Financial analysis of nations towards participation in decreasing CO_2 emissions over time

Project Report

Neeraj Vijay Bedmutha

December 14, 2019

- ## Introduction:

The Paris Agreement aims to respond to the threat of climate change by keeping the global temperature rise this century well below 2°C above pre-industrial levels. The agreement further challenges itself to limit this rise in temperature to 1.5°C. The agreement is signed by a total of 196 countries. To meet the goals of reducing carbon emissions and restricting this temperature rise, countries are expected to shift their energy mix in a way so that it depends more on renewable energy and cleaner sources. However, investments into renewable energy for every state are posed with several problems and dynamics. Although some countries have peaked their CO_2 emissions (not increasing CO_2 emission level further), there are many developing, low income states with dearth of clean energy consumption that need to progress in this direction. The first step for such states in reducing CO_2 emissions is to achieve stable or decreasing emission level. The aim of this project is to utilize parametric regression models and estimate if a country's economic conditions favor its participation towards net decrease in carbon emissions over time. The variables involved in the analysis are new renewable energy investments (REI), carbon emissions and gross domestic product (GDP).

- ## What Questions Does the Project try to Answer?

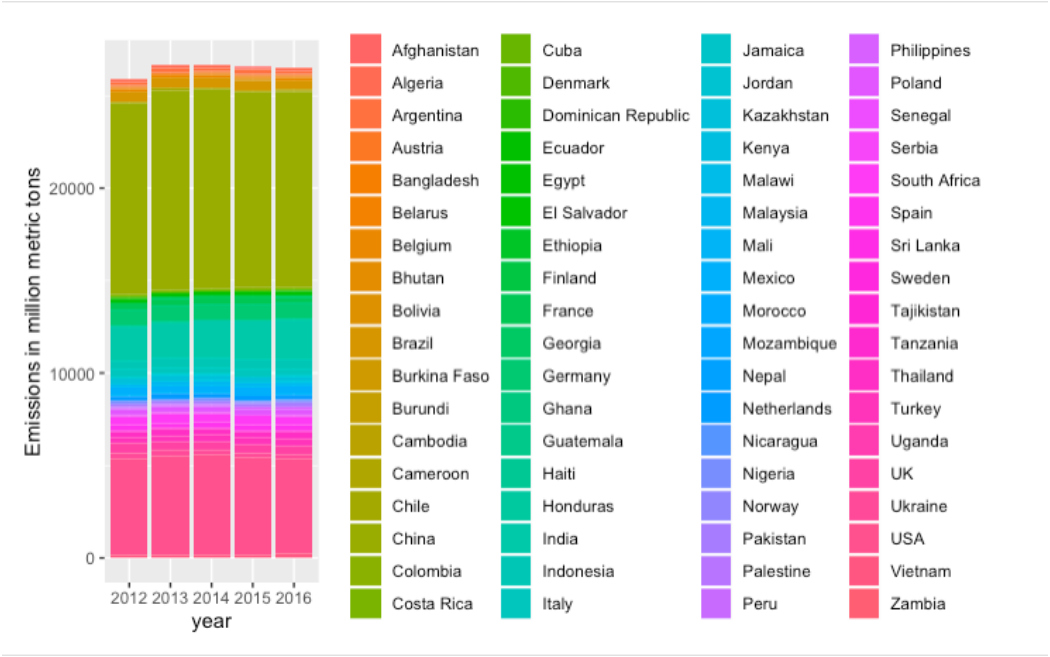
The questions below try to present the points the model tries to infer.

1. Do the new renewable energy investment (REI) of countries have a correlation to their GDP? If yes, what is it?
2. What is the cost/new renewable financial investment necessary for a country to participate towards a goal of net decrease in carbon emission over the previous year? The model tries to predict if the carbon emissions decreases with increasing renewable investments over the years.
3. What are the other possible factors that can be represented to have a relationship with carbon emissions through a parametric model? These are discussed under the section for omitted variables and omitted variable bias.

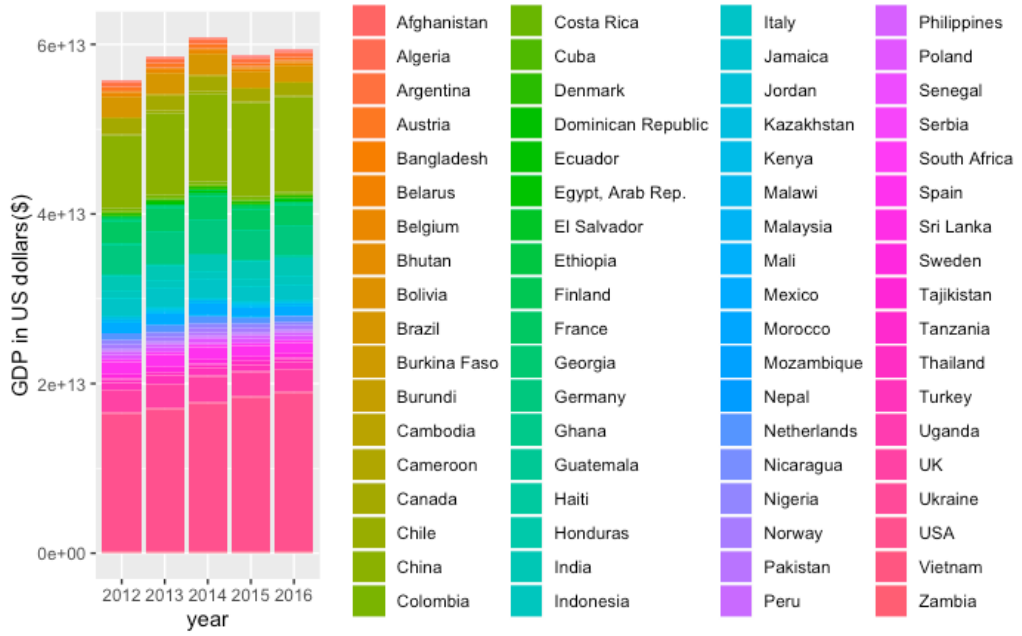
- **Data Description**

In order to carry out suitable analysis, the data is acquired from multiple sources. The data consists of:

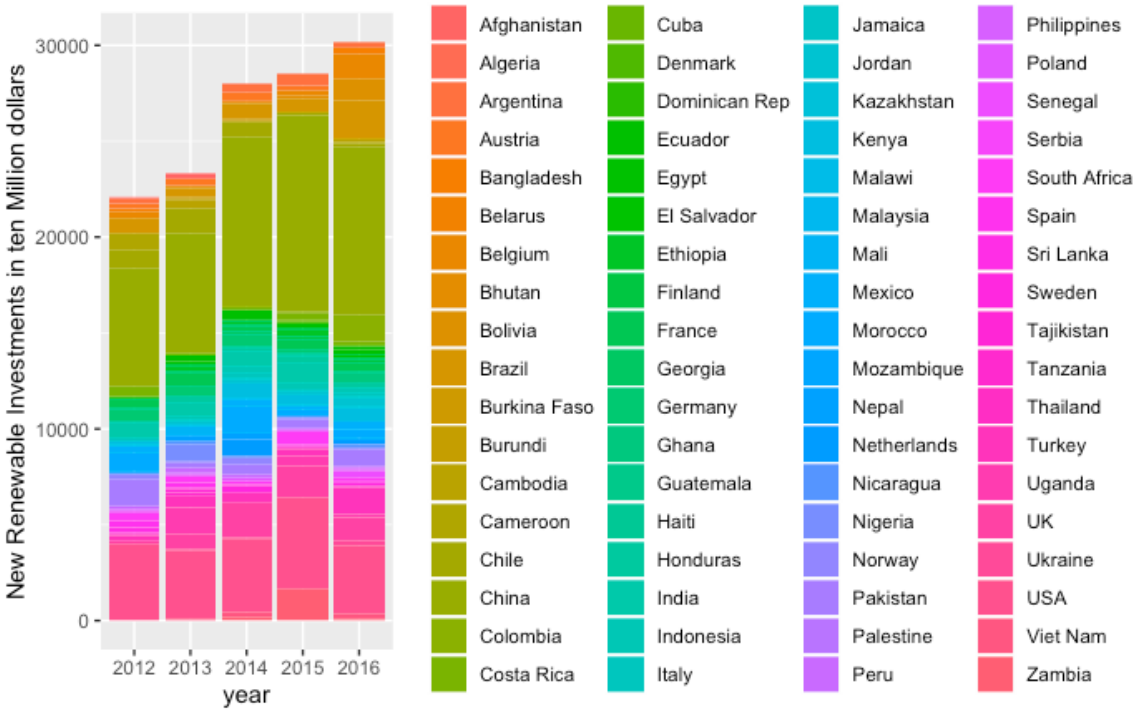
1. The total carbon di-oxide (CO2) emissions from the consumption of energy for 227 countries. The data used is from 2003 to 2016. The CO2 emissions are measured in million metric tons. We use the CO2 emissions from 2012 to 2016 (spanning till the maximum years that provide all the three data). The data used is for 72 countries of the world along with a cumulative world data for all the years. Below is a geom_column plot of the same.



2. The gross domestic product (GDP) for the same 72 countries from 2012 to 2018. The GDP is measured in current US dollars ($). Gross domestic product is a measure of the country's net value of all the finished goods and services throughout a year. We consider the data from 2012-2016 for all the countries along with the world cumulative. Below is a column bar graph of the same.



3. The new renewable energy investments for 220 countries from 2009 to 2018. The investments are measured in 2016 equivalent 10 million US dollars ($). The data has been later scaled for our convenience and visual purpose. The data for the year 2018 is available only for certain regions of the world. Because the data for carbon emissions is available for as recent as 2016, we will consider a common year span for all datasets not beyond 2016.



- ## Dependent Variable

Yearly CO_2 Emissions: The main analysis below is done to predict if a country's financial resources support its goal of decreasing CO_2 emissions over time. We establish the correlation of yearly CO_2 emissions with new REI and GDP variables using our models and estimate the possible result. The dependent variable *yearly CO_2* emissions is measured in Million metric tons. The emission for every country involved is a data points y_i. We establish the partial dependency of REI and GDP on CO_2 emissions before establishing our model for the two variables combined.

- ## Independent Variables

Renewable Energy Investments: Renewable energy consumption is one of the fundamental steps towards achieving the Paris Agreement goals. We use the new renewable energy investment as a metric to do a financial analysis of a country's progress. The investments are reflective of how much energy consumption is done through renewable built capacity. It is inclusive of solar energy (solar photovoltaic, concentrated solar power), hydropower, wind energy, biofuels, geothermal and other renewable resources. The investments are obtained from International Renewable Energy Agency (IRENA) and are measured in million US dollars (2016 equivalent). This variable is expected to have a negative correlation with the dependent variable, which is to say that increasing renewable investments should decrease CO_2 emissions

Gross Domestic Product (GDP): The GDP of a country is an indicator of its economy's health and progress. More GDP is expected to imply higher consumption of energy. For countries where the energy mix doesn't have enough renewable energy capacity, GDP is expected to negatively correlate to emissions, while countries with abundant renewable energy capacity in their energy mix are expected to have a positive relation.

Countries: The countries chosen for the analysis are a list of 72 countries between the year 2012-2016. Along with the countries, total world data for all these years is present to see the cumulative results and pattern. We want to assess the world cumulative data separately in order to observe results on the big scale. The 3 datasets were combined to test the models. Below is the data frame representing the first 10 rows of the combined dataset.

The models are generated in a way to proceed from a benchmark case to the best fit model using parametric models and generalized additive model. In order to check how well the model performs on test datasets, a cross validation model is analyzed.

