production-based CO2 intensity, energy-related CO2 per capita (Tonnes)

Energy

EnInt: Energy intensity, TPES per capita (Tonnes of oil equivalent (toe))
EnSup: Total primary energy supply (Tonnes of oil equivalent (toe), Millions)
ReEng1: Renewable energy supply, % total energy supply (%)
ReElect: Renewable electricity, % total electricity generation (%)
ReEng2: Renewable energy supply (excluding solid biofuels), % total energy supply (%)

Waste

Biomass: Biomass, % of DMC (%)

Biodiversity & Habitat

PestSal: Sales of pesticides per unit of agricultural land (Kg)

Air

PopPM: Mean population exposure to PM2.5 (micrograms per cubic metre)

Macro

RNDBud: Environmentally related government R&D budget, % total government R&D (%)
EnvTax: Environmentally related taxes, % GDP (%)
PopDens: Population density, inhabitants per km2

Others

GDPperCap: GDP per Capita (US$)

Dropped Columns

CO2Dem: Demand-based CO2 intensity, energy-related CO2 per capita (Tonnes)
CO2Tran: CO2 emissions from air transport per capita (Tonnes)
WasteGen: Municipal waste generated, kg per capita (missing 2019)
WasteRec: Municipal waste recycled or composted, % treated waste (%)
ReWater: Total renewable freshwater per capita (Cubic metres per capita)
BUAcap: Built up area per capita (Square metres)
ForestInt: Intensity of use of forest resource, Ratio
RegForest: Naturally regenerating forests, % total forest area (%)
TechDev: Development of environment-related technologies, % inventions worldwide (%)

APPENDIX B

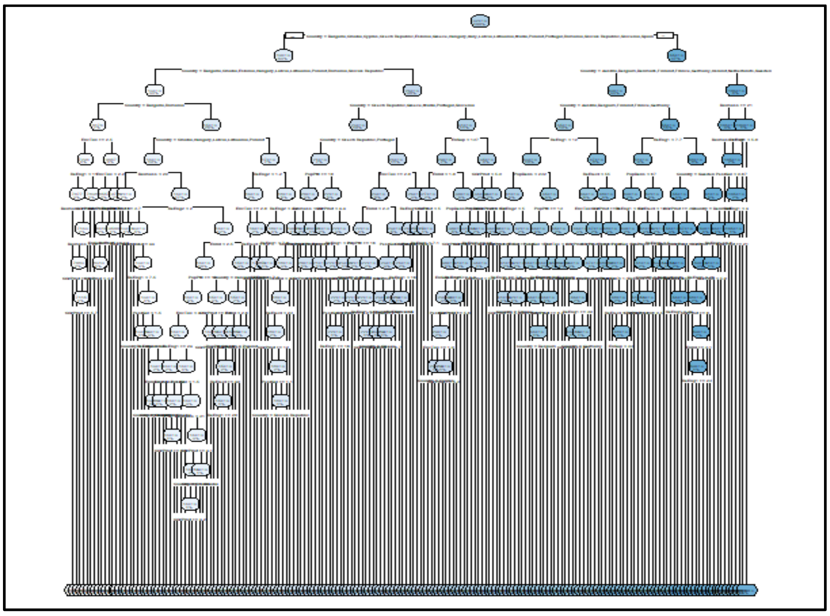


Figure 1.1: Grown Tree

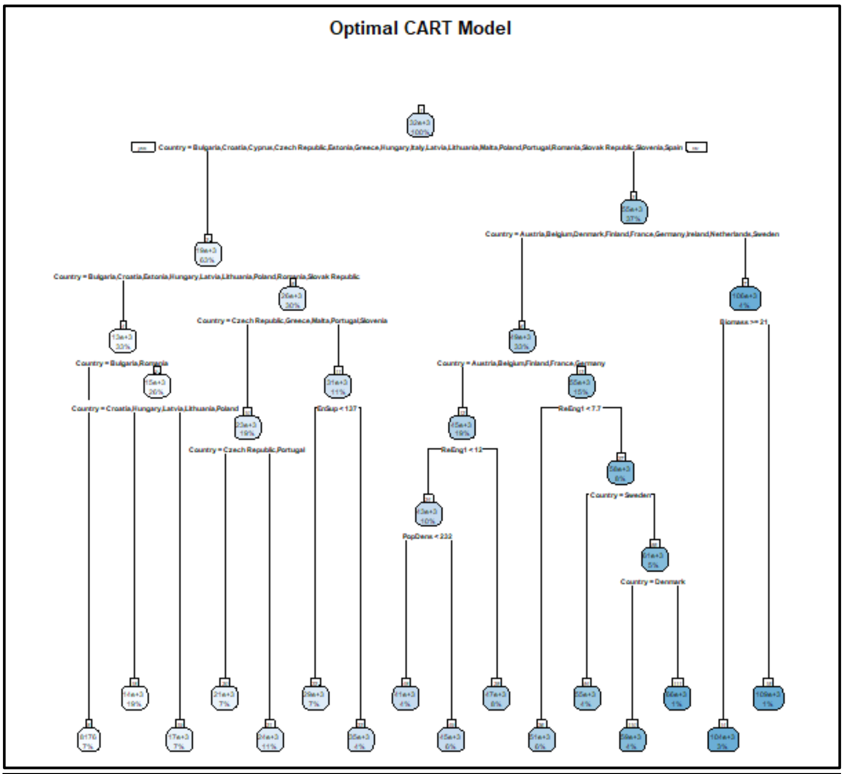


Figure 1.2: Optimal CART Model with Country as an indicator

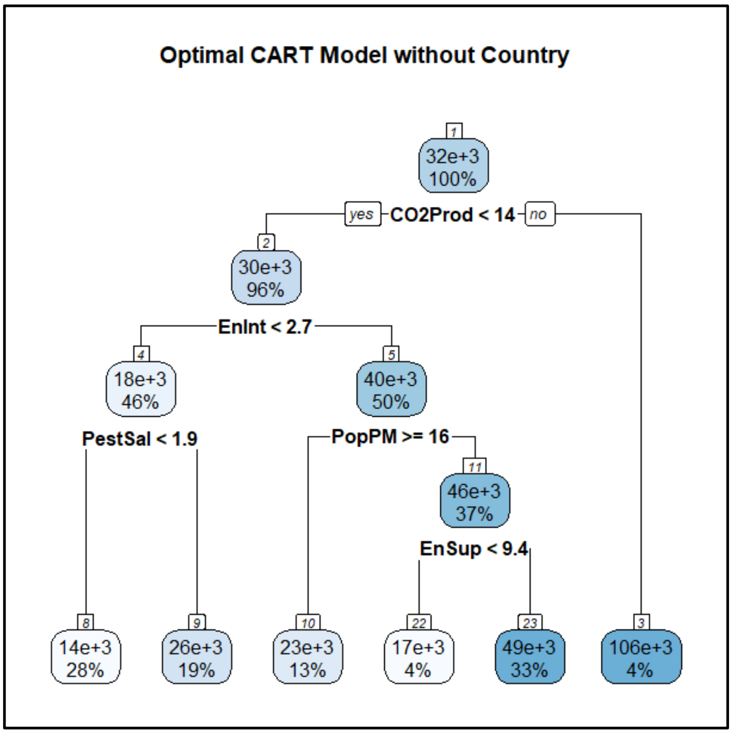


Figure 1.3: Optimal CART Model without Country as an indicator

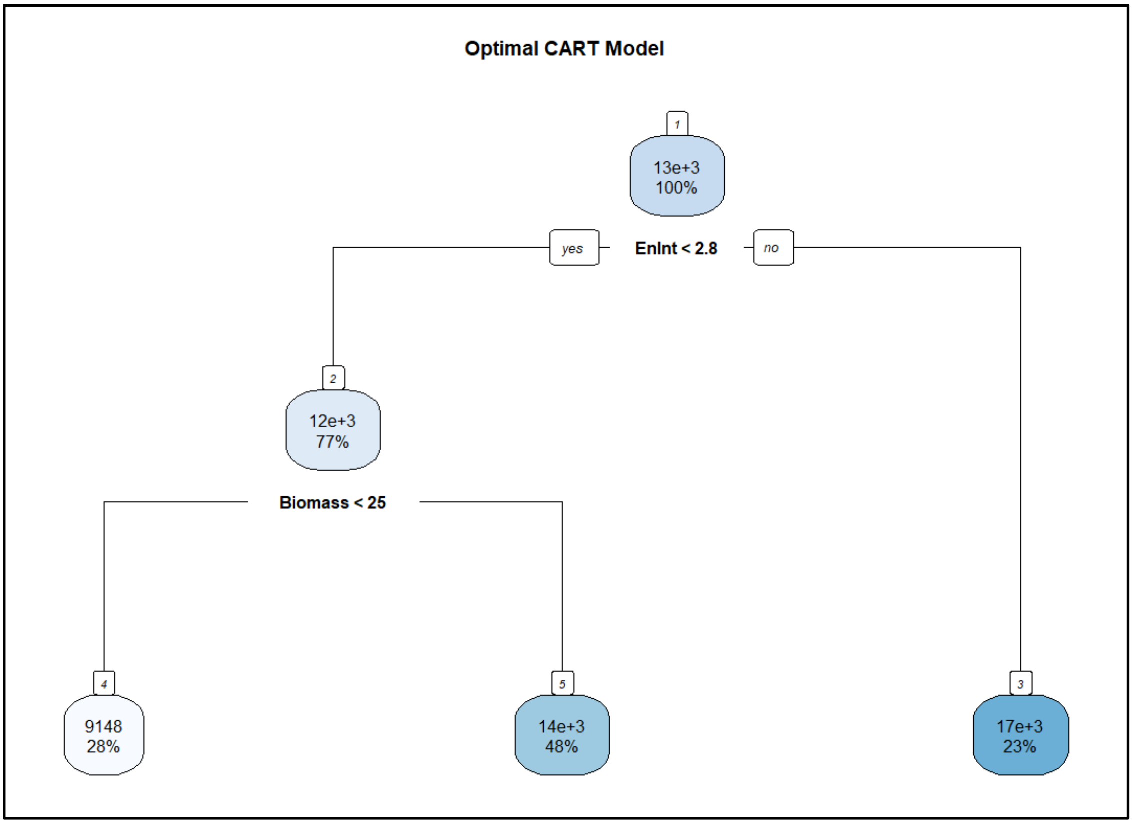
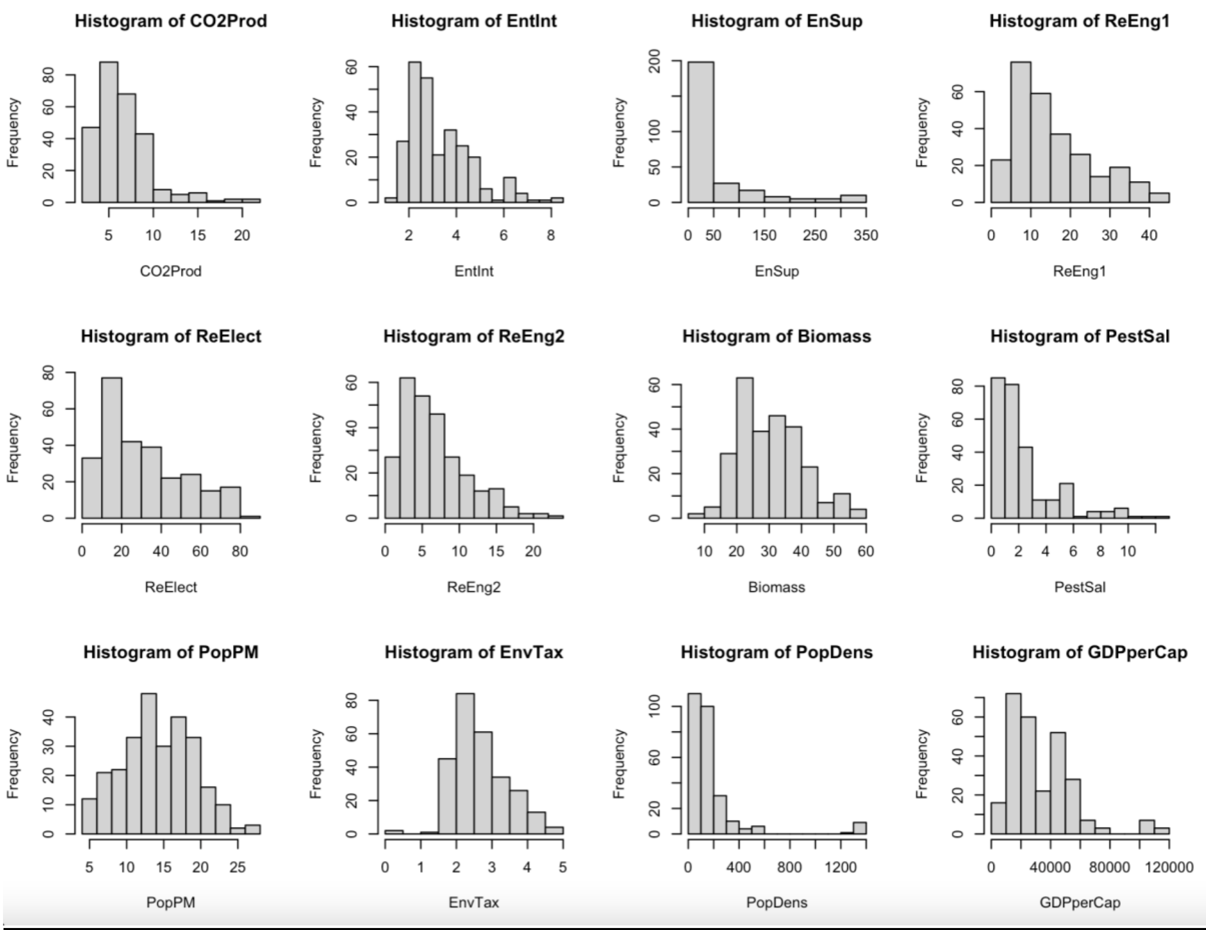