Country <chr>	year <fctr>	REI <dbl>	gdp_values <dbl>	emission_values <dbl>
Afghanistan	2012	59.90	20.001616	9.664710e+00
Algeria	2012	0.94	209.000000	1.296666e+02
Argentina	2012	243.57	546.000000	1.972963e+02
Austria	2012	252.06	409.000000	6.827096e+01
Bangladesh	2012	143.10	133.000000	6.338486e+01
Belarus	2012	2.83	65.685103	6.818174e+01
Belgium	2012	373.14	498.000000	1.328898e+02
Bhutan	2012	3.34	1.823692	6.056758e-01
Bolivia	2012	4.69	27.084498	1.590263e+01
Brazil	2012	766.00	2470.000000	5.041295e+02

1–10 of 360 rows

All countries combined dataset

- **Model Description**

The models below are used to establish the correlation between the dependent and the independent variables. These models are used for the above mentioned dependent. (yearly CO2 emissions). Cross validation is incorporated to prevent overfitting using the k-fold cross validation method. The RMSE error terms are expected to decrease as the model will better fit with cross validation.

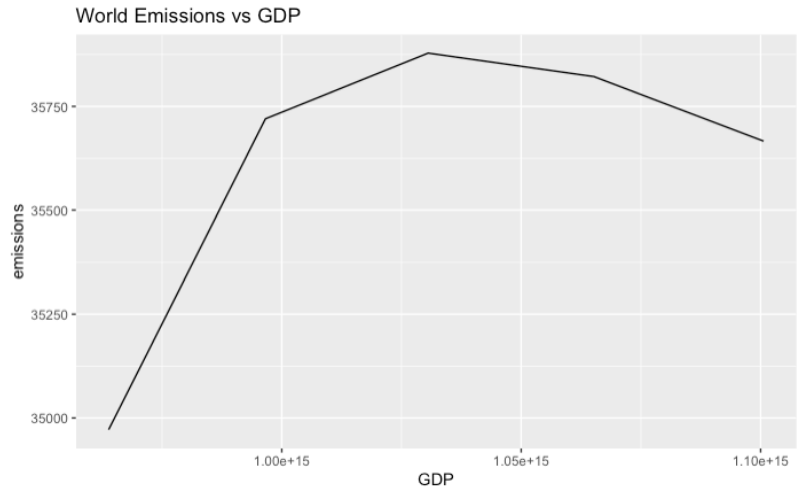
Dependent Variable: yearly CO_2 emission ($y_{i,t}$), *where i's* are the countries and *t* denotes the year.
Covariates / Regressors: Renewable energy investment $(REI)_{i,t}$, Gross Domestic Product (GDP) $_{i,t}$ *where i's* are the countries and *t* denotes the year.

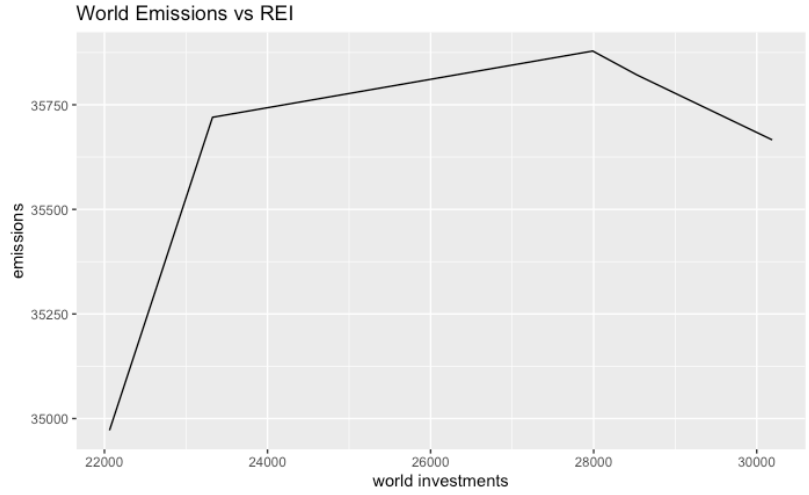
- **Preliminary Results**

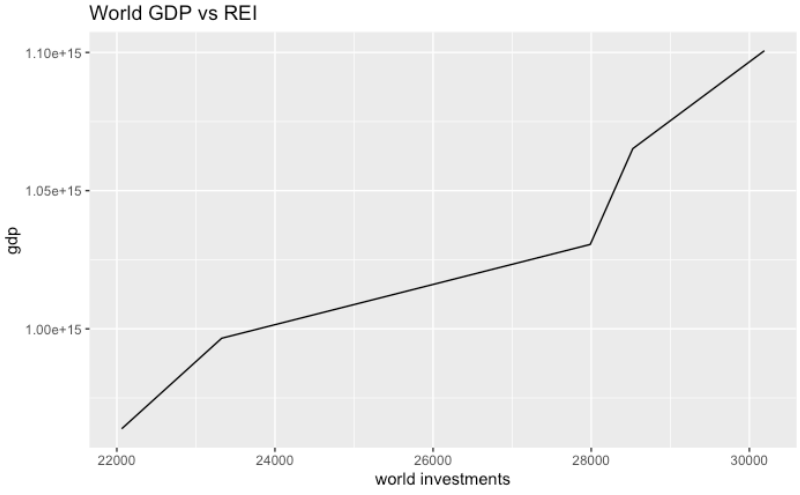
We establish some preliminary results before implementing our first model using all the possible regressors.

1. The world average of carbon emissions over the year has not been increasing in a linear fashion. Its dependence on renewable energy is estimated to show a negative relation.
2. The carbon emissions of a country are expected to increase as the GDP of countries increase. This is based on the idea of increased use of energy by a country following its growth. Even though the world emissions are not increasing as the years progress, the net total emissions per year for the entire world is a huge number.
3. Lastly, we plot the world GDP with the REI to assess if the world GDP has any effect on the new renewable energy investments made every year.

The three plots depicting these three relations are done below.





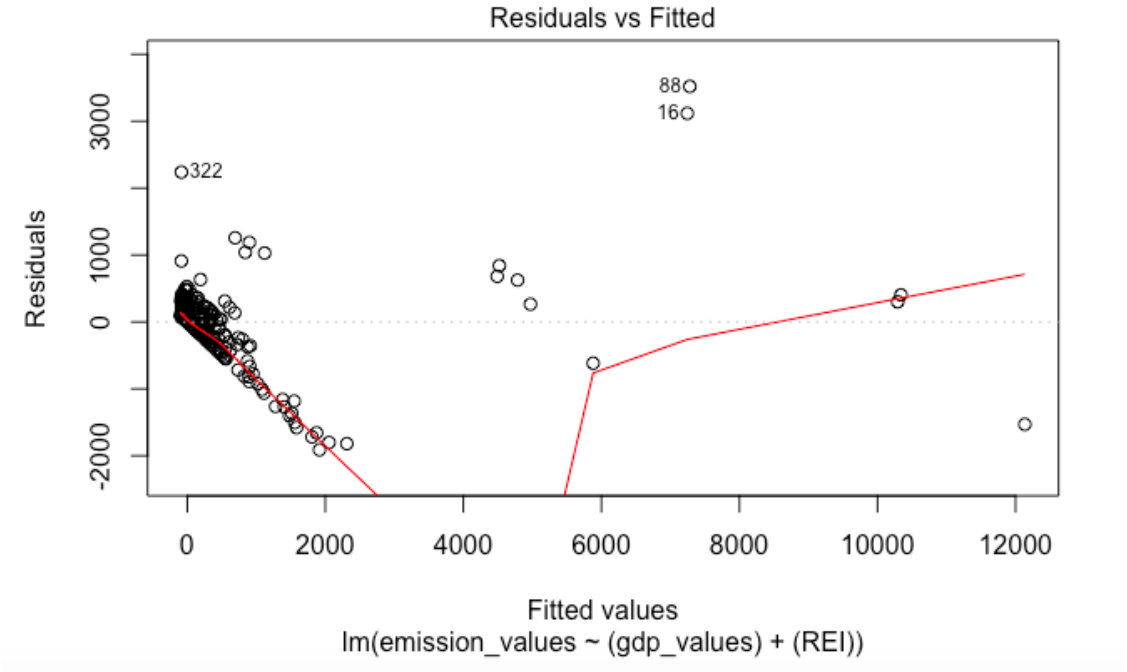


Model1: Parametric regression model (linear regression) is used as a benchmark model. We use a linear model as the benchmark model to verify our expected correlations.

Model Equation: $y = \beta_0 + \beta_1*(REI) + \beta_2*(GDP) + \varepsilon$,

*For a specific i*th year, *and time t,* $y_{i,t} = \beta_0 + \beta_1*(REI)_{i,t} + \beta_2*(GDP)_{i,t} + \varepsilon_{i,t}$

Where β*'s* are the parameters for the linear regression equation.



Observations of the linear model: The residual values of the dependent variable plotted with the fitted values shows that a linear model does not fit well on the data. The data when pointed for all the countries and all the years becomes haphazard in the linear model. Although the world cumulative shows the dependency of the variables amongst each other, it doesn't dawn upon the variability in countries. Some countries are big economies and their emissions are huge numbers as compared to smaller nations. At the same time, different countries have varying CO_2 emissions, GDP values and new renewable investments over the years. Thus, the data is skewed and cannot be explained by something as simple as a linear model. Below are the coefficients as observed in the linear model in R.

```
Coefficients:
(Intercept)   gdp_values          REI
 -8.524e+01    2.127e-15    1.187e+00


Call:
lm(formula = emission_values ~ (gdp_values) + (REI), data = all_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1912.0   -79.3    55.9   119.9  3520.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.524e+01  2.828e+01  -3.014  0.00277 **
gdp_values   2.127e-15  1.150e-15   1.850  0.06515 .
REI          1.187e+00  2.562e-02  46.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 495.2 on 357 degrees of freedom
Multiple R-squared:  0.8735,    Adjusted R-squared:  0.8728
F-statistic:  1232 on 2 and 357 DF,  p-value: < 2.2e-16
```

Model2: Generalized Additive Model (GAM)

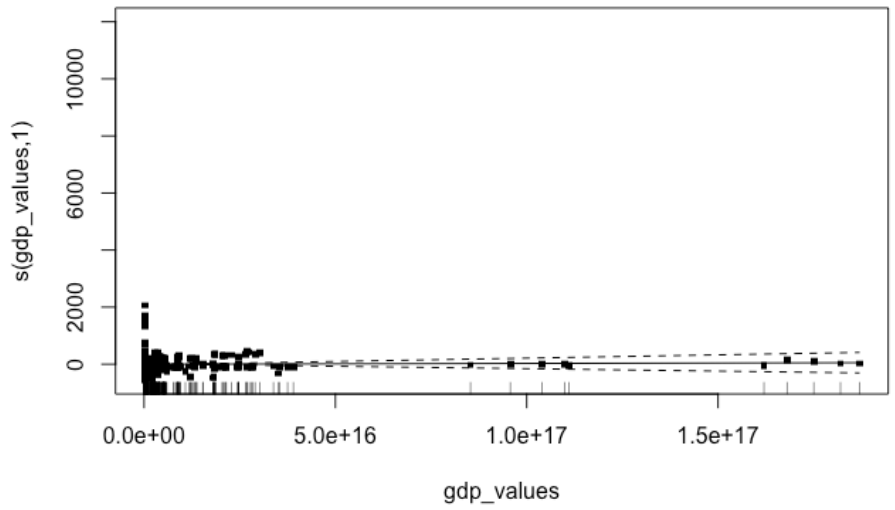
In order to determine the best possible polynomial (curve) that fits the data, we observe the smoothing splines generated by the GAM model. Usually the GAM model presents the non-linear relationships between variables. We can use the visual plots to estimate the best model that represents the smooth GAM function.