Figure 1.4: Optimal CART Model for "Low Capita" group

APPENDIX C

Histogram of all variables



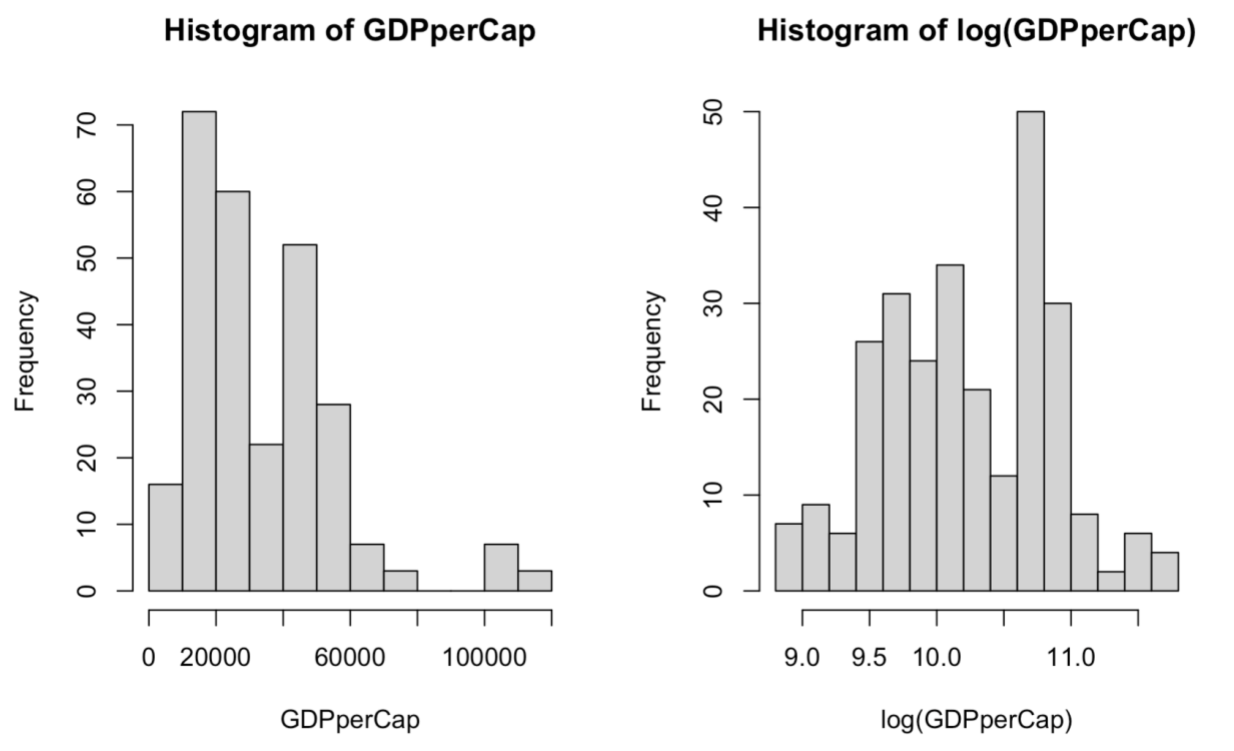
Skewness of variables before logarithmic transformation

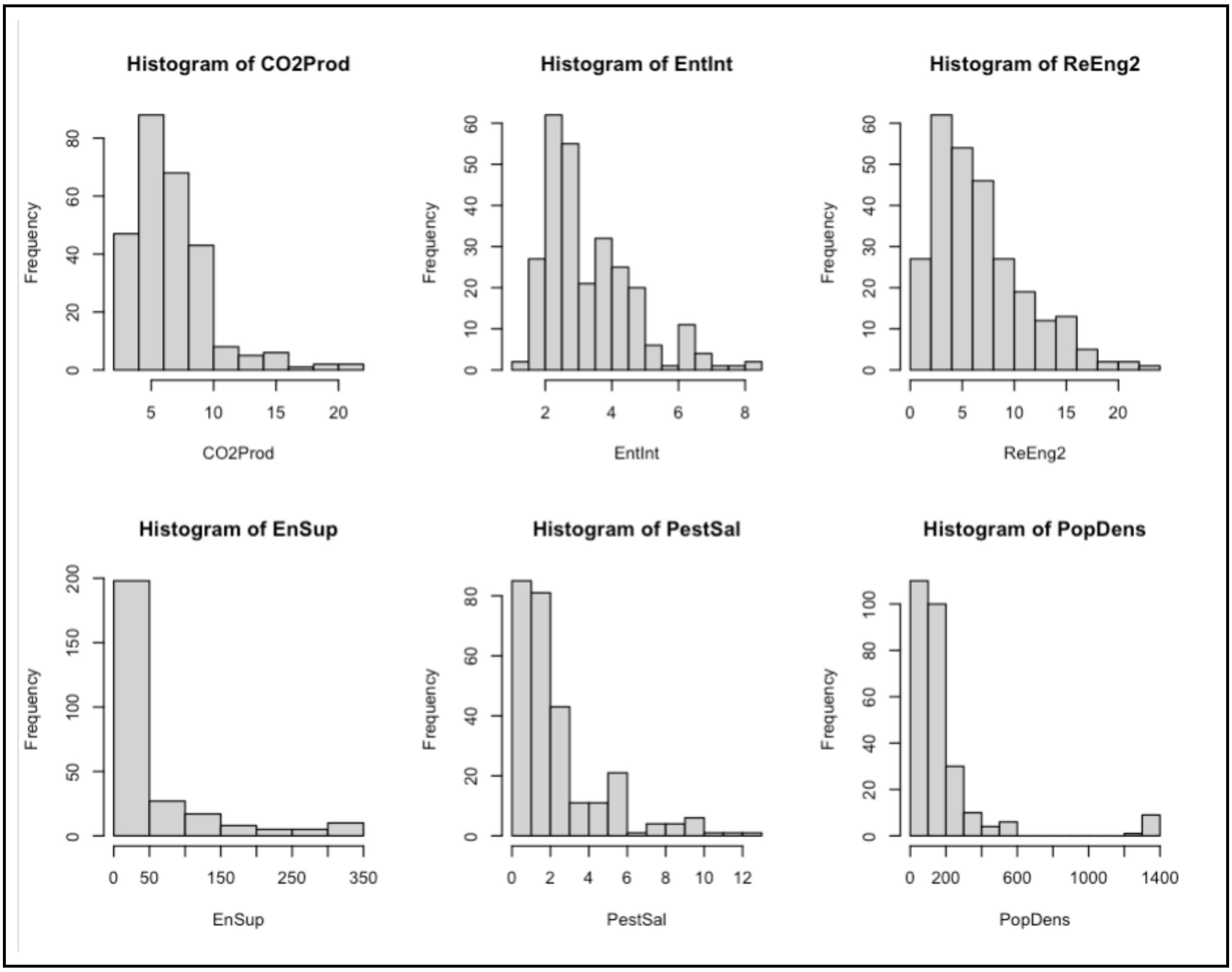
```
> sapply(dataset2[,3:14], function(x) ifelse(skewness(x) > 1 | skewness(x) < -1,
 skewness(x), NA))
  CO2Prod    EntInt     EnSup    ReEng1   ReElect    ReEng2   Biomass
 1.790067  1.158908  2.212798        NA        NA  1.065650        NA
  PestSal     PopPM    EnvTax   PopDens GDPperCap
 1.923122        NA        NA  3.737715  1.452018
> sapply(dataset2[,3:14], function(x) skewness(x))
  CO2Prod    EntInt     EnSup    ReEng1   ReElect    ReEng2   Biomass
1.7900672 1.1589084 2.2127980 0.9214792 0.6858368 1.0656502 0.4518886
  PestSal     PopPM    EnvTax   PopDens GDPperCap
1.9231222 0.1361381 0.4116548 3.7377154 1.4520178
```

Skewness of variables after logarithmic transformation

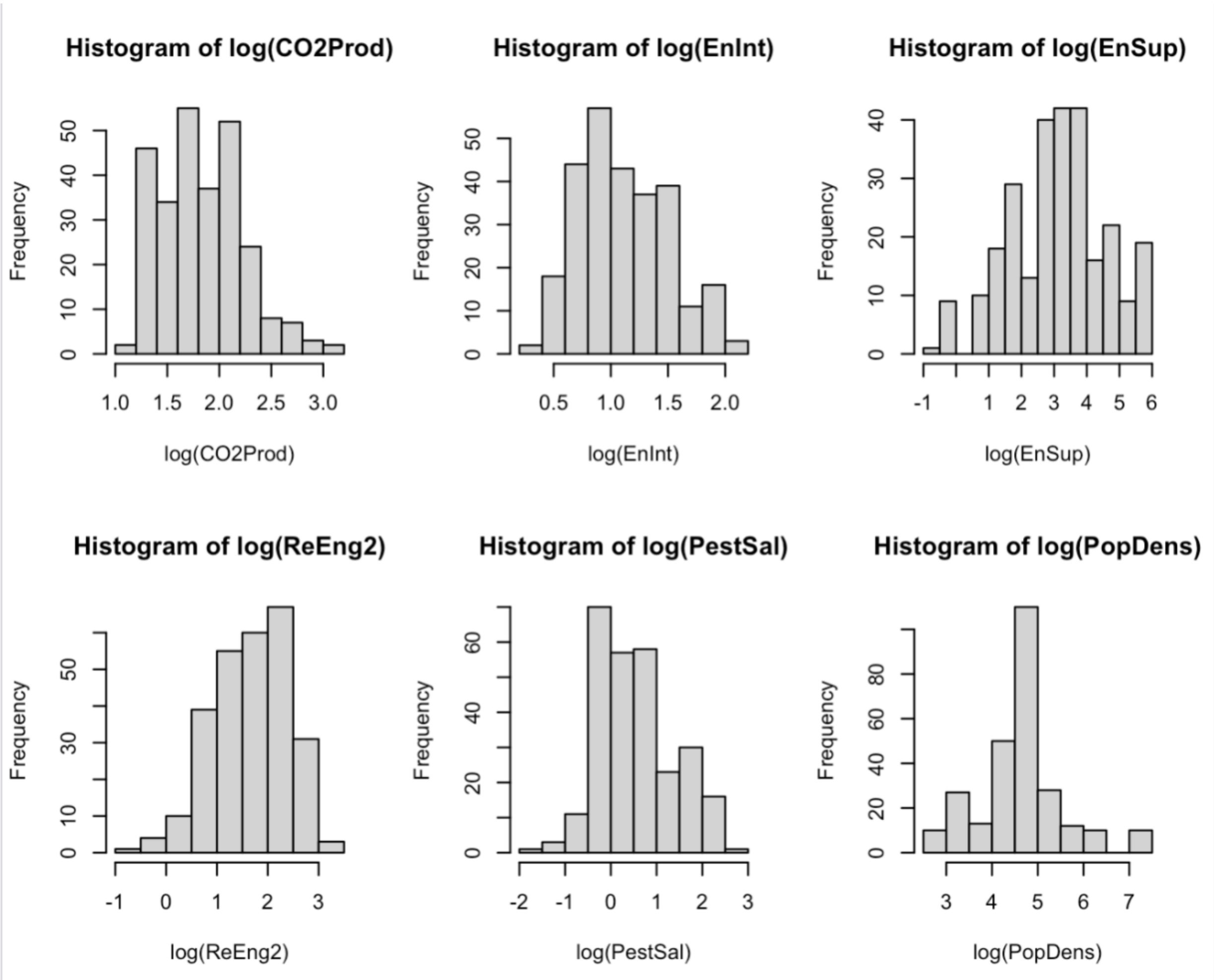
```
> sapply(logdataset[,3:14], function(x) ifelse(skewness(x) > 1 | skewness(x) < -
1, skewness(x), NA))
  CO2Prod     EnInt     EnSup    ReEng1   ReElect    ReEng2   Biomass
       NA        NA        NA        NA        NA        NA        NA
  PestSal     PopPM    EnvTax   PopDens GDPperCap
       NA        NA        NA        NA        NA
> sapply(logdataset[,3:14], function(x) skewness(x))
    CO2Prod       EnInt       EnSup      ReEng1     ReElect      ReEng2
 0.49740111  0.34354465 -0.23400138  0.92147919  0.68583683 -0.39040114
    Biomass     PestSal       PopPM      EnvTax     PopDens   GDPperCap
 0.45188860  0.34933259  0.13613812  0.41165476  0.46895608 -0.01440577
```

Histogram screenshots after logarithmic transformation

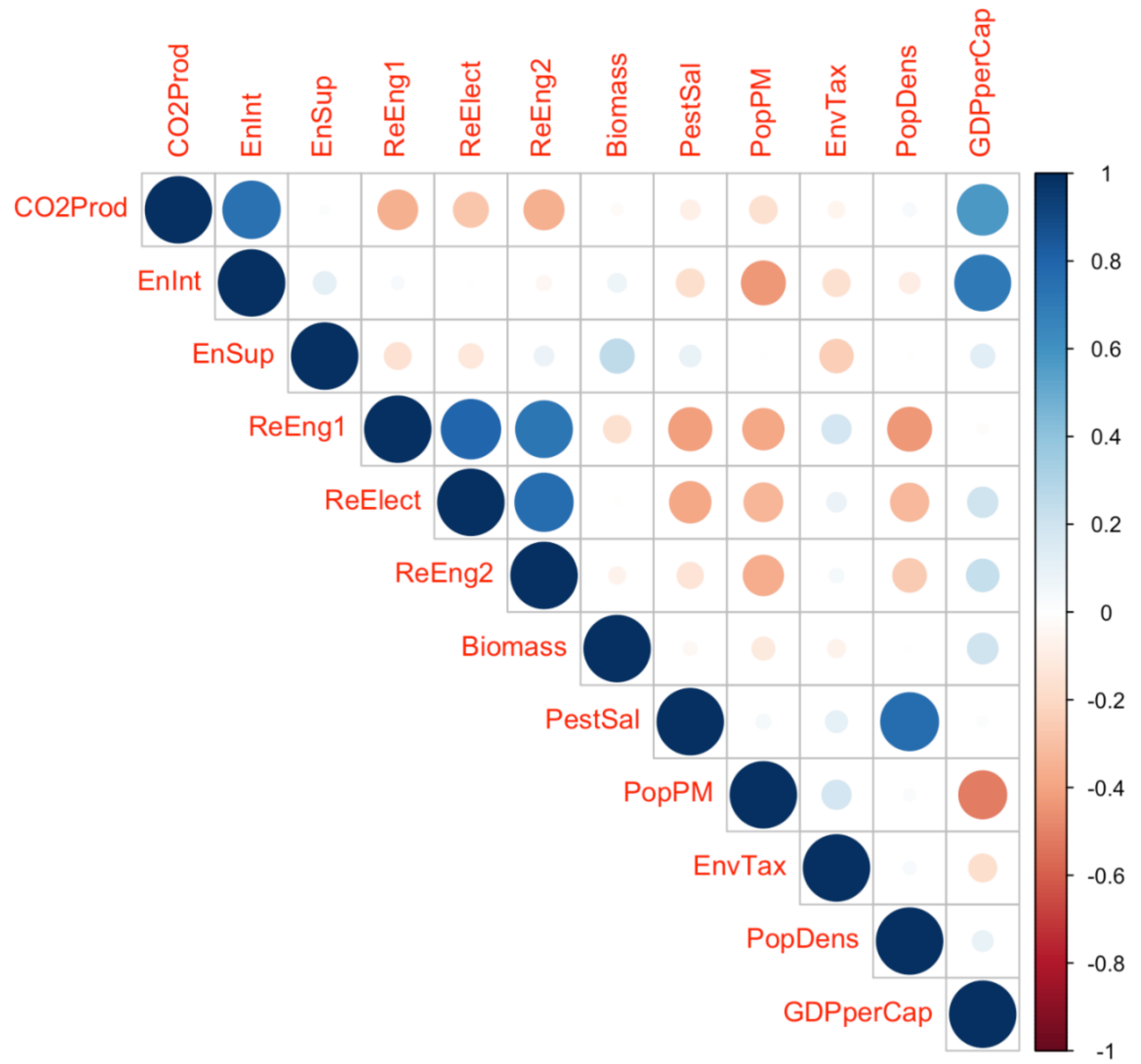




Histogram of CO2Prod
Frequency
CO2Prod
Histogram of EntInt
Frequency
EntInt
Histogram of ReEng2
Frequency
ReEng2
Histogram of EnSup
Frequency
EnSup
Histogram of PestSal
Frequency
PestSal
Histogram of PopDens
Frequency
PopDens



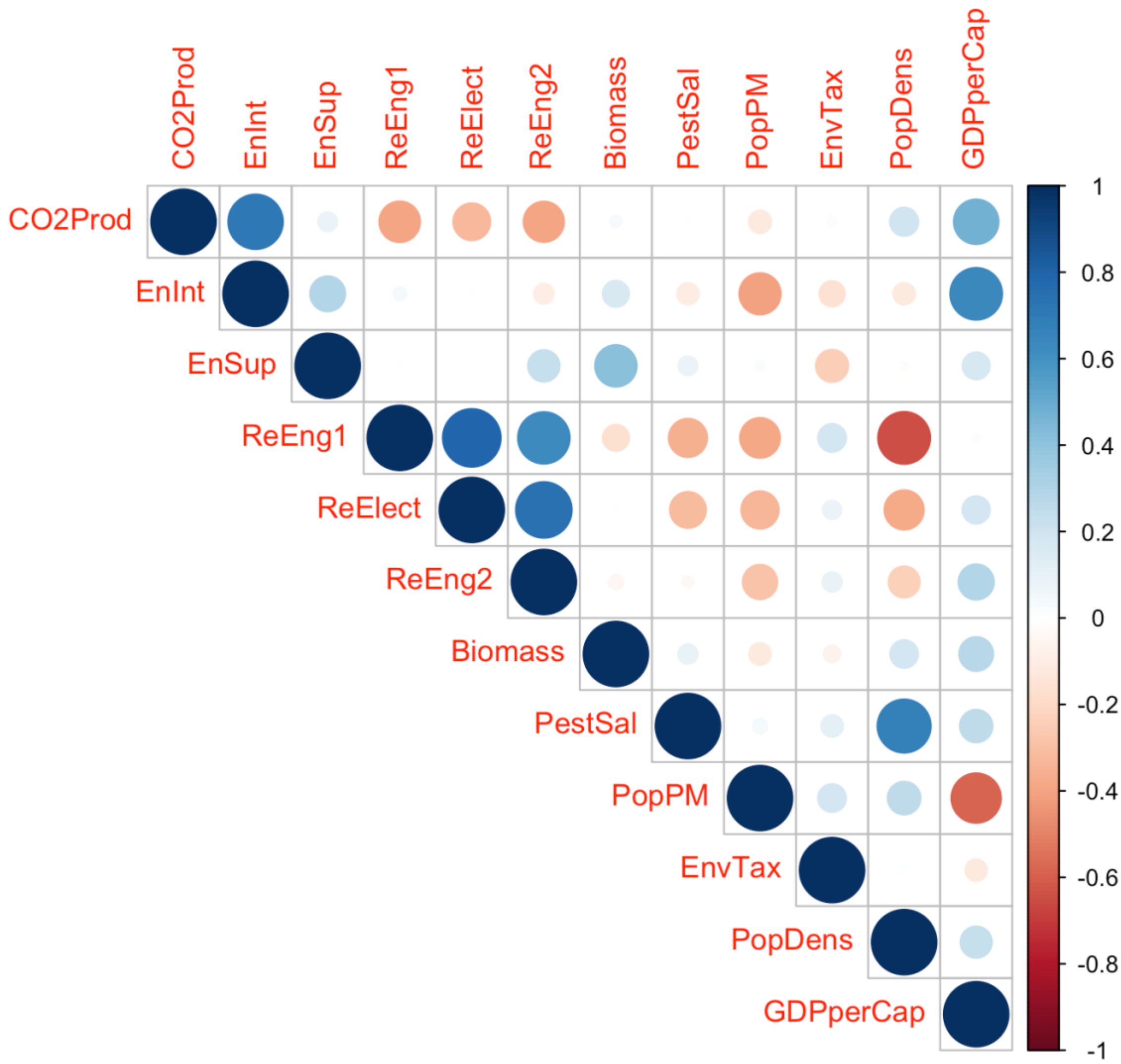
Correlation plot before logarithmic transformation



Correlation of all variables against GDPperCap	`> cor(dataset2[,3:14])[,12] # correlation of all variables before log against GDP percap`

```
     CO2Prod       EnInt       EnSup      ReEng1     ReElect      ReEng2
  0.57975370  0.70321843  0.12430459 -0.01979761  0.20257858  0.23440608
     Biomass     PestSal       PopPM      EnvTax     PopDens   GDPperCap
  0.20734093  0.01990438 -0.51933032 -0.17141252  0.09202393  1.00000000
```

Correlation plot after logarithmic transformation of required variables



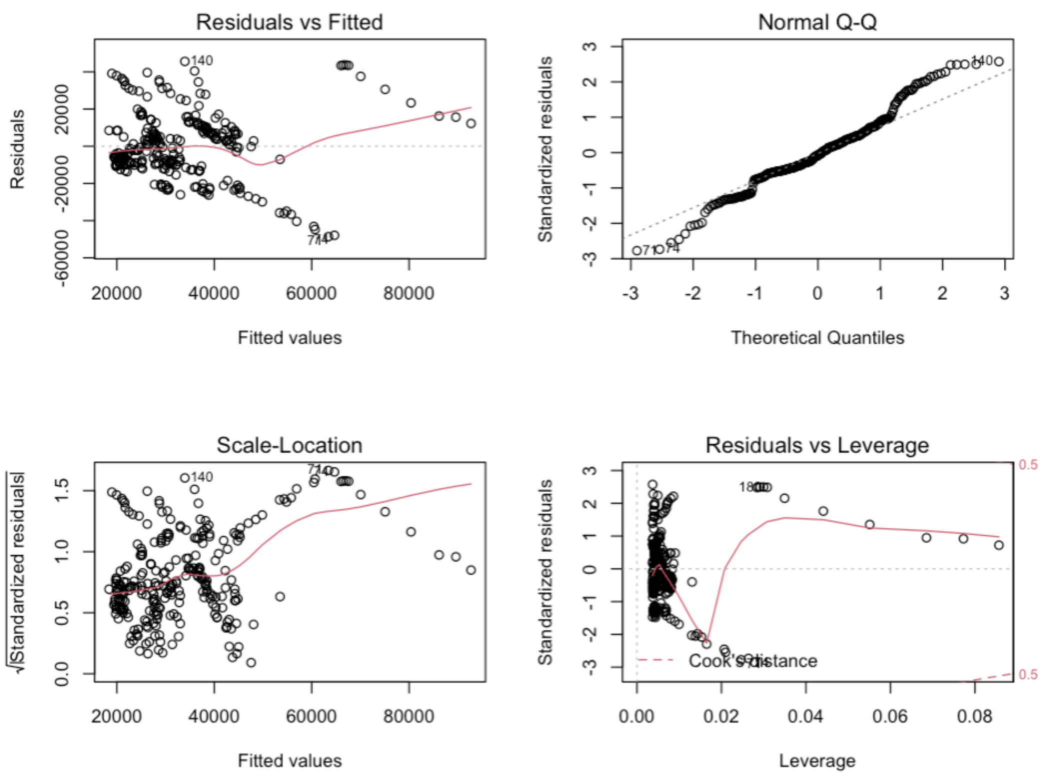
Correlation of all variables + transformed variables against GDPperCap