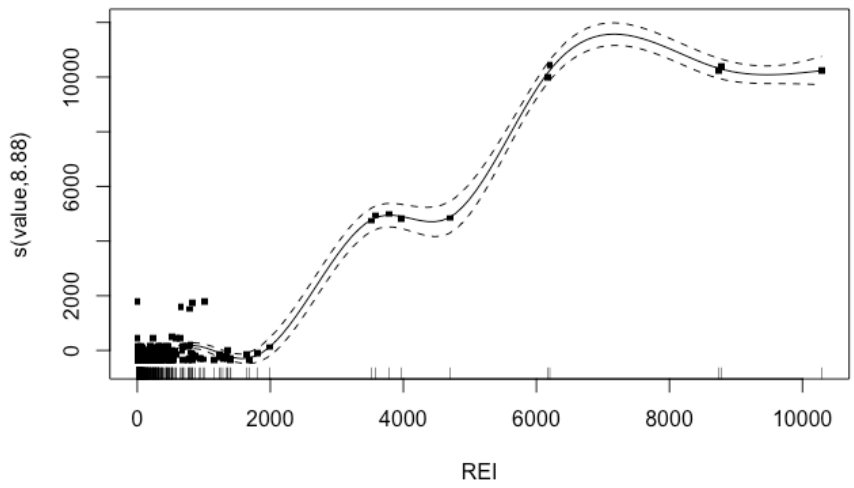
Model Equation: $y = \beta_0 + \beta_1 * s(REI) + \beta_2 * s(GDP) + \varepsilon$,

*For a specific i*th year, *and time t,* $y_{i,t} = \beta_0 + \beta_1 * s(REI)_{i,t} + \beta_2 * s(GDP)_{i,t} + \varepsilon_{i,t}$

Observations of the GAM model:

Below are the partial residual plots for the smoothing functions on the regressors.





The GAM model partial residual plot for emission values vs new renewable energy investments is a polynomial function. Although the function is a higher degree polynomial, it depicts a linearly positive correlation. To fit the regressor for REI, a linear or logarithmic relationship can be used as before in the lm model.

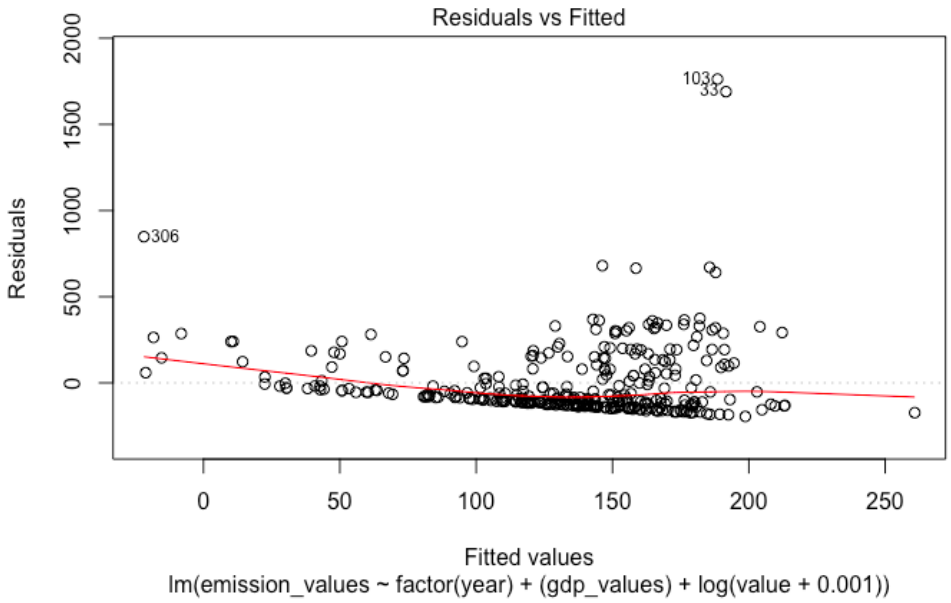
At the same time the GAM model looks linear and provides a clear correlation between the emission values and the GDP regressor. As is evident from the GAM model similar to the lm model is that the dependent variable and the regressor are being modeled invariant to the countries and the year. This disturbs the analysis by not capturing the clustering pattern in the data points.

Based on the inferences from the GAM model and the use need for factoring our dataset, the model3 is prepared. The summary from the GAM model in R is presented below.

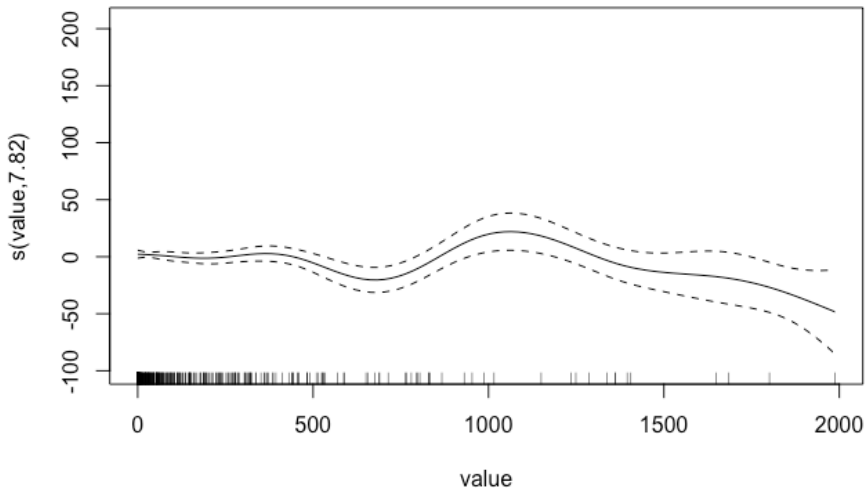
```
Formula:
emission_values ~ s(gdp_values) + s(REI)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   367.63      13.68   26.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                edf Ref.df      F p-value
s(gdp_values) 1.000  1.000   0.09   0.764
s(REI)        8.885  8.995 972.90  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.965   Deviance explained = 96.6%
GCV =  69519  Scale est. = 67417     n = 360
```

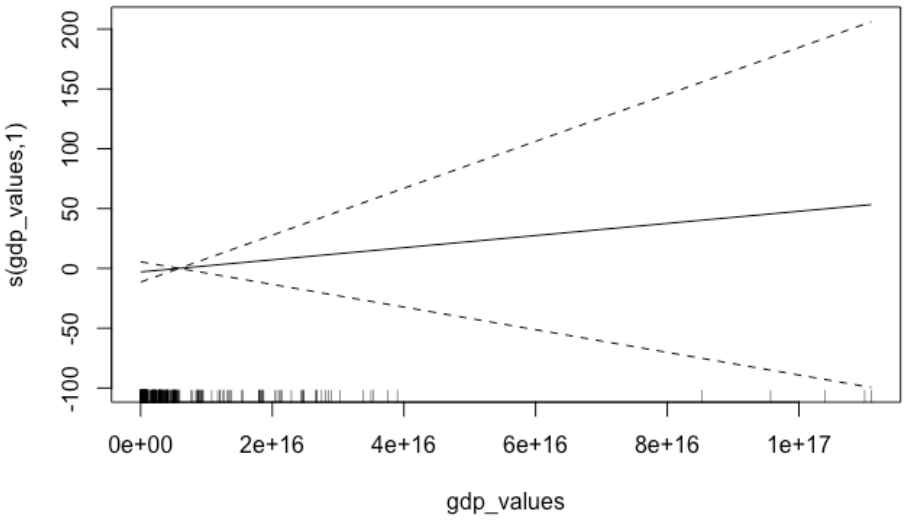
Model3: Parametric Model with Factors

We model our parametric model to be logarithmic in the renewable energy investment regressor and linear in the GDP regressor. The log (REI) is used because it fits much better as compared to the linear dependence used in model1. The plot with the log of REI regressor denotes the same. The residual vs fitted values plot moves closer to the zero-residual line. However, this is not the entire fix we need to do as the residuals still have a lot of scope for improvement. The model can thus be improved.



On simply including the factors for countries and year using the dummy variable in the GAM function, we try to observe the smoothing curves. The partial residual plots of the dependent variable are plotted. Below are the plots:





1. At the same time, the plots depict that global emissions for the world are decreasing with increase in the new renewable energy investments. This also aligns with our estimates keeping in mind that renewable or lean sources of energy will improve the energy mix and decrease the amount of carbon emissions.

2. The plots above depict the expected relationship between CO_2 emission values and the regressors. As is visible from the partial residual plot, the global emissions tend to increase as the GDP of the countries increase. This can clearly be attributed to higher net energy usage by the entities during the period of growth.

Finally, after we have made suitable updates based on our previous model, we can model a new dependency between the regressors and the dependent variable.

Model Equation: $y = \beta_0 + \Sigma\, factor(country) + \Sigma\, factor(year) + \beta_1 * log(REI) + \beta_2 * (GDP) + \varepsilon$,

*For a specific i*th country and year *t*,

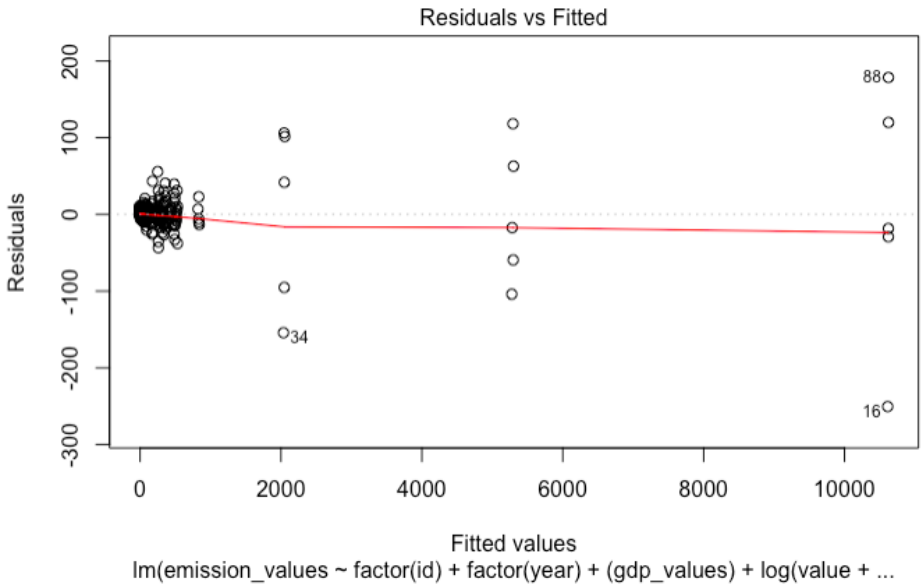
$$y_{i,t} = \beta_0 + \beta_1 * log(REI)_{i,t} + \beta_2 * (GDP)_{i,t} + \beta_i(factor(country_i) + \beta_t(factor(year_t)$$

$$y_{i,t} = \beta_0 + \beta_1 * log(REI)_{i,t} + \beta_2 * (GDP)_{i,t} + \gamma_i + \delta_t$$

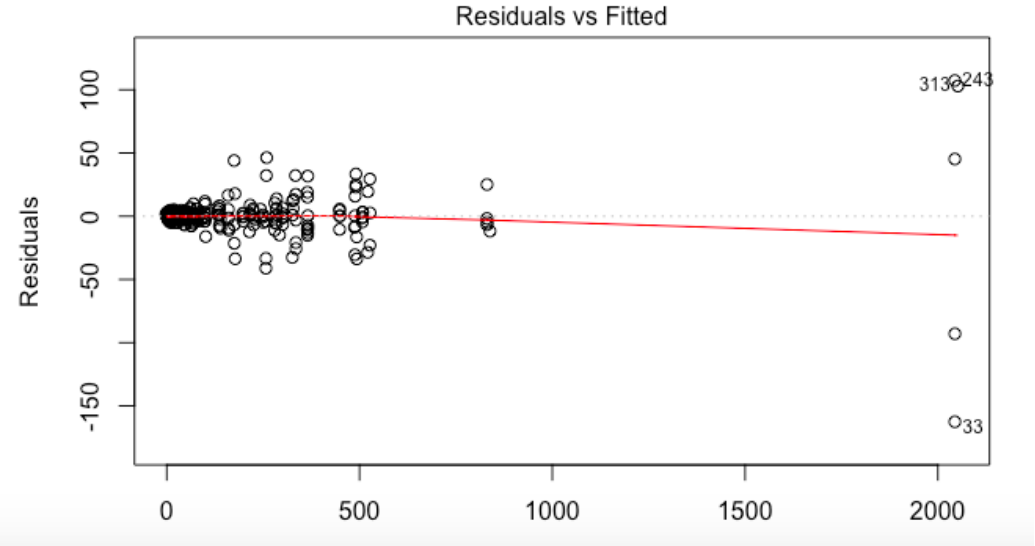
Note: Each factor in the regressors has a different coefficient. We use a small correction factor of 0.001 to fix the missing values in the new renewable energy investment dataset. We also account for China even though it behaves like an outlier for our data. However, we skip the point for China while arriving at the decisions for a need for factors.

Observations:

The fitted values have a decently better performance as compared to the previous models. The average residual value fits close to 0.



If we remove China from the analysis, the residual plot line further shifts to the zero line. Since, we think China behaves like an outlier for this dataset with most countries having very small emission values as compared to China, we believe it has a high leverage. It affects the way we look at the residual plot as well as the GAM plots.



Below is a table depicting the residuals values of the model. The coefficients can be observed as an entire list in the knitted file at the bottom.

```
Call:
lm(formula = emission_values ~ factor(id) + factor(year) + (gdp_values) +
    log(value + 0.001), data = renew_vs_gdp_emissions2)

Residuals:
     Min       1Q   Median       3Q      Max
-158.640   -2.566    0.144    2.465  105.461
```

Even if we do not remove the outlier behavioral value, this is a better model as compared to the previous models. We can thus conclusively say that the logarithmic transformation on the REI regressor, a linear relationship of the CO_2 emission values with the GDP regressor and the inclusion of factors for countries and years in our parametric model is a good set of changes.
The above model has a coefficient for each of the factors along with the other regressors.

Model4: *Cross Validation*

Cross validation 1:
We use cross validation to test our model performance on a set of data after training on a larger set to evaluate its performance on unseen data. We use both Model2 with factor for the years and Model3 to perform cross validation. The test set has all the factors from the training set as its regressors. Thus, no possible factor is missed out in the test set. We cannot regress over the country's as factors because the number of countries is large and they all the countries in the training set may not be present in the test set for the available years. Thus, our cross-validation process will be invariant to the countries but will still fit the dependent variable using years as factors. The REI and GDP regressor will be used as spline function regressors for model2 and will be remain as they were for model3. The sampling used for making the datasets is random.

The aim is to verify the best estimated model and the relationship between the dependent and the independent variable using the cross-validation process. If the complex model learnt in model3, fits better than GAM for cross validated sets, the model is a good fit. We carry out cross validation by using k-fold cross validation process. We use a randomly mixed set called the test set out and then estimating the error for the 2 models in place.

Cross validation 1 function attributes:
The k fold cross validation function uses k=5 folds for our dataset where 4 folds form the training set and 1-fold forms the test set.
The cross-validation process is run for 100 epochs or 100 runs before it finally calculates the rMSE for the specified models.

Cross Validation 2:
We also use cross validation to predict 2016 emissions data using the 2012-2015 as the train set. Cross validation for these manually distributed sets, behaves like a future prediction for the already known dataset of 2016. We calculate the residuals and subsequently the rMSE of the predictions by comparing them to the actual set.

Observations:
Corss-validation 1:
Below are the rMSE values observed as outputs (screenshot) for the 2 models used in cross validation1. The rMSE values are after 100 epochs.

```
[[1]]
[1] 1381.893

[[2]]
[1] 675.3221


Call:
lm(formula = emission_values ~ factor(Country) + gdp_values +
    log(REI + 0.01), data = train_beta)
```

RMSE for (GAM) model: 1381.893
RMSE for final parametric model: 675.3221

The rMSE values do not remain constant as the random sampling keeps changing the training set and the test set. Since the coefficients are a long list (one for each factor) they are not shown using a screenshot and can be found at the knit file in the bottom.

Cross-validation 2
We also use cross validation to predict the values in 2016 using 2012-2015 as the training set. This cross-validation process is performed in the R markdown file as well. As opposed to the first cross validation process, here we can use the factors for countries but cannot use the factors for years as we divide our dataset based on the year. Since cross validation has unseen data, it is a good test of how the model will perform on future datasets.

The rMSE observed on the test set for the year 2016 is put below: 769.0587

Cross Validation Inferences

Both the cross-validation results give a rMSE value, which is the root mean squared error between the predicted and the actual value. The CO_2 emission values span from a range of 0 to 12000 for the entire dataset. So, the rMSE is very small as compared to the highest value in the range. However, this is not the best metric of how well the model performs. Some countries which have really small actual emission value, but have a really large value predicted value, and result in a rMSE that is not comparable or sensible data.

A good metric for judging our model is to see its performance as compared to the GAM model. It can be seen from the results above, that the chosen parametric model, model3 works better on the data as compared to the GAM function for cross validation too. We carry out the remaining investigations necessary for analysis before concluding our results.

- ## Investigations

1. The above models are estimated based on plots and variations of partly selected data or world averages. I plan on investigating further variations using my GAM model to come up with the best fit parametric model if exists.

2. *Omitted variable bias:* Overall Energy Consumption of a country is a plausible omitted variable in deciding the carbon emissions. Despite increase in the new-renewable investments, a country might have increase in carbon emissions because it needs to consume more energy. At the same time the increase in energy consumption should be related to the gross domestic product of a country. If the residual plots do not obey any parametric model, I would include the omitted variable's contribution.

   Omitted Variable bias $\propto \quad \gamma \times \delta \quad$ where γ is the regression slope between the omitted variable and the dependent variable, and δ is the regression slope between the omitted variable and the other regressor.

3. *Gauss Markov Assumptions:*
   a. Random Sampling: We consider data from 80 countries with random choosing and sampling. There is no bias while choosing the dataset.
   b. Linearity in Parameters: It is assumed that the dependent variable can be modeled as a linear combination of the independent variables with the help of constant coefficients.

c. No perfect collinearity: Two variables that have perfect collinearity affect the dependent variable in the dame way. Therefore, we do not want to use such variables. The GDP, carbon emissions and new renewable investments are all different datasets which cannot have perfect collinearity and thus follow the Gauss Markov Assumption

- ## Results

We attempted to model the behavior of the world CO_2 emissions with the GDP of a country and the new renewable energy investments. We try to find a good model that fits the data well by working our way through a linear model, the generalized additive model. We finally test our model using cross validation methods.

The results for the residuals of our model tells us that it a good fit model and should depict the relationship between the regressors and the dependent variable quite accurately. The rMSE value from the cross-validation results obtained are listed in the cross validation section. We can comment on how the model is a good model because it performs better than GAM in the cross validation process,but cannot infer much from the rMSE values alone.

- ## Application

The model can be used with a lot more depth to understand the if countries and the world in general is ready to meet the goals of the Paris agreement and if they have the resources necessary for meeting the renewable energy investments required.

- ## References

1. *https://www.irena.org/publications/2019/Jul/Renewable-energy-statistics-2019*
2. *https://www.eia.gov/beta/international/data/browser/#/?c=41000000020000600000000000000g000200000000000000001&vs=INTL.44-1-AFRC-QBTU.A&vo=0&v=H&start=1980&end=2017&showdm=y*

- **Appendix**

The knit Rmd file is presented below:

Project

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see *http://rmarkdown.rstudio.com*.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
setwd("~/Downloads/possible data /csv files")

investments2 <- read.table("new_renewable_investment.csv", header = TRUE, sep
= ",",stringsAsFactors = FALSE, check.names = FALSE)
investments2$`2016` <- as.numeric(investments2$`2016`, "NA"=FALSE)

## Warning: NAs introduced by coercion

investments2[is.na(investments2)] <- 0
investments3<-investments2[,-c(2,3,4)]
investments <- investments3[1:72,]
df_melt = melt(investments, id = "(ten_million_dollar)")
colnames(df_melt) <- c('id','year','value')
#has all the new renewable energy investments.

plot3<-ggplot(df_melt, aes(year)) +
  geom_col(aes(x=year, y=value,fill = id))
plot3+labs(y="New Renewable Investments in ten Million dollars", x = "year")
```


```
#histogram plot of REI

world_investments<-investments3[80,1:6]
world_investments = melt(world_investments, id = "(ten_million_dollar)")
colnames(world_investments) <- c('id','year','value')
#has onl the world renewable investments


gdp2 <- read.table("gdp.csv", header = TRUE, sep = ",",stringsAsFactors = FAL
SE, check.names = FALSE)
gdp2[2]<-NULL
gdp <- gdp2[1:72,1:6]
gdp[is.na(gdp)] <- 0
gdp_melt=melt(gdp, id = "GDP")
colnames(gdp_melt) <- c('id','year','value')
#has all the gdp values

gdp_melt$year<- as.character(gdp_melt$year, "NA"=FALSE)
gdp_melt$year<- as.numeric(gdp_melt$year, "NA"=FALSE)
#changing factors to numeric

plot4<-ggplot(gdp_melt, aes(year)) +
```

```
  geom_col(aes(x=year, y=value,fill = id))
plot4+labs(y="GDP in US dollars($)", x = "year")
```

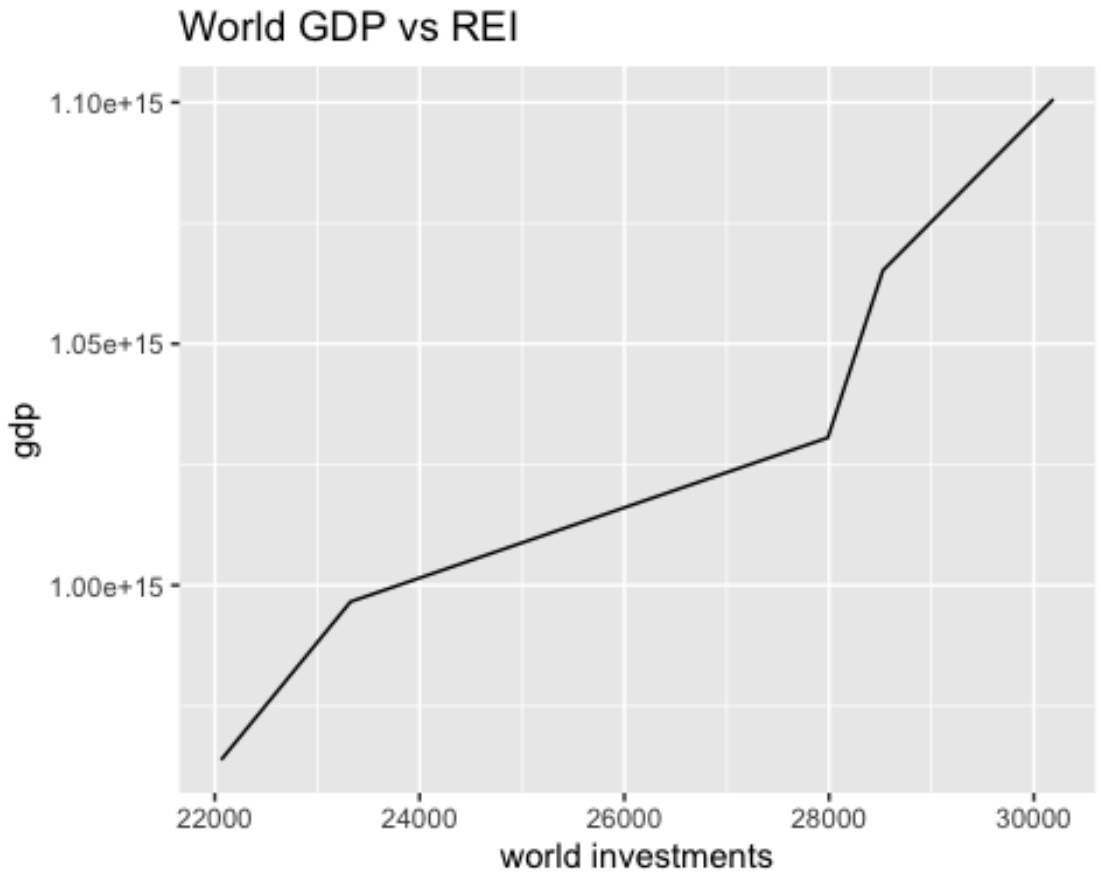