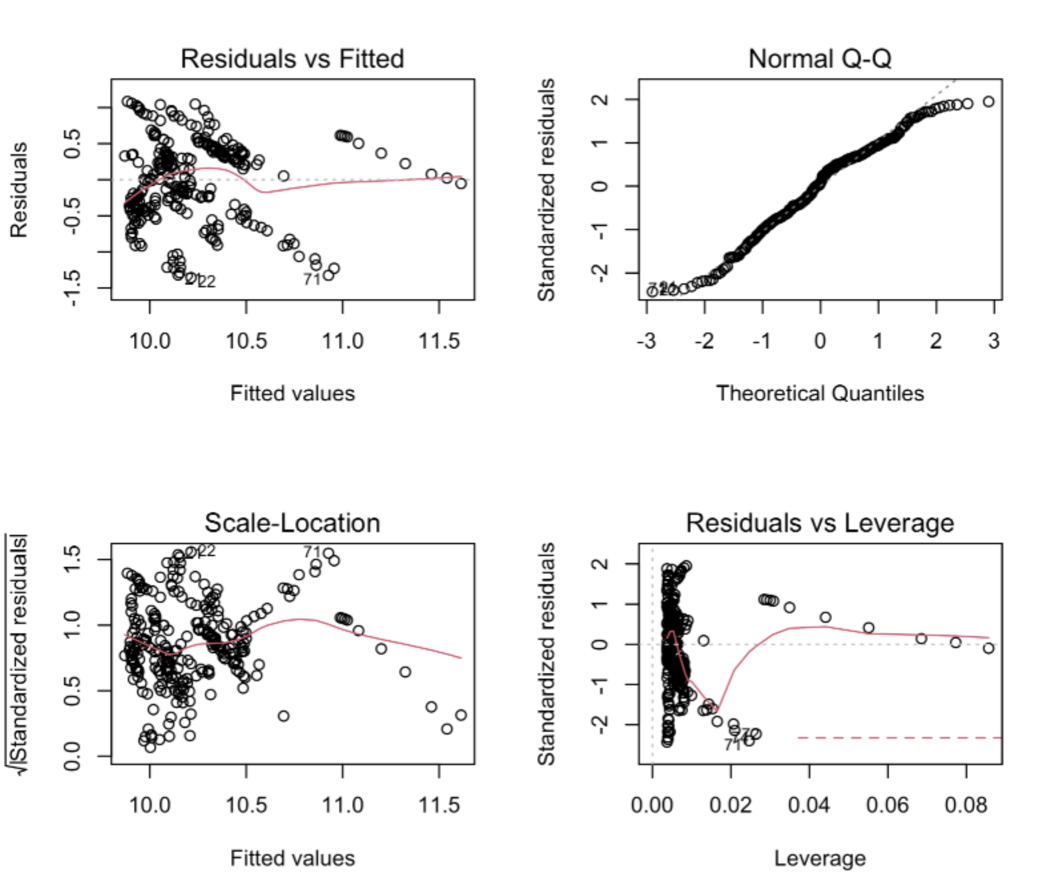
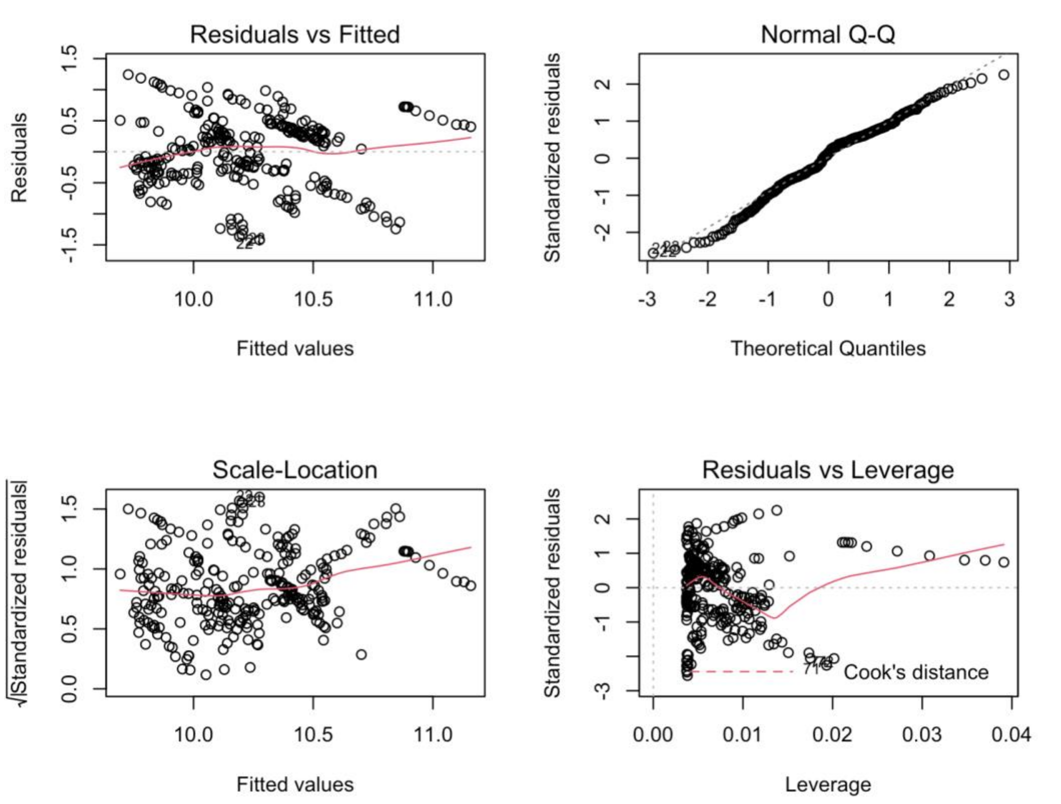
```
> cor(logdataset[,3:14])[,12] # correlation of all variables after log against GD
Ppercap
    CO2Prod       EnInt       EnSup      ReEng1     ReElect      ReEng2
 0.47866298  0.64343058  0.17597474  0.01195181  0.18055583  0.29956715
    Biomass     PestSal       PopPM      EnvTax     PopDens   GDPperCap
 0.27741678  0.25524368 -0.58748701 -0.11155091  0.23546465  1.00000000
```

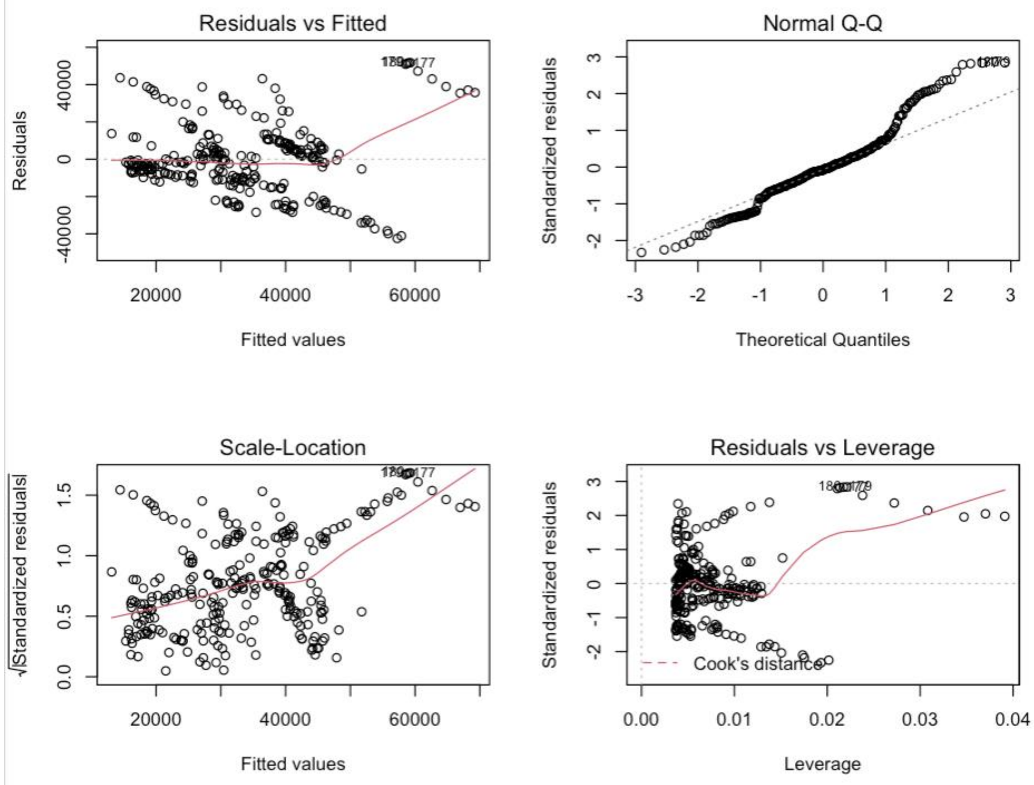
Variance Inflation Factor

```
> vif(VifReg)
log(CO2Prod)   log(EnInt)   log(EnSup)       ReEng1      ReElect  log(ReEng2)
    4.595917     3.970464     2.377785     8.024271     5.832085     3.716253
     Biomass log(PestSal)        PopPM       EnvTax log(PopDens)
    1.896571     2.678826     2.188162     1.774183     3.841035
```

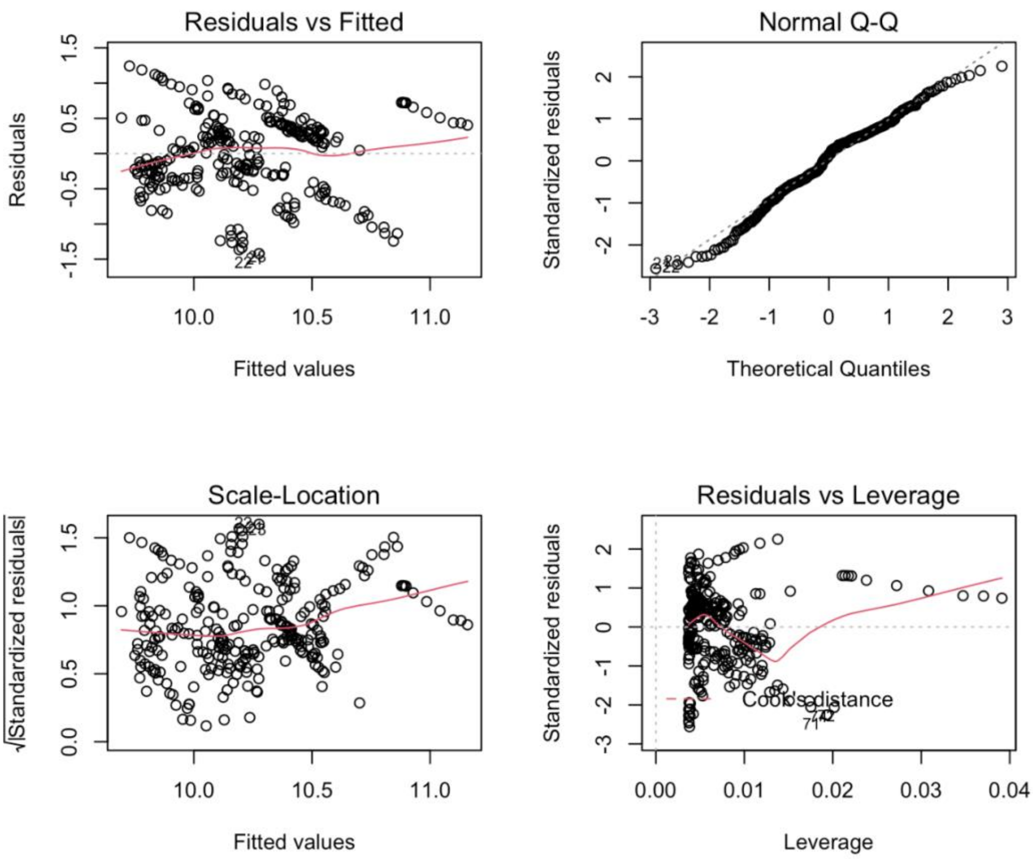
Model diagnostics plot of model relationship between GDP and CO2Prod