```
#histogram plot of gdp

world_gdp<-gdp2[80,1:6]
world_gdp = melt(world_gdp, id = "GDP")
colnames(world_gdp) <- c('id','year','value')
#has the world gdp values only

emissions2 <- read.table("50 countries emissions.csv", header = TRUE, sep = "
,",stringsAsFactors = FALSE, check.names = FALSE)
emissions2[2]<-NULL
emissions<-emissions2[,-c(2,3,4,5,6)]
emissions_world<-emissions[80,1:6]
emissions <- emissions[1:72,1:6]
emissions_melt=melt(emissions, id = "Emissions")
colnames(emissions_melt) <- c('id','year','value')
#data for emissions

emissions_world=melt(emissions_world, id="Emissions")
colnames(emissions_world)<-c('id','year','emission_value')
#has the world emissions
```

```
plot2<-ggplot(emissions_melt, aes(year)) +
  geom_col(aes(x=year, y=value,fill = id))
plot2+labs(y="Emissions in million metric tons", x = "year")
```

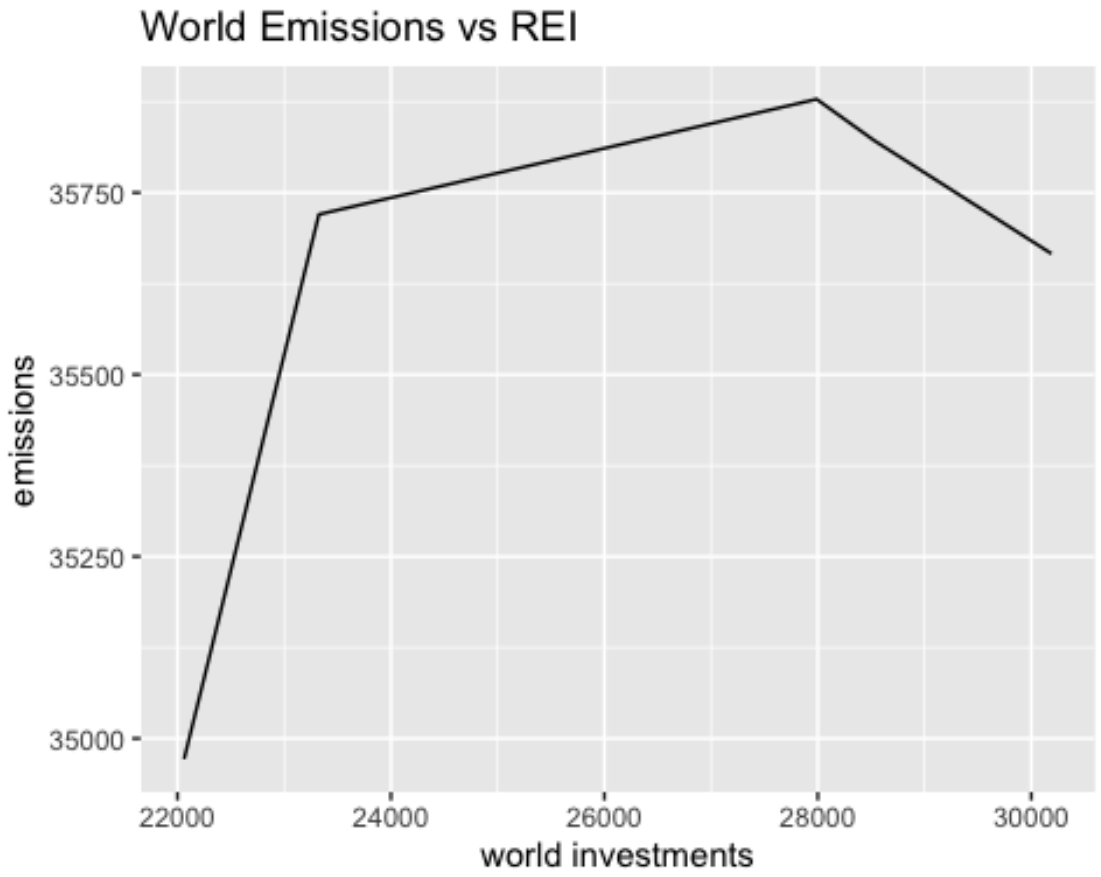

```
#Histogram Plot for emissions


world_gdp_investments<-world_gdp%>%
  mutate(investments=world_investments$value)
#for world's gdp vs investment relations

world_gdp_investments[5]<-emissions_world[3]
world_gdp_investments
```

```
##      id year       value investments emission_value
## 1 World 2012 9.63845e+14    22061.97       34971.76
## 2 World 2013 9.96597e+14    23326.20       35720.13
## 3 World 2014 1.03057e+15    27988.07       35878.52
## 4 World 2015 1.06521e+15    28526.24       35821.84
## 5 World 2016 1.10064e+15    30190.08       35666.08
```

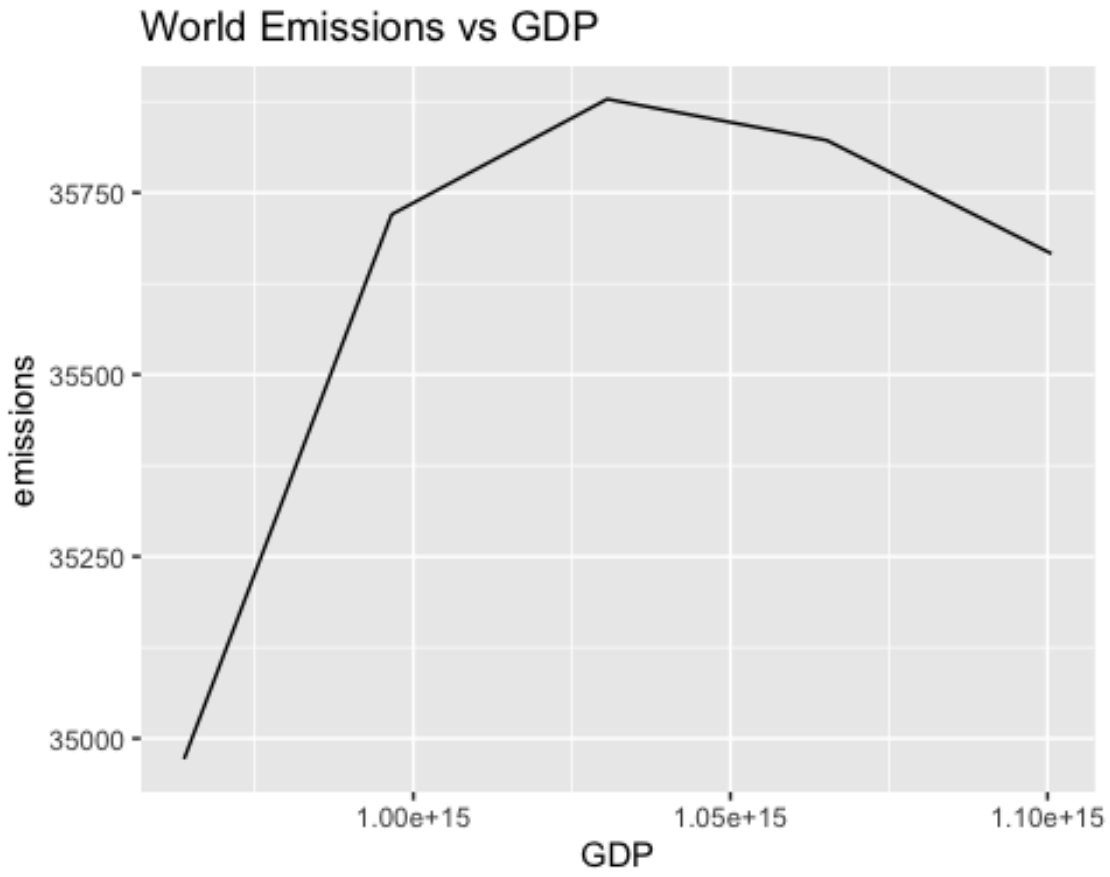
```
#has combined the three world datasets

plot1<-ggplot(aes(x=investments, y=value), data=world_gdp_investments)+geom_l
ine()
plot1+ labs(title= "World GDP vs REI",
                      y="gdp", x = "world investments")
```




```
plot5<-ggplot(aes(x=investments, y=emission_value), data=world_gdp_investment
s)+geom_line()
plot5+ labs(title= "World Emissions vs REI",
                      y="emissions", x = "world investments")
```

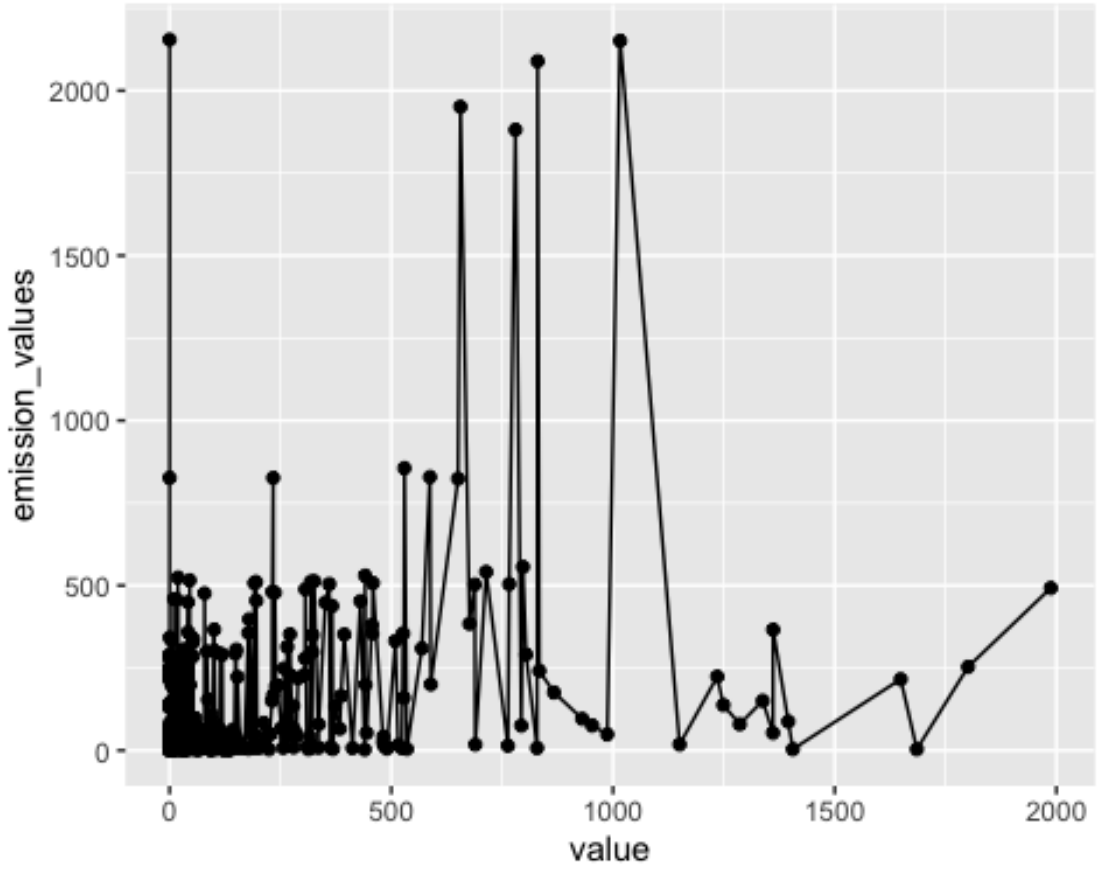