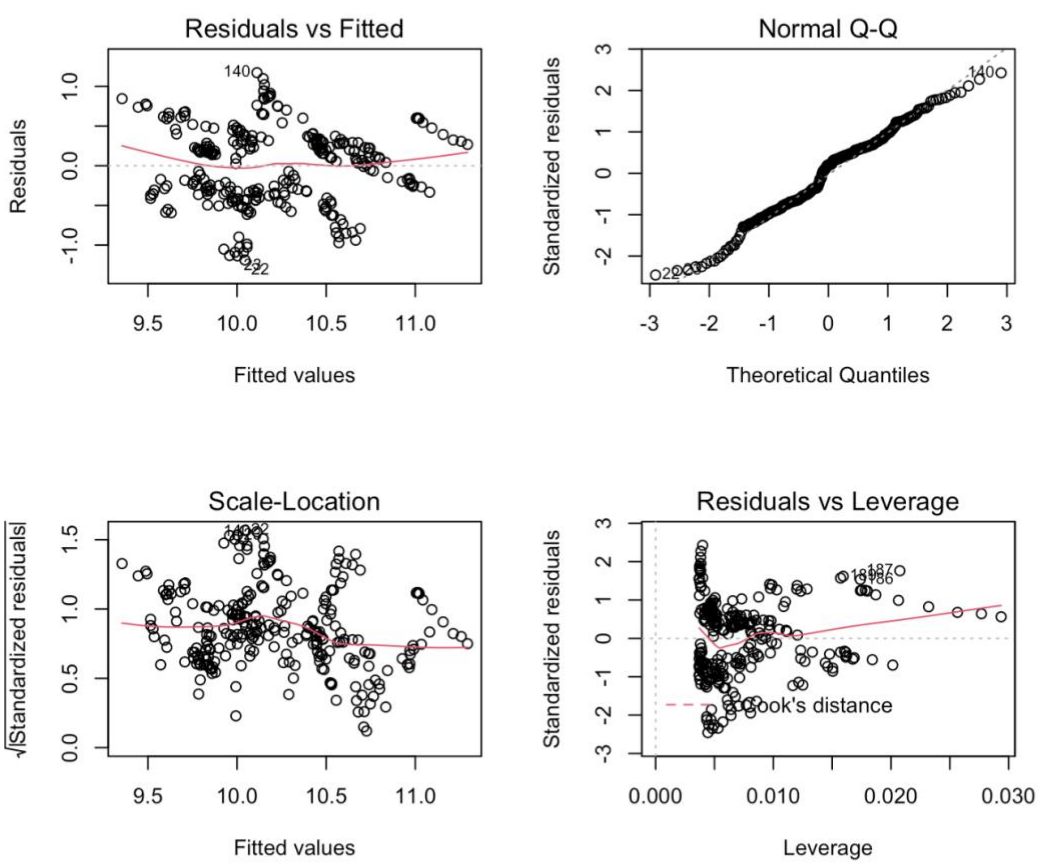
Variables	Model Diagnostics Plot
GDP against CO2Prod	

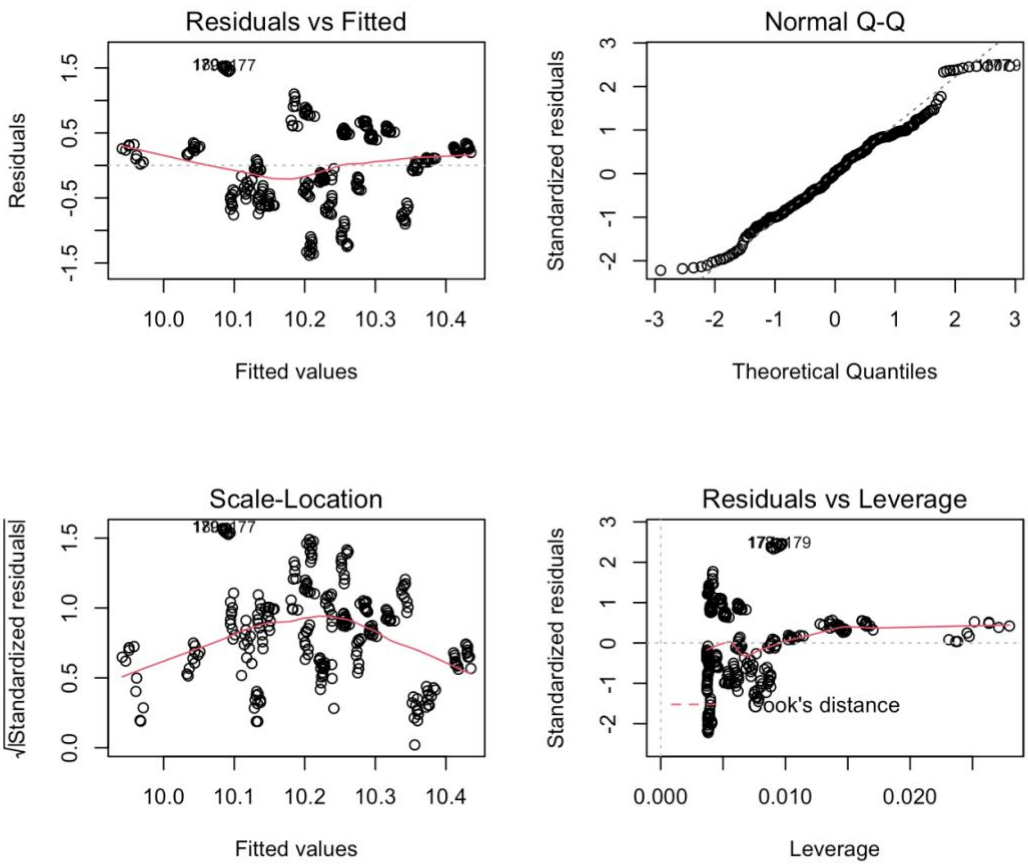
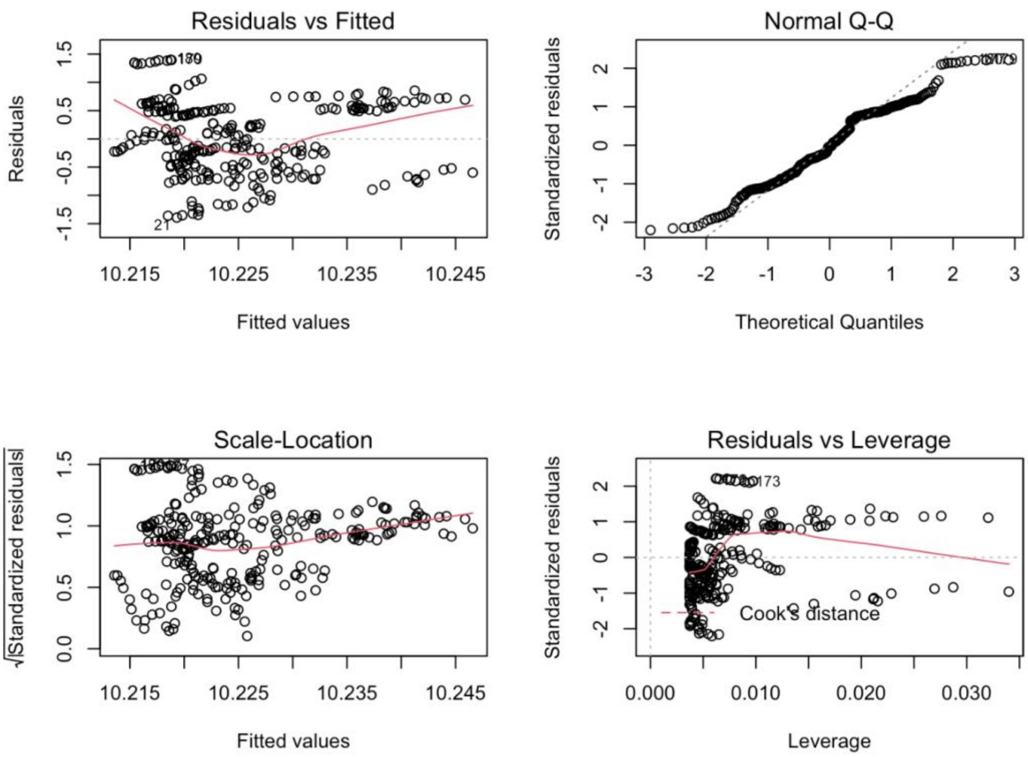
log GDP against CO2Prod	
log GDP against log(CO2Prod)	

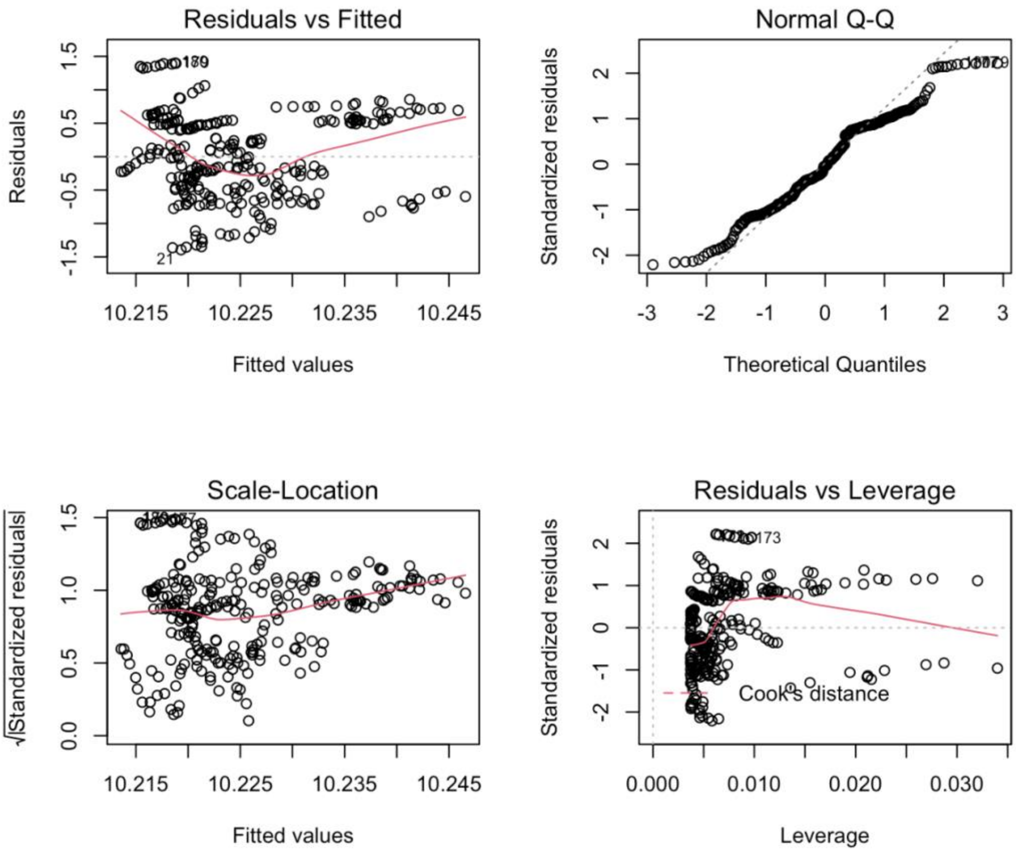
GDP against log CO2Prod	

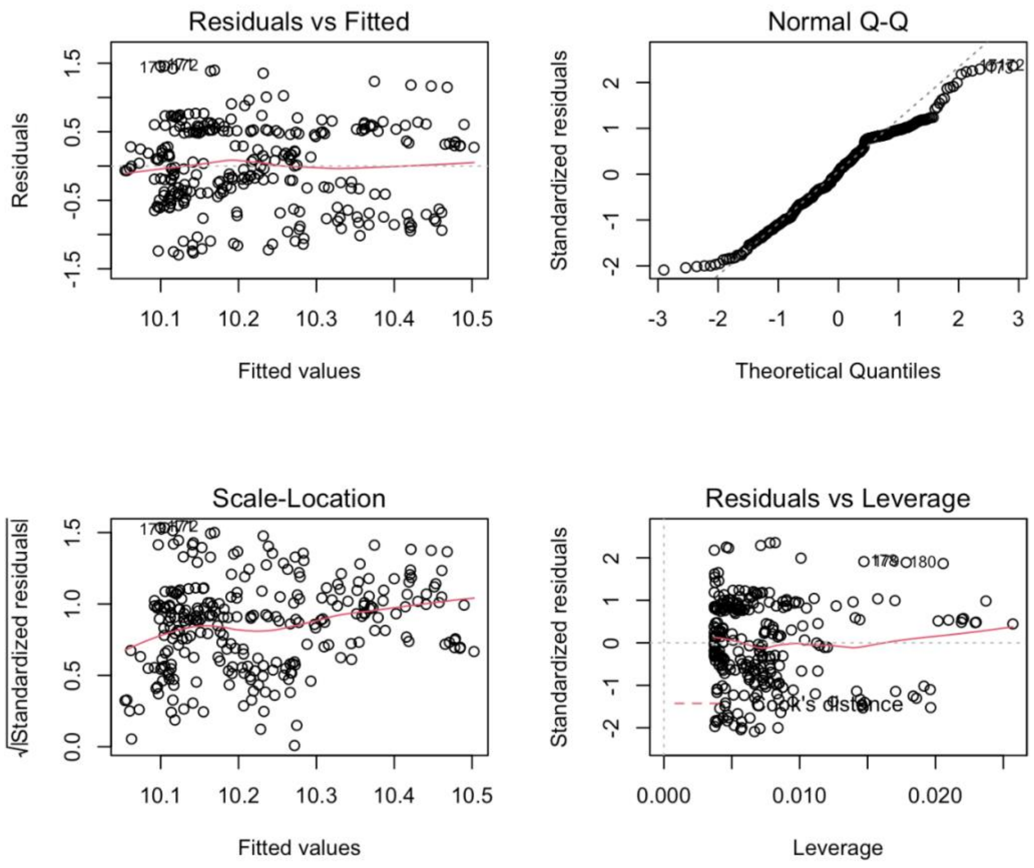
Final Analysis of Model Diagnostic Plots

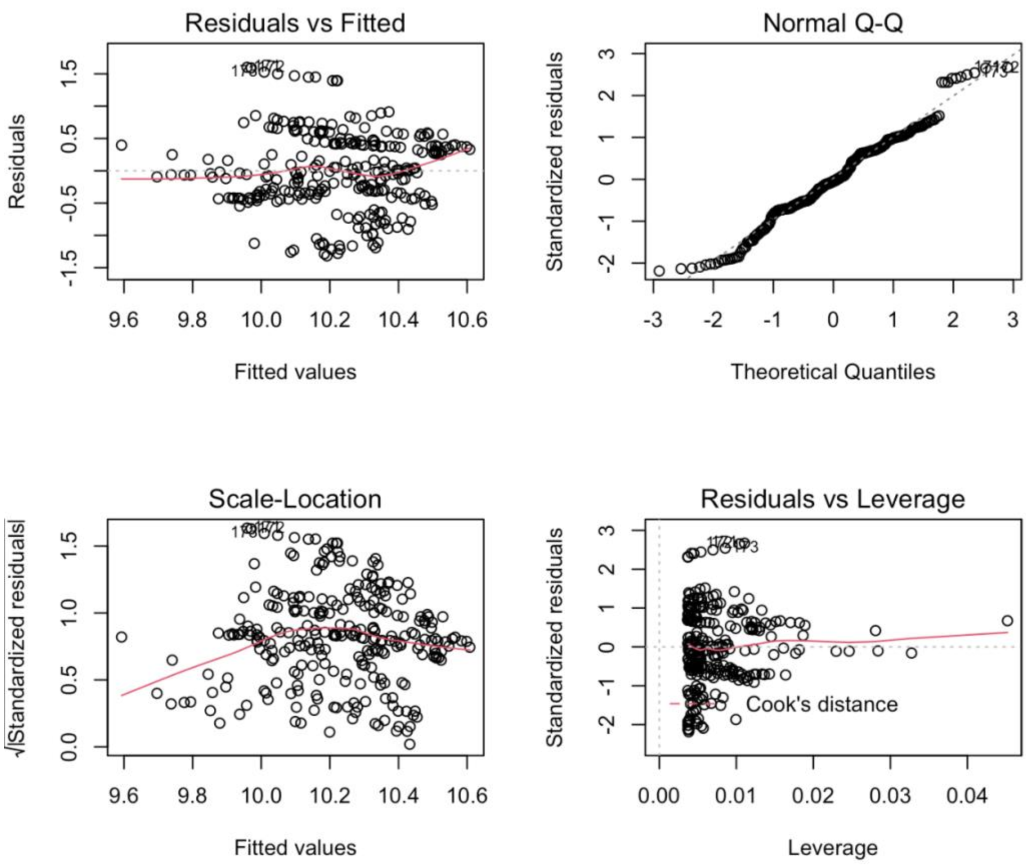
Variables	Model Diagnostics Plot
GDP against CO2Prod	