```
plot6<-ggplot(aes(x=value, y=emission_value), data=world_gdp_investments)+geo
m_line()
plot6+ labs(title= "World Emissions vs GDP",
                      y="emissions", x = "GDP")
```

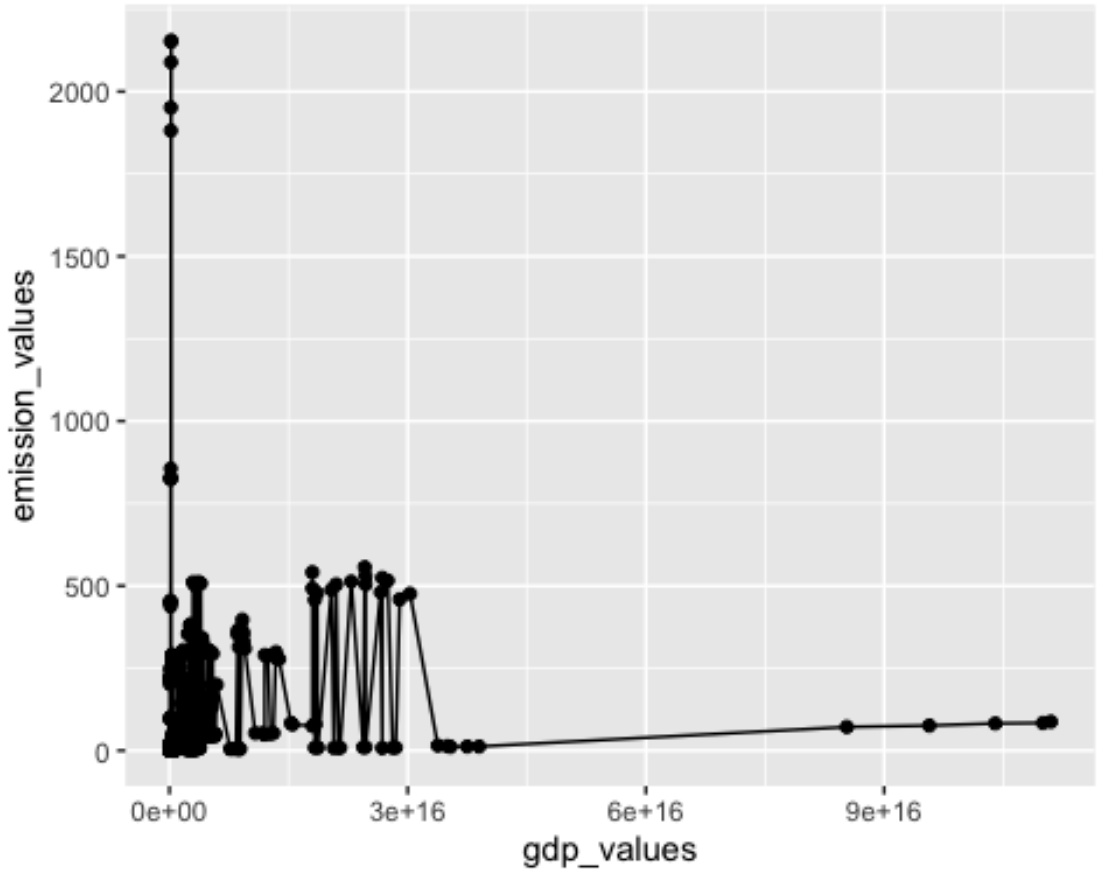



```
#World plots done


gdp_values<-gdp_melt$value
investment_values<-df_melt$value
renew_vs_gdp3<-df_melt%>%
  mutate(gdp_values=gdp_values)
#has all the gdp and investment values



#this section had additional plots


#Emissions vs gdp vs investments below
renew_vs_gdp3[,4]<-renew_vs_gdp3[,4]*10e+3
emissions_values<-emissions_melt$value
renew_vs_gdp_emissions<-renew_vs_gdp3%>%
  mutate(emission_values=emissions_values)

renew_vs_gdp_emissions<-renew_vs_gdp_emissions%>%
  mutate(lsinvest=log(value))
renew_vs_gdp_emissions<-renew_vs_gdp_emissions%>%
```

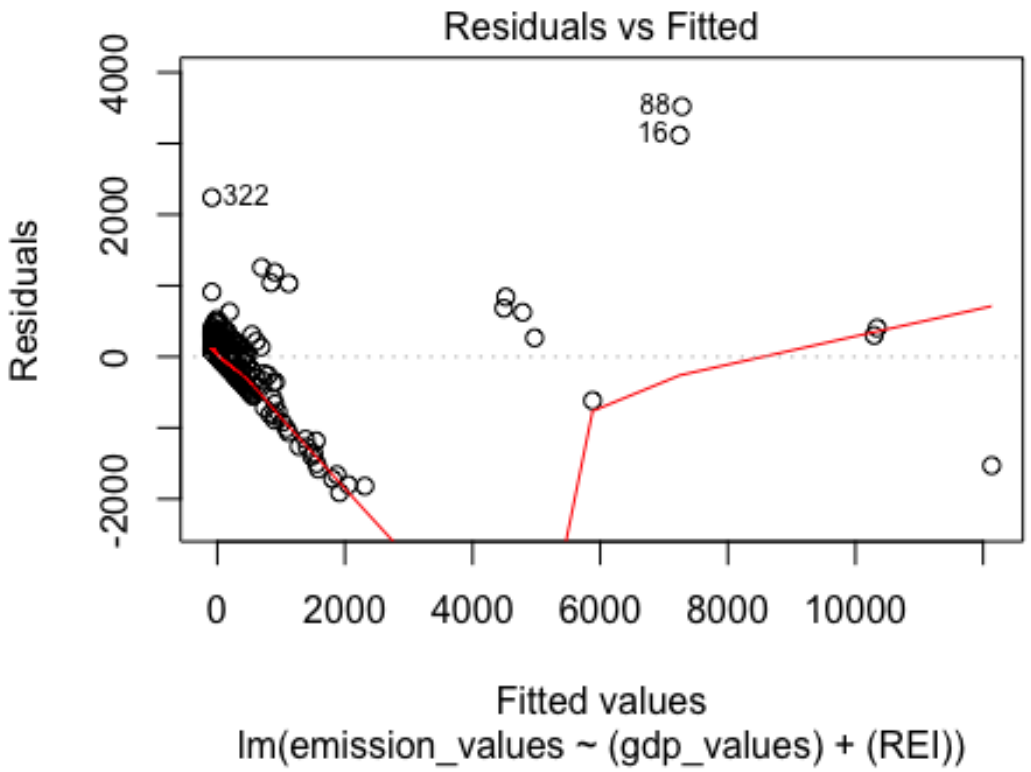
```
  mutate(lgdp=log(gdp_values))
renew_vs_gdp_emissions<-renew_vs_gdp_emissions%>%
  mutate(lemissions=log(emission_values))

all_data<-renew_vs_gdp_emissions
names(all_data)[1]<-"Country"
names(all_data)[3]<-"REI"
#all_data #the combined data

renew_vs_gdp_emissions2<-renew_vs_gdp_emissions%>%
  filter(emission_values<4000) #removing outliers

#renew_vs_gdp_emissions2


#plot for emissions and renewable investment values
ggplot(aes(y =emission_values, x =value),data=renew_vs_gdp_emissions2) + geom
_point()+geom_line()
```




```
#plot for emissions and gdp_values
ggplot(aes(y =emission_values, x =gdp_values),data=renew_vs_gdp_emissions2) +
geom_point()+geom_line()
```




```
#linear model for all regressors
lm_all<-lm(emission_values~(gdp_values)+(REI), data=all_data)
plot(lm_all, which=1)
```



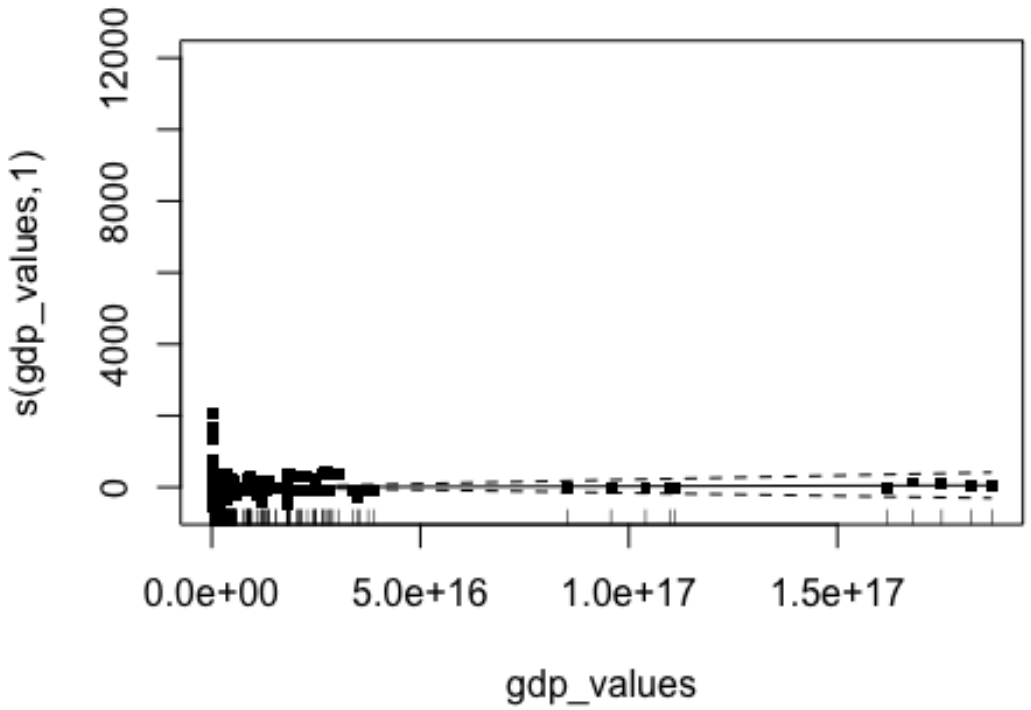