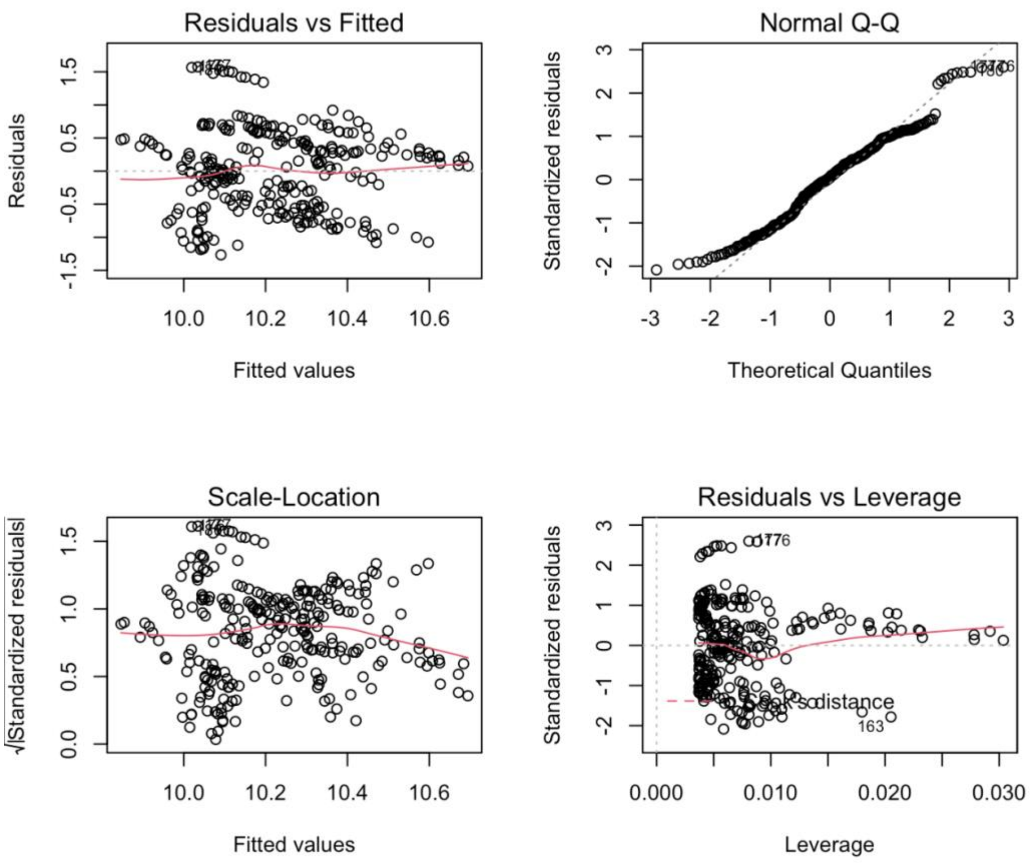
GDP against EnInt	

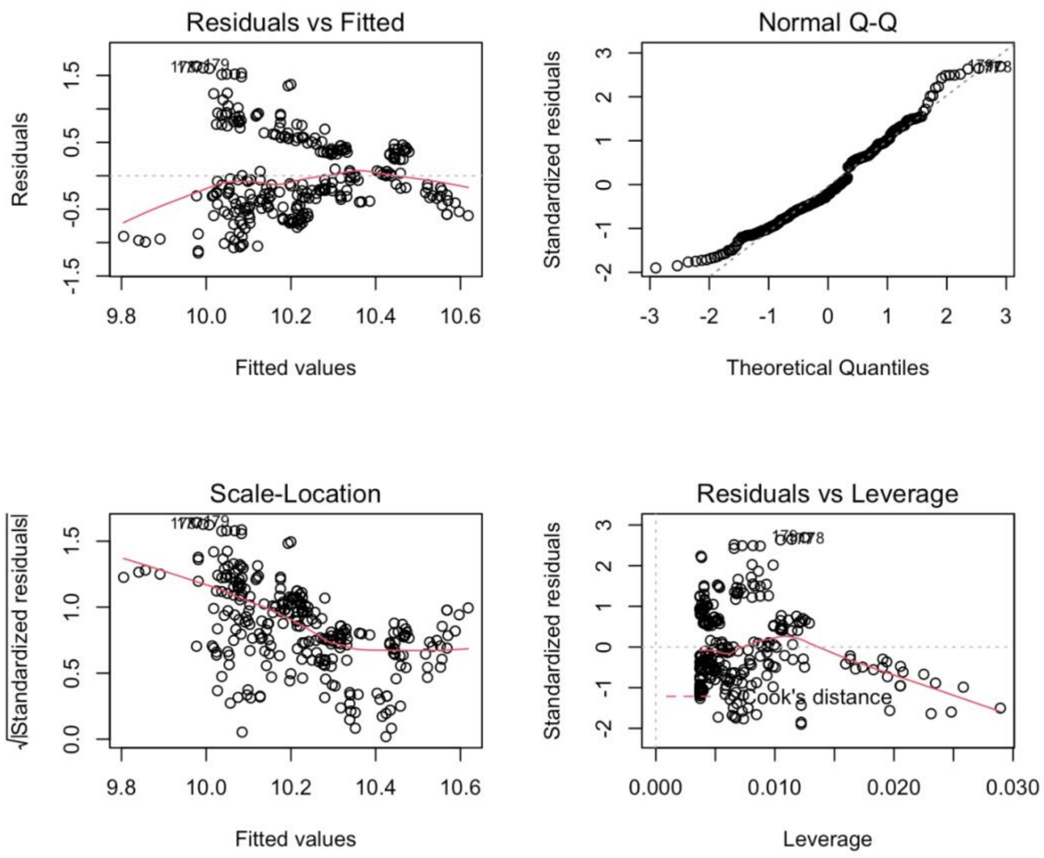
GDP against EnSup	
GDP against ReEng1	

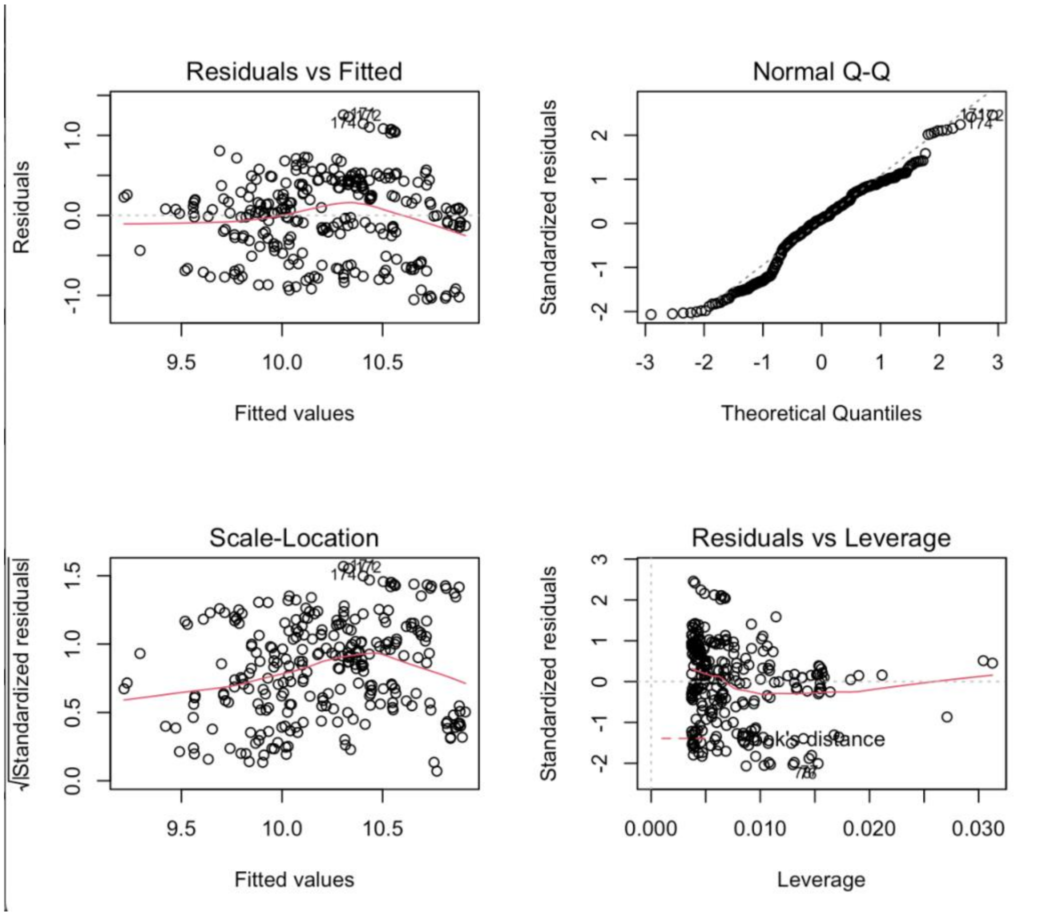
GDP against ReEng1	

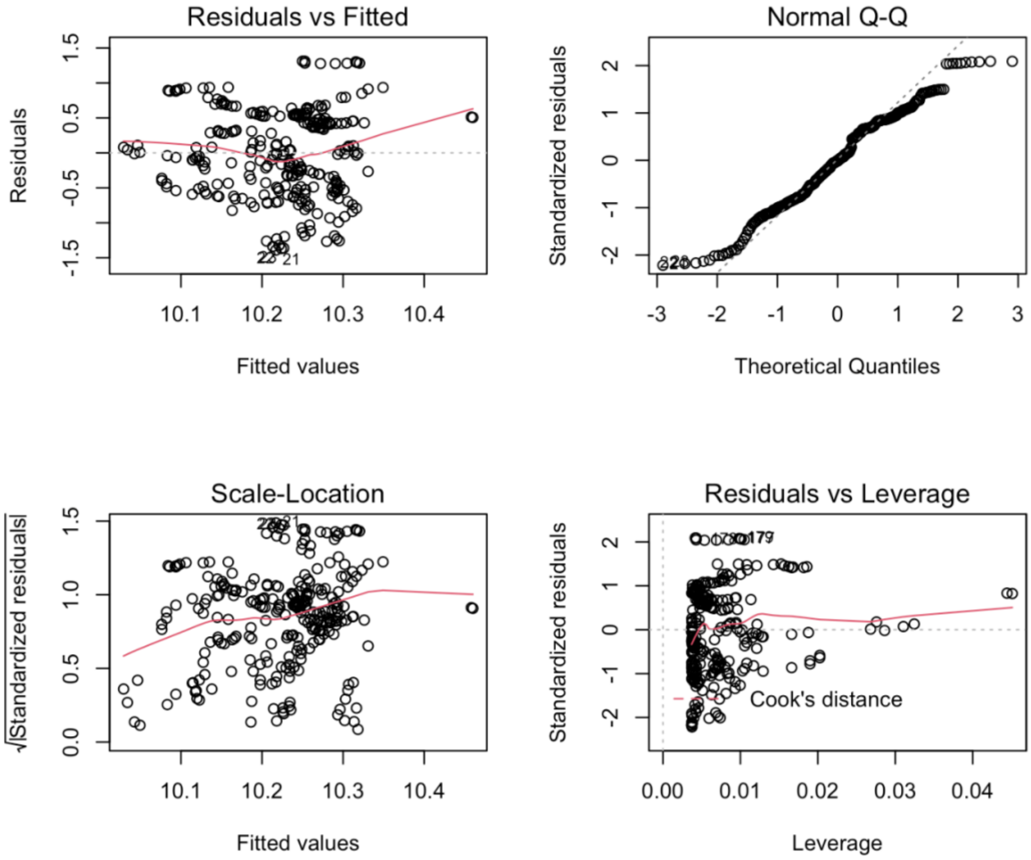
GDP against ReElect	

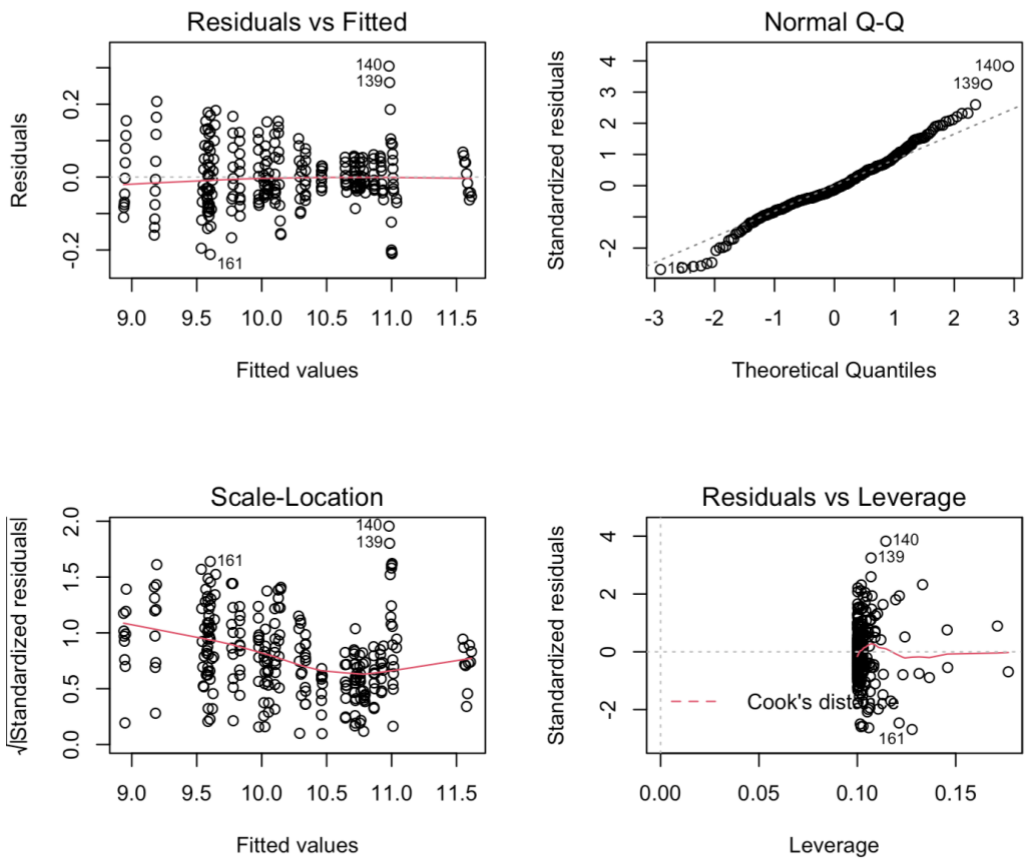
GDP against ReEng2	

GDP against Biomass	

GDP against PestSal	

GDP against PopPm	

GDP against EnvTax	

GDP against PopDens	

APPENDIX D

Regression Model

RegTrainSet:

Regression on trainset, factored country

```
Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(EnSup) +
    ReEng1 + ReElect + log(ReEng2) + Biomass + log(PestSal) +
    PopPM + EnvTax + log(PopDens) + Country, data = trainset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.077210 -0.020306 -0.002957  0.018763  0.136364

Coefficients:
                        Estimate Std. Error t value             Pr(>|t|)
(Intercept)           15.4936800  1.5446300  10.031 < 0.0000000000000002 ***
log(CO2Prod)          -0.2432251  0.0918823  -2.647             0.008979 **
log(EnInt)             0.7795463  0.6164758   1.265             0.207992
log(EnSup)            -0.2642207  0.6128559  -0.431             0.666989
ReEng1                 0.0027639  0.0035325   0.782             0.435197
ReElect               -0.0002438  0.0008522  -0.286             0.775189
log(ReEng2)            0.0880422  0.0267778   3.288             0.001256 **
Biomass               -0.0031290  0.0011149  -2.807             0.005667 **
log(PestSal)           0.0031934  0.0171489   0.186             0.852527
PopPM                 -0.0132057  0.0033433  -3.950             0.000120 ***
EnvTax                -0.0994224  0.0188578  -5.272     0.00000045853125 ***
log(PopDens)          -0.8904946  0.5316439  -1.675             0.096007 .

Residual standard error: 0.03881 on 155 degrees of freedom
Multiple R-squared:  0.997,     Adjusted R-squared:  0.9963
F-statistic:  1551 on 33 and 155 DF,  p-value: < 0.00000000000000022
```

RegTrainSet

Backward Elimination

```
Step:  AIC=-1197.63
log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(ReEng2) + Biomass +
    PopPM + EnvTax + log(PopDens) + Country

               Df Sum of Sq     RSS     AIC
<none>                       0.2335 -1197.6
- Biomass        1    0.0116  0.2451 -1190.5
- log(CO2Prod)   1    0.0144  0.2479 -1188.3
- log(ReEng2)    1    0.0214  0.2549 -1183.0
- log(EnInt)     1    0.0241  0.2575 -1181.1
- log(PopDens)   1    0.0294  0.2629 -1177.2
- PopPM          1    0.0306  0.2640 -1176.4
- EnvTax         1    0.0412  0.2747 -1168.9
- Country       26   12.0194 12.2528  -501.1

Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(ReEng2) +
    Biomass + PopPM + EnvTax + log(PopDens) + Country, data = trainset)
```

RegTrainSet2:

Regression on trainset, factored country after backward elimination

```
Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(ReEng2) +
    Biomass + PopPM + EnvTax + log(PopDens) + Country, data = trainset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.078214 -0.021190 -0.002164  0.019948  0.139285

Coefficients:
                        Estimate Std. Error t value             Pr(>|t|)
(Intercept)            15.963748   1.273608  12.534 < 0.0000000000000002 ***
log(CO2Prod)           -0.269222   0.087001  -3.094             0.002340 **
log(EnInt)              0.542080   0.135599   3.998  0.00009869909327403 ***
log(ReEng2)             0.088641   0.023490   3.774             0.000229 ***
Biomass                -0.002977   0.001074  -2.773             0.006235 **
PopPM                  -0.014261   0.003166  -4.505  0.00001299258644557 ***
EnvTax                 -0.096026   0.018361  -5.230  0.00000054173553624 ***
log(PopDens)           -1.096571   0.248220  -4.418  0.00001864079850787 ***
CountryBelgium          1.377171   0.324507   4.244  0.00003768851354117 ***
CountryBulgaria        -1.944028   0.149845 -12.974 < 0.0000000000000002 ***
CountryCroatia         -1.135408   0.122496  -9.269 < 0.0000000000000002 ***
CountryCyprus           0.112185   0.083652   1.341             0.181854
CountryCzech Republic  -0.290107   0.073340  -3.956             0.000116 ***
CountryDenmark          0.796564   0.081588   9.763 < 0.0000000000000002 ***
CountryEstonia         -2.148423   0.315545  -6.809  0.00000000020391853 ***
CountryFinland         -2.175253   0.415254  -5.238  0.00000052097773761 ***
CountryFrance          -0.057412   0.065924  -0.871             0.385169
CountryGermany          0.890661   0.196734   4.527  0.00001184532888443 ***
CountryGreece          -0.499692   0.103476  -4.829  0.00000326409432158 ***
CountryHungary         -0.812946   0.071487 -11.372 < 0.0000000000000002 ***
CountryIreland         -0.135171   0.141616  -0.954             0.341321
CountryItaly            0.744832   0.159326   4.675  0.00000634921749637 ***
CountryLatvia          -2.267567   0.319824  -7.090  0.00000000004457863 ***
CountryLithuania       -1.859242   0.240379  -7.735  0.00000000000123317 ***
CountryLuxembourg       1.583385   0.219807   7.204  0.00000000002394882 ***
CountryMalta            2.631505   0.590203   4.459  0.00001574992276107 ***
CountryNetherlands      2.064576   0.404430   5.105  0.00000095786923606 ***
CountryPoland          -0.473529   0.071011  -6.668  0.00000000042999222 ***
CountryPortugal        -0.582400   0.058926  -9.884 < 0.0000000000000002 ***
CountryRomania         -1.527272   0.125609 -12.159 < 0.0000000000000002 ***
CountrySlovak Republic -0.670053   0.046231 -14.494 < 0.0000000000000002 ***
CountrySlovenia        -0.394529   0.045206  -8.727  0.00000000000000387 ***
CountrySpain           -0.490308   0.070635  -6.941  0.00000000009985158 ***
CountrySweden          -1.933801   0.361391  -5.351  0.00000030919615275 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03881 on 155 degrees of freedom
Multiple R-squared:  0.997,     Adjusted R-squared:  0.9963
F-statistic:  1551 on 33 and 155 DF,  p-value: < 0.00000000000000022
```

RMSE for RegTrainset2

```
> RMSE.RegTrainSet2.train
[1] 0.0351469
> RMSE.RegTrainSet2.test
[1] 4131.066
```

RegTrainSetB

Regression on transit without factored country

```
Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(EnSup) +
    ReEng1 + ReElect + log(ReEng2) + Biomass + log(PestSal) +
    PopPM + EnvTax + log(PopDens), data = trainset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57353 -0.17980  0.02536  0.15629  0.57335

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)   7.475802   0.265795  28.126 < 0.0000000000000002 ***
log(CO2Prod)  0.430618   0.092536   4.654        0.00000637344 ***
log(EnInt)    0.822134   0.087040   9.445 < 0.0000000000000002 ***
log(EnSup)   -0.116510   0.018645  -6.249        0.00000000299 ***
ReEng1       -0.002254   0.005946  -0.379             0.705139
ReElect      -0.004732   0.002644  -1.790             0.075175 .
log(ReEng2)   0.562738   0.049069  11.468 < 0.0000000000000002 ***
Biomass       0.015505   0.002367   6.551        0.00000000060 ***
log(PestSal)  0.127576   0.032527   3.922             0.000125 ***
PopPM        -0.031033   0.005169  -6.004        0.00000001066 ***
EnvTax       -0.067840   0.034297  -1.978             0.049480 *
log(PopDens)  0.158200   0.037604   4.207        0.00004106889 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2283 on 177 degrees of freedom
Multiple R-squared:  0.8807,    Adjusted R-squared:  0.8733
F-statistic: 118.8 on 11 and 177 DF,  p-value: < 0.00000000000000022
```

RegTrainSetB Backward Elimination

```
Step:  AIC=-548.59
log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(EnSup) + ReElect +
    log(ReEng2) + Biomass + log(PestSal) + PopPM + EnvTax + log(PopDens)

               Df Sum of Sq     RSS     AIC
<none>                       9.2329 -548.59
- EnvTax        1    0.3359  9.5688 -543.83
- ReElect       1    0.5909  9.8238 -538.86
- log(PestSal)  1    0.7945 10.0275 -534.98
- log(PopDens)  1    1.3880 10.6210 -524.12
- log(CO2Prod)  1    1.6184 10.8514 -520.06
- PopPM         1    2.0336 11.2666 -512.96
- log(EnSup)    1    2.3272 11.5602 -508.10
- Biomass       1    3.0477 12.2806 -496.67
- log(EnInt)    1    5.0743 14.3073 -467.80
- log(ReEng2)   1    7.9839 17.2169 -432.82

Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(EnSup) +
    ReElect + log(ReEng2) + Biomass + log(PestSal) + PopPM +
    EnvTax + log(PopDens), data = trainset)
```

<table>
<tr>
<td>RegTrainSet2:

Regression on trainset, after backward elimination</td>
<td>
<pre>
Call:
lm(formula = log(GDPperCap) ~ log(CO2Prod) + log(EnInt) + log(EnSup) +
    ReElect + log(ReEng2) + Biomass + log(PestSal) + PopPM +
    EnvTax + log(PopDens), data = trainset)

Residuals:
     Min   1Q   Median   3Q   Max
-0.57091 -0.17948 0.02592 0.15751 0.57520

Coefficients:
              Estimate Std. Error t value   Pr(>|t|)
(Intercept)   7.402527   0.181956 40.683 < 0.0000000000000002 ***
log(CO2Prod) 0.447985   0.080200   5.586   0.00000008581615 ***
log(EnInt)   0.811315   0.082027   9.891 < 0.0000000000000002 ***
log(EnSup)   -0.118663   0.017716 -6.698   0.00000000026700 ***
ReElect   -0.005518   0.001635 -3.375   0.000906 ***
log(ReEng2)   0.569221   0.045881 12.406 < 0.0000000000000002 ***
Biomass   0.015930   0.002078   7.665   0.00000000000111 ***
log(PestSal) 0.126175   0.032238   3.914   0.000129 ***
PopPM   -0.030367   0.004850 -6.261   0.00000000277508 ***
EnvTax   -0.074553   0.029298 -2.545   0.011787 *
log(PopDens) 0.165627   0.032018   5.173   0.00000061642915 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2278 on 178 degrees of freedom
Multiple R-squared: 0.8806,   Adjusted R-squared: 0.8739
F-statistic: 131.3 on 10 and 178 DF, p-value: < 0.00000000000000022
</pre>
</td>
</tr>
<tr>
<td>RMSE for RegTrainsetB2</td>
<td>
<pre>
> RMSE.RegTrainSetB2.train
[1] 0.2210238
> RMSE.RegTrainSetB2.test
[1] 12577.92
</pre>
</td>
</tr>
</table>

Comparison between models	

CART Base Model	`> mean(predict1MAE)/mean(predictresults1$GDPperCap)    #8.6%`
	`[1] 0.0861372`
	`> mean(predict1RMSE)/mean(predictresults1$GDPperCap)   #11.7%`
	`[1] 0.1171627`

CART Base Model without Country	`> mean(predict2MAE)/mean(predictresults2$GDPperCap)    #17.58%`
	`[1] 0.1757902`
	`> mean(predict2RMSE)/mean(predictresults2$GDPperCap)   #25.84%`
	`[1] 0.2584181`

CART low capita	`> mean(predictLMAE)/mean(predictresultsL$GDPperCap)    #17.36%`
	`[1] 0.1736157`
	`> mean(predictLRMSE)/mean(predictresultsL$GDPperCap)   #20.67%`
	`[1] 0.2067256`

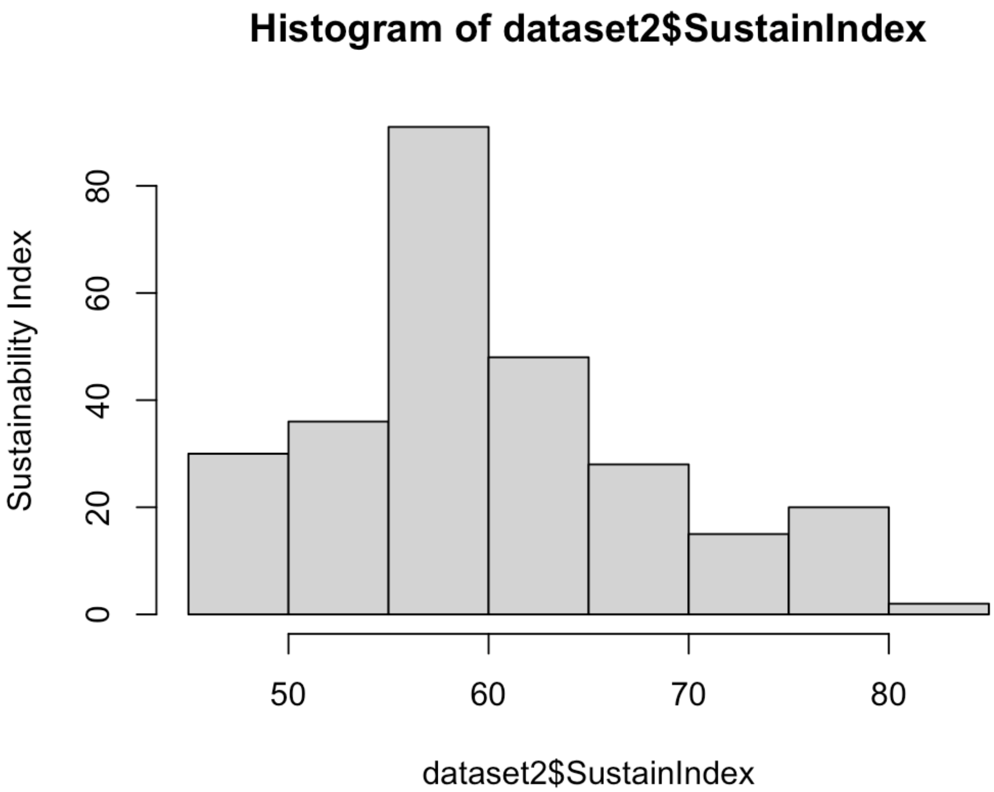
CART medium capita	`> mean(predictMMAE)/mean(predictresultsM$GDPperCap)    #7.36%`
	`[1] 0.07366934`
	`> mean(predictMRMSE)/mean(predictresultsM$GDPperCap)   #8.53%`
	`[1] 0.08530858`

CART high capita	`> mean(predictHMAE)/mean(predictresultsH$GDPperCap)    #11.95%`
	`[1] 0.1195285`
	`> mean(predictHRMSE)/mean(predictresultsH$GDPperCap)   #17.56%`
	`[1] 0.1755786`

Linear Regression Model RegTrainSet2	`> mean(MAE.RegTrainSet2.test)/mean(predictresult.RegTrainSet2.test$GDPperCap)`
	`#6.89%`
	`[1] 0.06894254`
	`> mean(RMSE.RegTrainSet2.test)/mean(predictresult.RegTrainSet2.test$GDPperCap)`
	`#11.50%`
	`[1] 0.1150013`

Linear Regression Model RegTrainSetB2	`> mean(MAE.RegTrainSetB2.test)/mean(predictresult.RegTrainSetB2.test$GDPperCap)`
	`#22.39%`
	`[1] 0.2239364`
	`> mean(RMSE.RegTrainSetB2.test)/mean(predictresult.RegTrainSetB2.test$GDPperCap)`
	`#35.01%`
	`[1] 0.3501463`

APPENDIX E

<table>
<tr><th colspan="2">Creation of ESI</th></tr>
<tr><td>Weightage assigned</td><td>log(CO2Prod) log(EnInt) log(ReEng2) Biomass PopPM EnvTax log(PopDens)
0.128 0.257 0.042 0.001 0.007 0.046 0.520</td></tr>
<tr><td>Distribution of the ESI</td><td>> summary(dataset2$SustainIndex)
Min. 1st Qu. Median Mean 3rd Qu. Max.
45.24 55.22 57.51 60.18 64.85 80.57
</td></tr>
</table>

<table>
<tr><td>Testing of the ESI</td><td>

```
Call:
lm(formula = log(GDPperCap) ~ Country + Year + SustainIndex,
    data = dataset2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.138338 -0.026427  0.000125  0.022808  0.203668

Coefficients:
                        Estimate Std. Error t value            Pr(>|t|)
(Intercept)             9.168543   0.280395  32.699 < 0.0000000000000002 ***
CountryBelgium         -0.304288   0.049046  -6.204   0.0000000024875838 ***
CountryBulgaria        -1.587987   0.051187 -31.023 < 0.0000000000000002 ***
CountryCroatia         -0.917206   0.056633 -16.196 < 0.0000000000000002 ***
CountryCyprus          -0.269282   0.046248  -5.823   0.0000000190986013 ***
CountryCzech Republic  -0.853532   0.024116 -35.393 < 0.0000000000000002 ***
CountryDenmark          0.237421   0.022951  10.345 < 0.0000000000000002 ***
CountryEstonia         -0.818201   0.041774 -19.587 < 0.0000000000000002 ***
CountryFinland          0.135947   0.037600   3.616             0.000367 ***
CountryFrance          -0.087832   0.024832  -3.537             0.000488 ***
CountryGermany         -0.193066   0.031736  -6.083   0.0000000047840443 ***
CountryGreece          -0.530844   0.044282 -11.988 < 0.0000000000000002 ***
CountryHungary         -0.976452   0.044550 -21.918 < 0.0000000000000002 ***
CountryIreland          0.401623   0.042017   9.559 < 0.0000000000000002 ***
CountryItaly           -0.312772   0.023006 -13.595 < 0.0000000000000002 ***
CountryLatvia          -0.802972   0.081413  -9.863 < 0.0000000000000002 ***
CountryLithuania       -0.769132   0.074917 -10.266 < 0.0000000000000002 ***
CountryLuxembourg       0.462103   0.064017   7.218   0.0000000000073717 ***
CountryMalta           -0.756464   0.029586 -25.568 < 0.0000000000000002 ***
CountryNetherlands     -0.227486   0.059683  -3.812             0.000177 ***
CountryPoland          -1.070060   0.033083 -32.345 < 0.0000000000000002 ***
CountryPortugal        -0.549816   0.045216 -12.160 < 0.0000000000000002 ***
CountryRomania         -1.231294   0.071465 -17.229 < 0.0000000000000002 ***
CountrySlovak Republic -0.835227   0.031674 -26.370 < 0.0000000000000002 ***
CountrySlovenia        -0.655824   0.023506 -27.900 < 0.0000000000000002 ***
CountrySpain           -0.258034   0.041731  -6.183   0.0000000027866307 ***
CountrySweden           0.384636   0.050416   7.629   0.0000000000005999 ***
Year2011                0.026256   0.014053   1.868             0.062960 .
Year2012                0.027241   0.014224   1.915             0.056704 .
Year2013                0.036834   0.014506   2.539             0.011762 *
Year2014                0.065362   0.014990   4.360   0.0000194862474121 ***
Year2015                0.098303   0.014940   6.580   0.0000000003075586 ***
Year2016                0.118137   0.014675   8.050   0.0000000000000423 ***
Year2017                0.144856   0.014281  10.143 < 0.0000000000000002 ***
Year2018                0.175171   0.014322  12.231 < 0.0000000000000002 ***
Year2019                0.210206   0.014955  14.056 < 0.0000000000000002 ***
SustainIndex            0.023680   0.004276   5.538   0.0000000824445753 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05118 on 233 degrees of freedom
Multiple R-squared:  0.9943,    Adjusted R-squared:  0.9934
F-statistic:  1132 on 36 and 233 DF,  p-value: < 0.00000000000000022
```

</td></tr>
</table>