```
lm_all
```

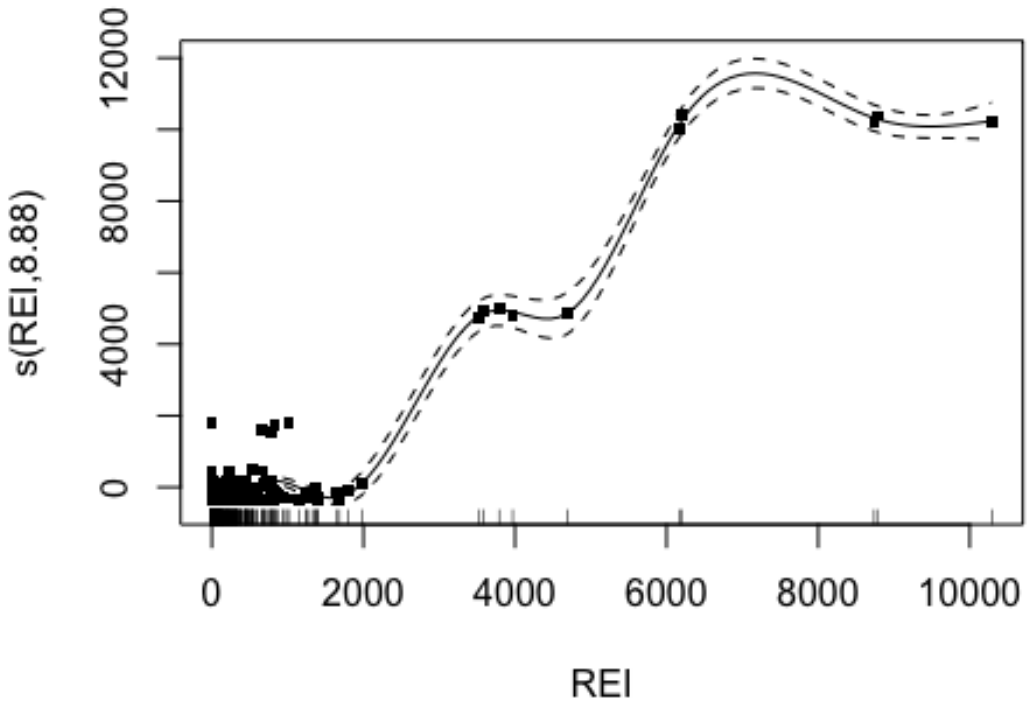
```
##
## Call:
## lm(formula = emission_values ~ (gdp_values) + (REI), data = all_data)
##
## Coefficients:
## (Intercept)   gdp_values          REI
##  -8.524e+01    2.127e-15    1.187e+00
```

```
summary(lm_all)
```

```
##
## Call:
## lm(formula = emission_values ~ (gdp_values) + (REI), data = all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1912.0   -79.3    55.9   119.9  3520.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.524e+01  2.828e+01  -3.014  0.00277 **
## gdp_values   2.127e-15  1.150e-15   1.850  0.06515 .
```

```
## REI          1.187e+00  2.562e-02  46.325  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 495.2 on 357 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8728
## F-statistic:  1232 on 2 and 357 DF,  p-value: < 2.2e-16

#gam model for all the regressors
gam_all<-gam(emission_values~s(gdp_values)+s(REI), data=all_data)
plot(gam_all, residuals=TRUE, cex=5)
```


s(gdp_values,1)
0
4000
8000
12000
0.0e+00
5.0e+16
1.0e+17
1.5e+17
gdp_values




```
gam_all
```

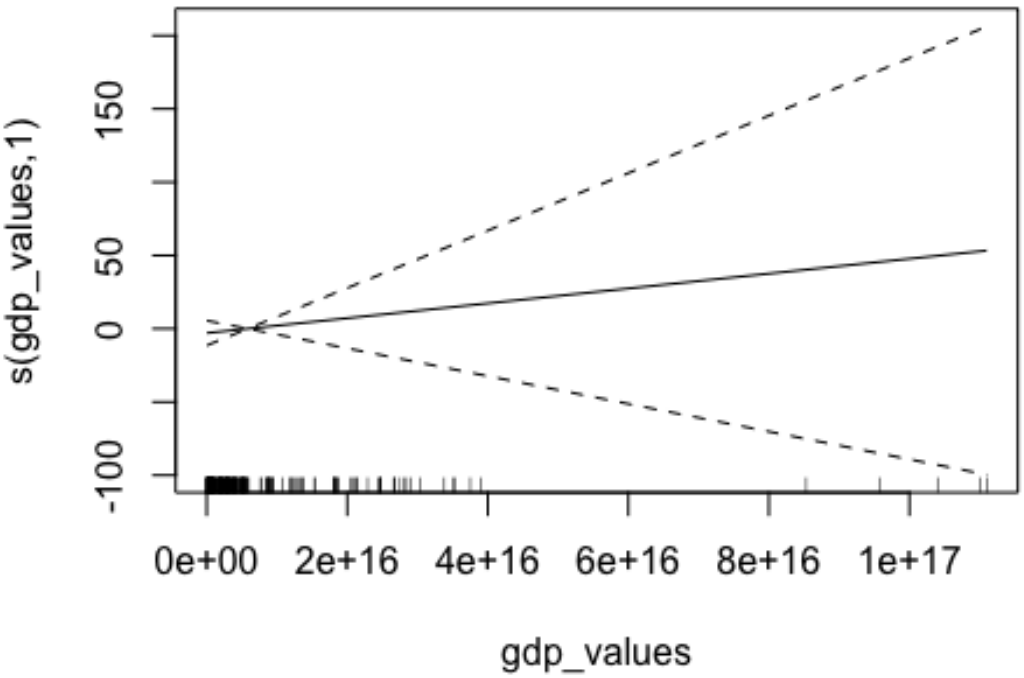
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## emission_values ~ s(gdp_values) + s(REI)
##
## Estimated degrees of freedom:
## 1.00 8.88  total = 10.88
##
## GCV score: 69518.82
```

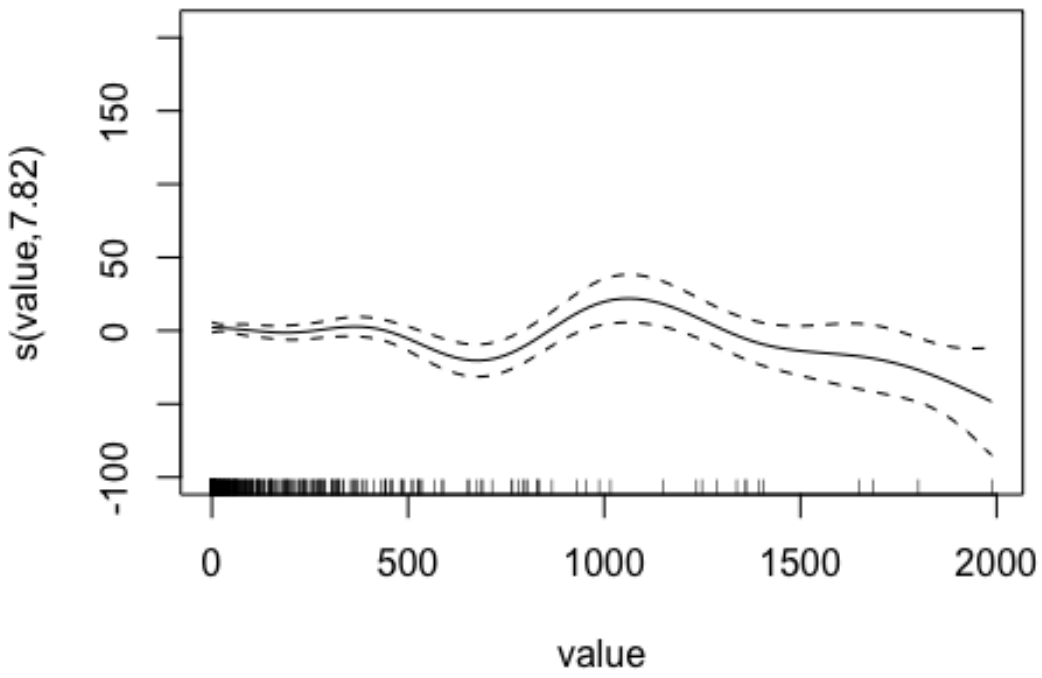
```
summary(gam_all)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## emission_values ~ s(gdp_values) + s(REI)
##
## Parametric coefficients:
```

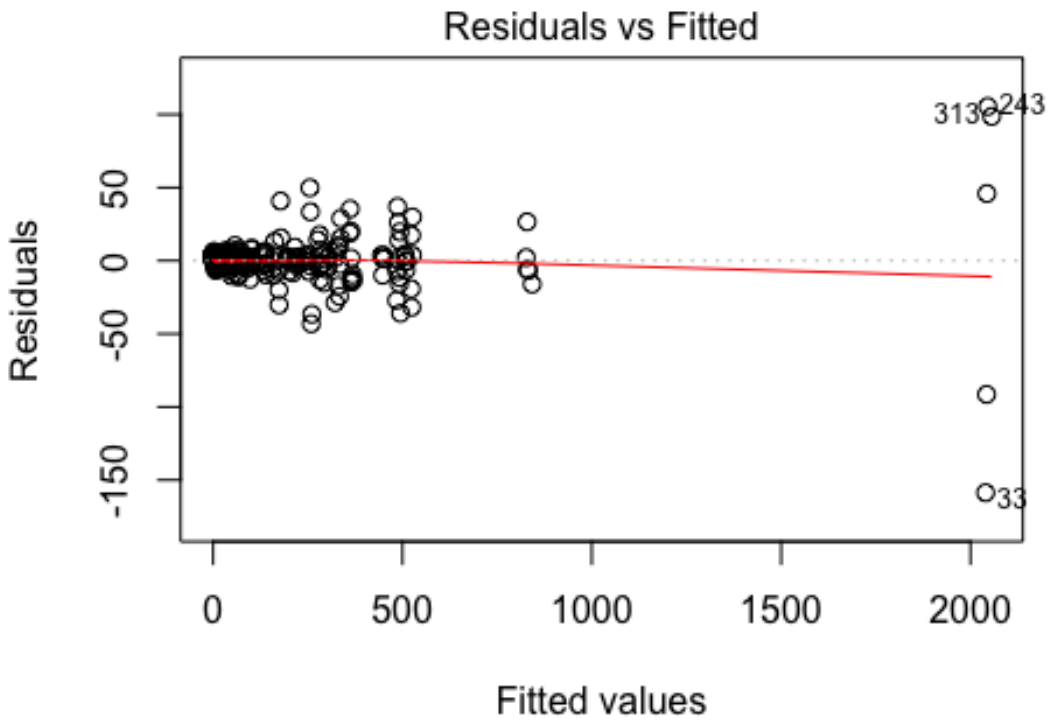
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   367.63      13.68   26.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                 edf Ref.df      F p-value
## s(gdp_values) 1.000  1.000   0.09   0.764
## s(REI)        8.885  8.995 972.90  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.965   Deviance explained = 96.6%
## GCV =  69519  Scale est. = 67417     n = 360
```

```
#Possible linear and log combinations
lm_log_gdp<-lm(emission_values ~ log(gdp_values)+(value), data=renew_vs_gdp_e
missions2)
#plot(lm_log_gdp, which=1)


lm_invest<-lm(emission_values ~ log(value+0.01)+ (gdp_values), data=renew_vs_
gdp_emissions2)
#plot(lm_invest,which=1)


lm_log_all<-lm(emission_values~lgdp+log(value+0.01), data=renew_vs_gdp_emissi
ons2)
#plot(lm_log_all,which=1)


lm_log_depend<-lm(lemissions~(gdp_values)+log(value+0.001), data=renew_vs_gdp
_emissions2)
#plot(lm_log_depend,which=1)


#GAM model with factors
gam_fac<-gam(emission_values~factor(id)+factor(year)+s(gdp_values)+s(value),d
ata=renew_vs_gdp_emissions2)
plot(gam_fac, residuals=FALSE)
```


s(gdp_values,1)
150
50
0
-100
0e+00
2e+16
4e+16
6e+16
8e+16
1e+17
gdp_values

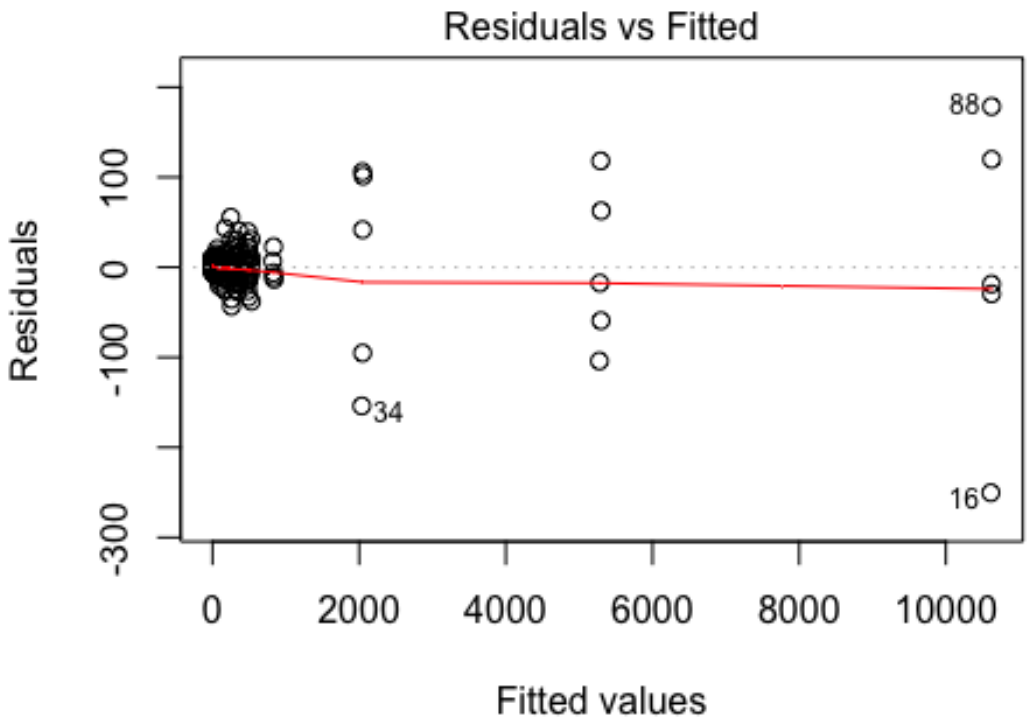



```
#the best model used with removing outlier China
lm_accurate<-lm(emission_values~factor(id)+factor(year)+(gdp_values)+log(valu
e+0.001),data=renew_vs_gdp_emissions2)
plot(lm_accurate, which=1)
```



ı(emission_values ~ factor(id) + factor(year) + (gdp_values) + log(va

```
#The best model used without removing the outlier china
lm_accurate2<-lm(emission_values~factor(Country)+factor(year)+(gdp_values)+lo
g(REI+0.001),data=all_data)
plot(lm_accurate2, which=1)
```




```
#single factor models , applicable for cross validation
lm_facone<-lm(emission_values~factor(year)+(gdp_values)+log(value+0.001),data
=renew_vs_gdp_emissions2)
#plot(lm_facone,which=1)

lm_facid<-lm(emission_values~factor(id)+(gdp_values)+log(value+0.001),data=re
new_vs_gdp_emissions2)
#plot(lm_facid,which=1)


summary(lm_accurate)

##
## Call:
## lm(formula = emission_values ~ factor(id) + factor(year) + (gdp_values) +
##     log(value + 0.001), data = renew_vs_gdp_emissions2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -158.640   -2.566    0.144    2.465  105.461
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)              6.063e+00  8.187e+00   0.741 0.459542
## factor(id)Algeria        1.254e+02  1.134e+01  11.063  < 2e-16 ***
## factor(id)Argentina      1.874e+02  1.188e+01  15.773  < 2e-16 ***
## factor(id)Austria        5.818e+01  1.158e+01   5.022 9.21e-07 ***
## factor(id)Bangladesh     6.154e+01  1.127e+01   5.460 1.07e-07 ***
## factor(id)Belarus        5.182e+01  1.127e+01   4.600 6.47e-06 ***
## factor(id)Belgium        1.216e+02  1.175e+01  10.353  < 2e-16 ***
## factor(id)Bhutan        -8.188e+00  1.121e+01  -0.731 0.465642
## factor(id)Bolivia        8.826e+00  1.120e+01   0.788 0.431481
## factor(id)Brazil         5.019e+02  1.936e+01  25.929  < 2e-16 ***
## factor(id)Burkina Faso  -4.908e+00  1.120e+01  -0.438 0.661568
## factor(id)Burundi       -8.302e+00  1.120e+01  -0.741 0.459260
## factor(id)Cambodia      -1.211e+00  1.120e+01  -0.108 0.913938
## factor(id)Cameroon      -7.887e-01  1.120e+01  -0.070 0.943915
## factor(id)Chile          5.921e+01  1.658e+01   3.571 0.000421 ***
## factor(id)Colombia      -8.922e+00  7.411e+01  -0.120 0.904264
## factor(id)Costa Rica    -1.564e+00  1.148e+01  -0.136 0.891714
## factor(id)Cuba           1.783e+01  1.122e+01   1.589 0.113263
## factor(id)Denmark        2.966e+01  1.122e+01   2.644 0.008668 **
## factor(id)Dominican Rep  8.541e+00  1.150e+01   0.743 0.458220
## factor(id)Ecuador        3.254e+01  1.123e+01   2.897 0.004069 **
## factor(id)Egypt          2.166e+02  1.121e+01  19.310  < 2e-16 ***
## factor(id)El Salvador   -2.986e+00  1.140e+01  -0.262 0.793479
## factor(id)Ethiopia       3.561e+00  1.123e+01   0.317 0.751512
## factor(id)Finland        3.917e+01  1.122e+01   3.491 0.000560 ***
## factor(id)France         3.574e+02  1.140e+01  31.360  < 2e-16 ***
## factor(id)Georgia       -1.970e+01  2.210e+01  -0.891 0.373696
## factor(id)Germany        8.247e+02  1.120e+01  73.602  < 2e-16 ***
## factor(id)Ghana         -2.418e+01  2.837e+01  -0.852 0.394875
## factor(id)Guatemala      3.990e+00  1.122e+01   0.356 0.722356
## factor(id)Haiti         -5.942e+00  1.120e+01  -0.530 0.596281
## factor(id)Honduras       2.355e+00  1.124e+01   0.210 0.834208
## factor(id)India          2.038e+03  1.121e+01 181.804  < 2e-16 ***
## factor(id)Indonesia      4.657e+02  1.832e+01  25.416  < 2e-16 ***
## factor(id)Italy          3.520e+02  1.290e+01  27.280  < 2e-16 ***
## factor(id)Jamaica       -1.704e+01  1.833e+01  -0.930 0.353411
## factor(id)Jordan         1.667e+01  1.120e+01   1.488 0.137984
## factor(id)Kazakhstan     2.723e+02  1.120e+01  24.313  < 2e-16 ***
## factor(id)Kenya          7.260e+00  1.132e+01   0.641 0.521815
## factor(id)Malawi        -8.152e+00  1.121e+01  -0.727 0.467706
## factor(id)Malaysia       2.055e+02  1.128e+01  18.219  < 2e-16 ***
## factor(id)Mali          -7.303e+00  1.140e+01  -0.641 0.522259
## factor(id)Mexico         4.422e+02  1.125e+01  39.310  < 2e-16 ***
## factor(id)Morocco        3.587e+01  1.412e+01   2.541 0.011618 *
## factor(id)Mozambique    -1.036e+00  1.122e+01  -0.092 0.926529
## factor(id)Nepal         -2.569e+00  1.122e+01  -0.229 0.819017
## factor(id)Netherlands    2.363e+02  1.132e+01  20.876  < 2e-16 ***
## factor(id)Nicaragua     -8.574e+00  1.265e+01  -0.678 0.498443
## factor(id)Nigeria        9.062e+01  1.122e+01   8.080 2.08e-14 ***
## factor(id)Norway         3.236e+01  1.172e+01   2.761 0.006153 **
```

```
## factor(id)Pakistan       1.509e+02  1.170e+01  12.892  < 2e-16 ***
## factor(id)Palestine     -8.420e+00  1.134e+01  -0.742 0.458597
## factor(id)Peru           4.265e+01  1.130e+01   3.773 0.000197 ***
## factor(id)Philippines    8.964e+01  1.136e+01   7.893 7.10e-14 ***
## factor(id)Poland         2.898e+02  1.174e+01  24.684  < 2e-16 ***
## factor(id)Senegal       -7.979e-01  1.120e+01  -0.071 0.943251
## factor(id)Serbia         3.692e+01  1.120e+01   3.296 0.001110 **
## factor(id)South Africa   4.992e+02  1.149e+01  43.434  < 2e-16 ***
## factor(id)Spain          2.679e+02  1.456e+01  18.395  < 2e-16 ***
## factor(id)Sri Lanka      1.146e+01  1.121e+01   1.022 0.307564
## factor(id)Sweden         3.959e+01  1.183e+01   3.348 0.000929 ***
## factor(id)Tajikistan    -5.084e+00  1.123e+01  -0.452 0.651281
## factor(id)Tanzania       3.950e+00  1.121e+01   0.352 0.724782
## factor(id)Thailand       3.158e+02  1.155e+01  27.356  < 2e-16 ***
## factor(id)Turkey         3.219e+02  1.291e+01  24.935  < 2e-16 ***
## factor(id)Uganda        -2.297e+00  1.124e+01  -0.204 0.838219
## factor(id)UK             2.515e+02  1.134e+01  22.184  < 2e-16 ***
## factor(id)Ukraine        4.613e+02  2.307e+01  19.997  < 2e-16 ***
## factor(id)Viet Nam       1.670e+02  1.126e+01  14.823  < 2e-16 ***
## factor(id)Zambia        -4.697e+00  1.120e+01  -0.419 0.675221
## factor(year)2013         2.369e+00  3.002e+00   0.789 0.430722
## factor(year)2014         3.174e+00  3.022e+00   1.050 0.294481
## factor(year)2015         5.711e+00  2.999e+00   1.904 0.057936 .
## factor(year)2016         7.007e+00  3.004e+00   2.333 0.020394 *
## gdp_values               8.039e-16  7.254e-16   1.108 0.268747
## log(value + 0.001)      -6.611e-01  3.851e-01  -1.717 0.087166 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.71 on 274 degrees of freedom
## Multiple R-squared:  0.997,  Adjusted R-squared:  0.9961
## F-statistic:  1204 on 75 and 274 DF,  p-value: < 2.2e-16

#cross validation1: (Removed the outlier point China)

benchmark <- c()
gam_error <- c()

cv<- function(folds = 5){
  # Construct the folds
  fold_num <- rep(1:folds, length.out = nrow(all_data))
  fold_ran <- sample(fold_num)

for(i in 1:folds){
  # Construct training and test sets
  train <- all_data[fold_ran != i, ]
  test <- all_data[fold_ran == i, ]

  # Fit models to training data
  benchmark_model <- lm(emission_values ~ factor(year)+gdp_values + log(REI+0
```

```
.01), data = train)

  gam_cross <- gam(emission_values ~ s(gdp_values) + s(REI), data = train)

  # Test error
    bench_test <- (log(test$emission_values) - predict(benchmark_model, newda
ta = test))^2
    gam_test <- (log(test$emission_values) - predict(gam_cross, newdata = tes
t))^2

  # Store results
    benchmark <- append(benchmark, bench_test)
    gam_error <- append(gam_error, gam_test)

  }

  # Test rmse
  rmse_benchmark <- sqrt(sum(benchmark)/(length(benchmark)))
  rmse_gam <- sqrt(sum(gam_error)/(length(gam_error)))

  return(list(rmse_gam,rmse_benchmark))
}

# Replicate the 5-fold cross-validation 100 times
  cvs <- replicate(100, cv())

# Pull out the average rMSE for each model
  cv.gam <- mean(sapply(cvs[1, ], mean))
  cv.bench <- mean(sapply(cvs[2, ], mean))

cv.bench

## [1] 669.9501

cv.gam

## [1] 1431.767

#Cross Validation2
train_beta<-all_data[1:288,]
test_beta<-all_data[289:360,]

train_model<-lm(emission_values ~ factor(Country)+ gdp_values + log(REI+0.01)
, data = train_beta)

test_pred<-predict(train_model, newdata=test_beta)

residual<-test_beta$emission_values-test_pred

rmse_test<-mean(residual^2)
```

```
print(rmse_test)
```

```
## [1] 769.0587
```

Including Plots

You can also embed plots, for example: