**Usman Khan, Edwin Aguirre, Maruthi Mutnuri**

**Data 602 -Statistical Data Analysis of GDP Growth Impact on CO2 emissions**

Jim Stallard

1. *Purpose:*

Post-industrialization era has been marked with increased GDP growth for countries as well as an improved standard of living for the people. However, throughout the past years, people have not realized the importance of climate and the implications of climatic changes on our society. With increasing globalization, drastic visible climatic changes – and their impact on the society— have been observed that has led to people realizing the gravity of the situation at hand. Globalization leads to countries increasing their production in order to meet the local demand as well as the international demand. Such increased global demands have led to countries using more fossil fuel run machinery, that has led to increased air pollution that has started affecting our ecosystem.

Many of the sources of outdoor air pollution are also sources of high CO2 emissions. A major source of carbon dioxide is the burning of fossil fuel—mostly by the energy and transport sectors. In regions that are prone to temperature and precipitation pattern changes due to climate change, it is very much likely that the frequency and severity of forest fires will also increase, destroying the ecosystem and the habitants, while simultaneously releasing more air pollutants.[1] Household air pollution from cooking with solid fuels has accounted for 3.8 millions deaths in the year 2016.[2] Similar to household pollution, outdoor air pollution in both cities and rural areas was estimated to cause 4.2 million premature deaths worldwide in the year 2016.[3]

1. *Dataset:*

We are using four datasets for our study, annual-co2-emissions-per-country, fossil-fuel-consumption-by-fuel-type, greenhouse-gas-emissions-by-gas, maddison-data-gdp-per-capita-in-2011us, modern-renewable-energy-consumption, and CO2-by-source.

1. annual-co-emissions-by-region: This dataset has data on the Annual CO2 emissions in tonnes by entity, code (ISO-code) and year. country column is comprised of world, continents and regions along with all the countries.
2. fossil-fuel-consumption-by-fuel-type: This dataset has data on Fossil fuel consumption namely Oil, Gas and Coal in terawatt-hours by entity, code (ISO-code) and year. Entity is comprised of world, continents and regions along with all the countries.
3. maddison-data-gdp-per-capita-in-2011us: This dataset has data on GDP in 2011 USD by entity, code (ISO-code) and year. Entity is comprised of world, continents and regions along with all the countries.
4. modern-renewable-energy-consumption: This dataset has data on various renewable energy consumption namely Solar, Wind, Hydro, Other renewables (modern biofuels; geothermal; wave & tidal) in terawatt-hours by entity, code (ISO-code) and year. Entity is comprised of world, continents and regions along with all the countries.

For our study, we will be looking at the G20 countries from 1985 to 2015. Our parameter for choosing the G20 countries is based on the fact that these countries contribute to 80% of the world trade with high GDPs and high consumption of fossil fuels.[4] The dataset used is in .csv format. The data itself is considered as structured data and can also be considered as tabular data. We’ve filtered the original dataset to only include the G20 countries. Since these countries are considered as the most advanced economies in the world, we can extrapolate the analysis/visualization performed in this project to rest of the countries in world. The G20 countries are:  Argentina, Australia, Brazil, Canada, China, European Union, France, Germany, India, Indonesia, Italy, Japan, Mexico, Russia, Saudi Arabia, South Africa, South Korea, Turkey, United Kingdom and the United States. Limitations to the availability of data has led us to exclude the European Union. Furthermore, we’re using data for only 30 years (from 1985 until 2015).

Lastly, we obtained the data-in-use from *(https://ourworldindata.org/)* which is a free open source website.

1. *Topics:*

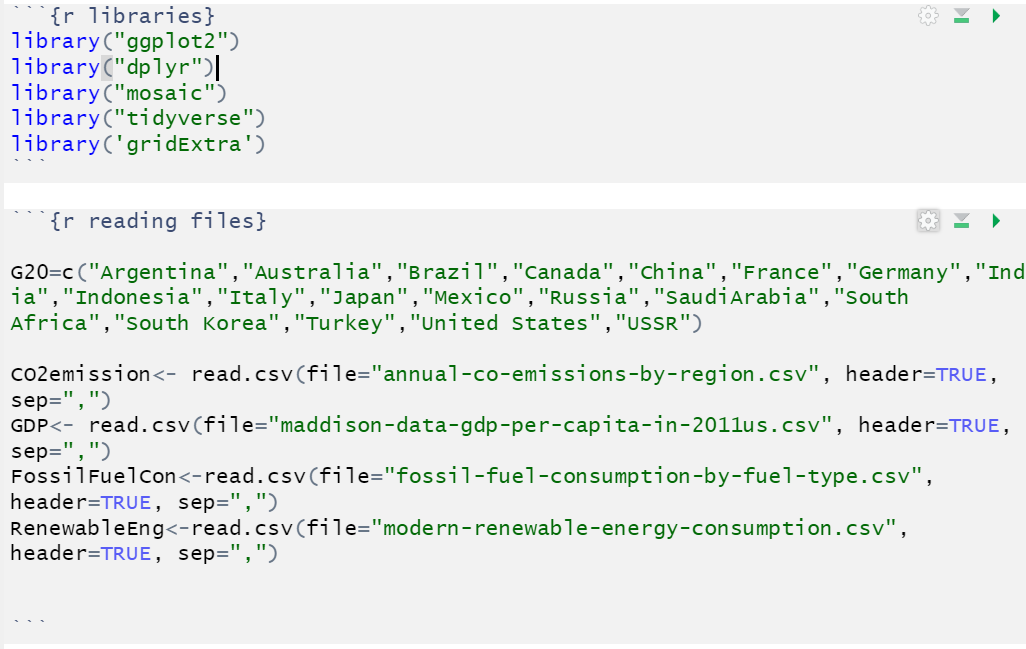
One of the main concerns faced globally is the amount of fossil fuels burnt and consequently the amount of air pollutants released into the atmosphere by burning these fossil fuels. Many types of health ailments and consequently deaths are linked to the extent of exposure to air pollutants. As mentioned earlier, the global economy is majorly influenced by the G20 nations which are highly developed economies in the world and to meet their expanding need for energy, these countries are also the largest consumers of fossil fuels. Hence, we want to analyze how their economic growth influences the consumption of fossil fuels and air pollutants like CO and CO2 emissions. Another point of interest in this study is to measure how much of renewable resources the G20 countries are using for their energy needs and has this decreased the amount of air pollutants being released into the atmosphere. Below are some of the areas that we will be targeting:

* How did the GDP grow in G20 countries between 1985 and 2010? Is there any impact of this on the fossil fuel consumption and air pollutant emissions by these countries?
* What is the correlation between renewable sources of energy used and CO2 emissions in G20 countries between time period 1985 and 2010? The data from 2010 to 2015 will be used to compare the prediction model and the true values.

*Statistical Method to be used:*

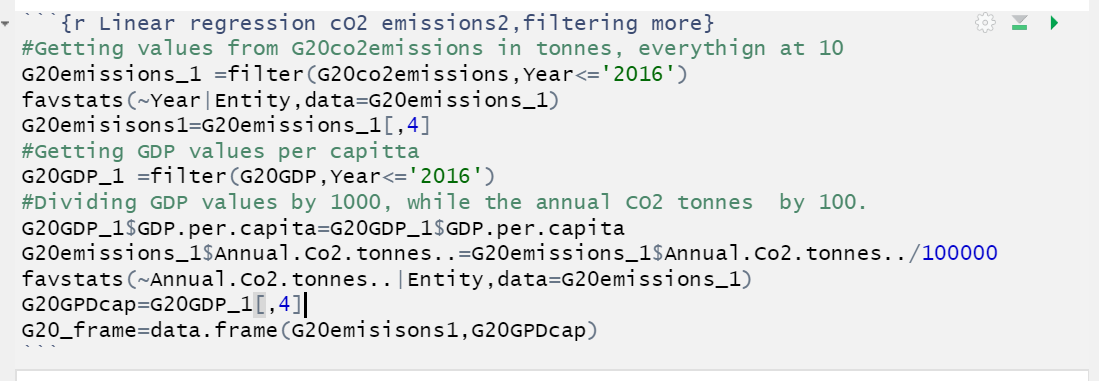
1. Use the Hypothesis testing method to come up with a Null Hypothesis, and an Alternate Hypothesis to test the impact of the GDP change on the CO2 emissions by the G20 countries the time period of 1985 to 2015.
2. Use Regression statistical method to analyze the correlations between the renewable sources of energy used and CO2 emissions.
3. *Data: Wrangling:*

Creating a Vector called G20 countries, which has all the G20 countries that we’re interested in our analysis. Then, four csv files are read, which have all the datasets that we need. The data set that I’m interested in is called Co2 emissions and GDP (It can be observed below)



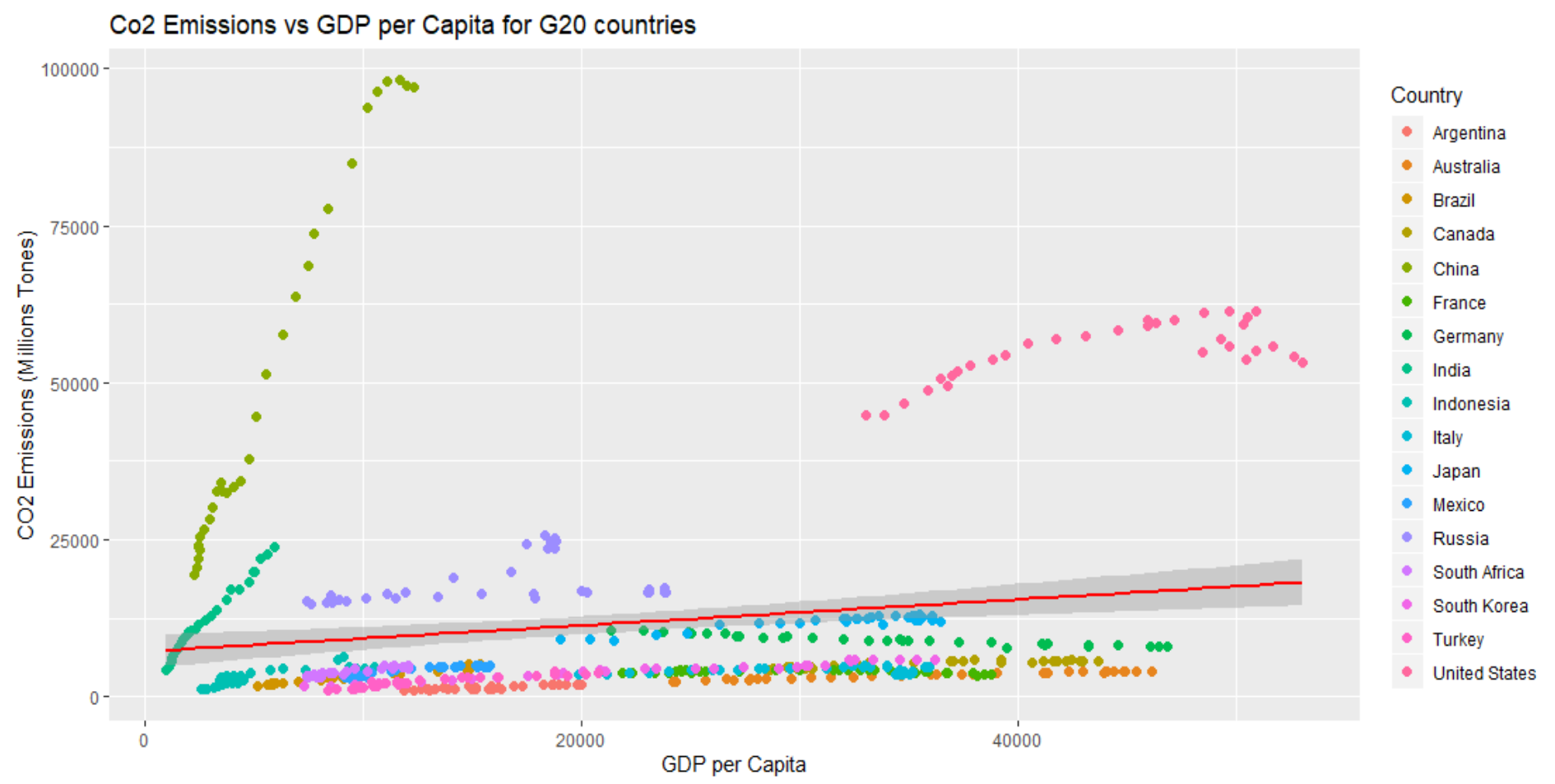
Then, we create a new temporal data frame that filters the previously created data frame (CO2emissions) by the year and entity column. This is performed by doing a for loop that goes from 1 to the length of the previously defined G20 countries vector. This temporal data frame is then added the actual data frame (which uses rbind function). This same process is repeated for both the CO2 emissions and GDP per capita. 

We then, create a new data frame which filters everything by the years that are equal to less than 2016, and divided the column of annual CO2 tonnes by 100,000.Finally, we obtain a G20\_frame, which has two columns, the first one is G20 emissions1, and the second one is G20GPDcap.

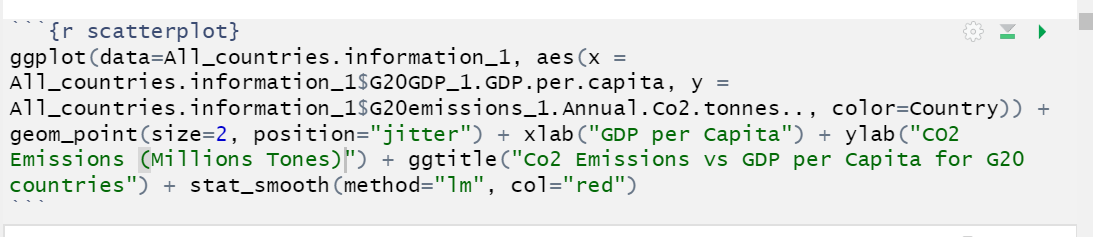


1. *Bivariate Data, and its relationship(correlation):*

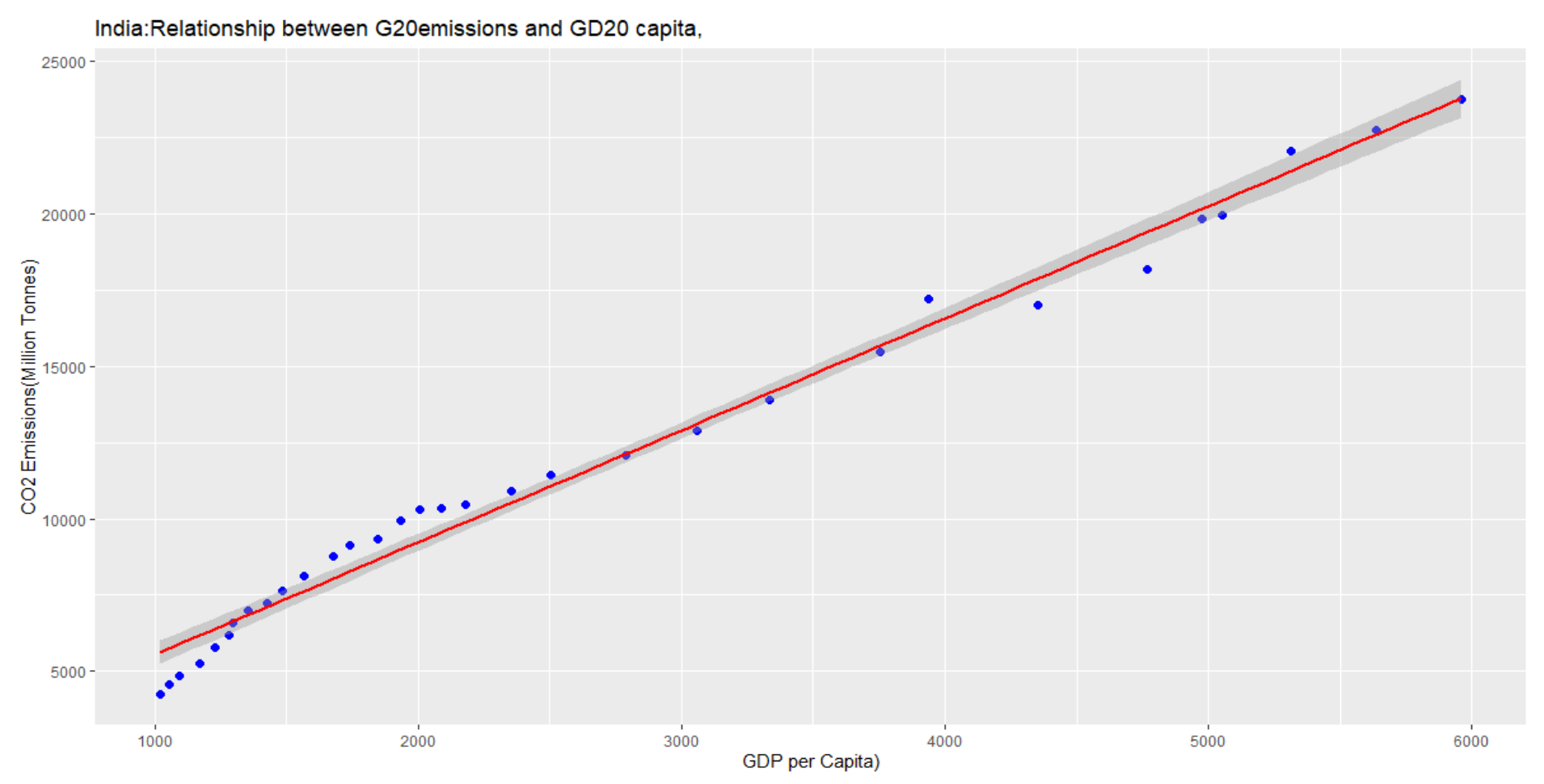
The data that we’re interested in this project is CO2 emissions and GDP per capita. Then, we defined x to be the GDP per capita, while the is CO2 variable is the y axis. The scatter plot of all the countries is given by :



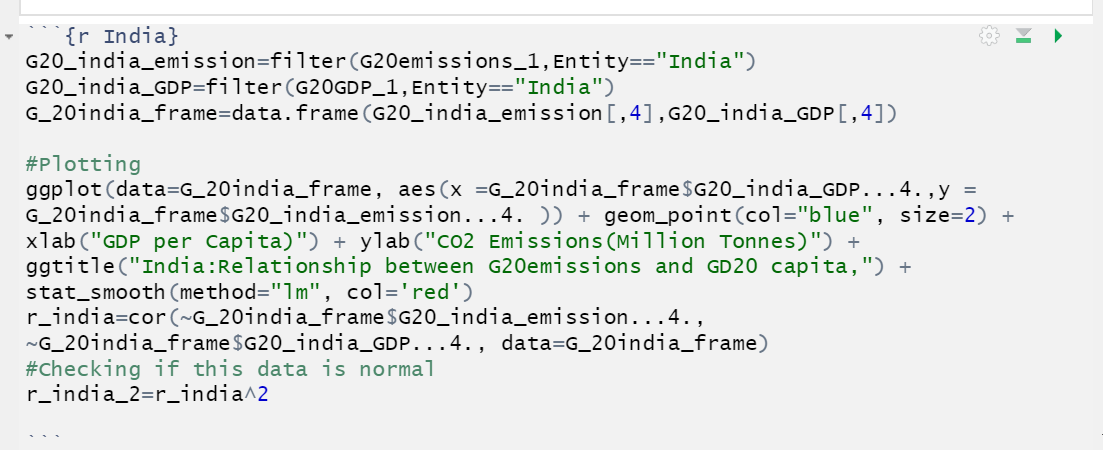
The code in R can be seen from the snippet below:



It can be seen from the scatterplot that it is not feasible to obtain a relationship between CO2 emissions and GDP per capita. Therefore, we looked at the variable GDP per capita (variable x) and the variable CO2 (variable y) individually per country. This can be seen underneath:



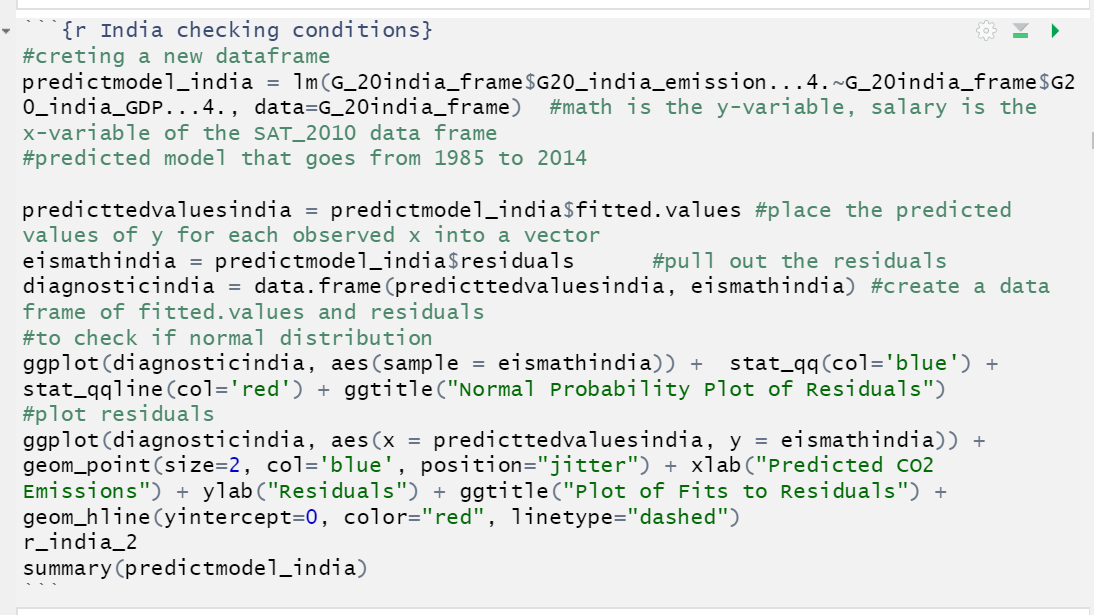
The code which produces the plot is given by:



1. *Modeling the Relationship between GDP per Capita and Co2 Emissions:*

The general probabilistic model used in this dataset is given by:

Yi=A+ (B∗Xi) + ei for i=1,2,⋯,n

In order to produce a linear model from the datasets, we use a linear model function in R. This can be observed below:  


From the training data, we have an estimate of the model is:

**CO2i = 1872.581507 + 3.676635\*GDP per Capitai + ei**

**where i = years 1985,…2016**

**SSE=∑ (Yi−Yˆi)2 where i = 1,2,3….n**

*Se* = √(*SSE/n-*2)

**Se =** 751.1

For India, the standard Residual standard error is given by 751.1

1. *Quantifying the relationship between GDP per capita and Correlation Coefficient:*

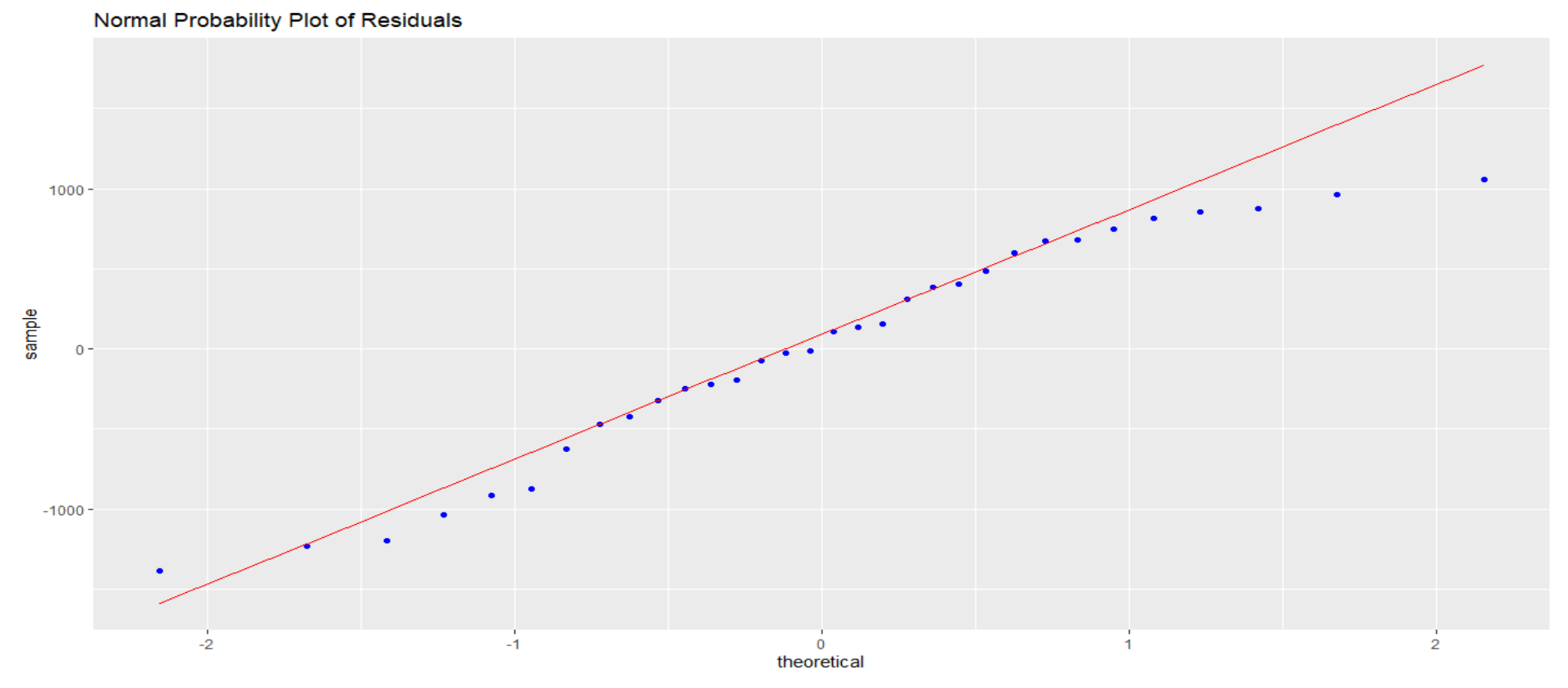
The correlation coefficient is calculated in this model (for India) was 0.9924497

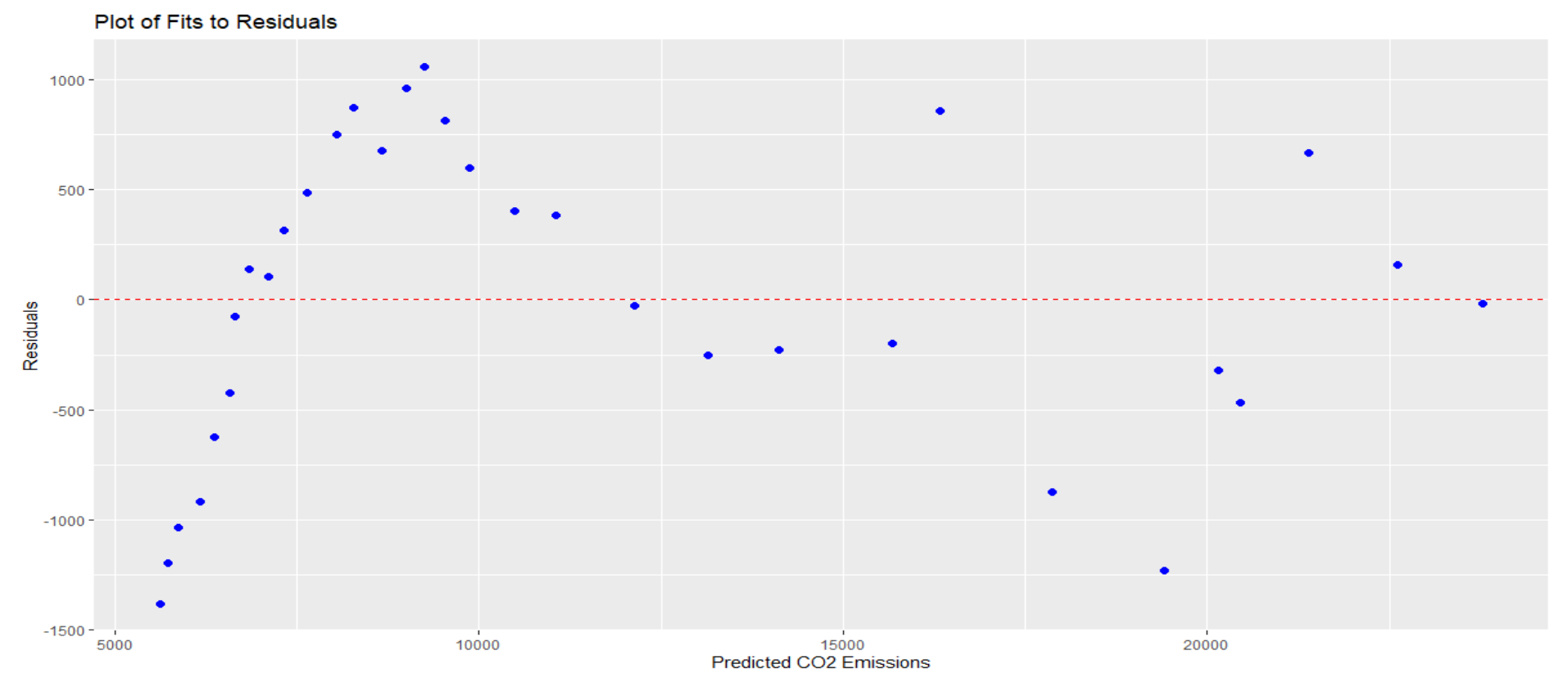
1. *Conditions of the model:*

For the model to be linear, two conditions must be met. These conditions are:

1. The y variable, must be Normally distributed with a mean μ and standard deviation of   
   σ.
2. For each distinct value of the x variable (the predictor variable), the y variable has the same standard deviation σ .

Condition one is referred, as the normality of the residuals of the condition, while the second one is called homoscedasticity.





1. *Coefficient of determination (***r2)***:*

The Coefficient of determination(r2), is a statistic that determines how well the statistical model obtained represents the data obtained. Lastly, this variable is expressed from 0 to 1. It can be obtained by squaring the coefficient of correlation. For India, the coefficient of determination is given by **0.9849563**.

**r2 = 0.9849563**

1. *Statistical Significance of Model:*

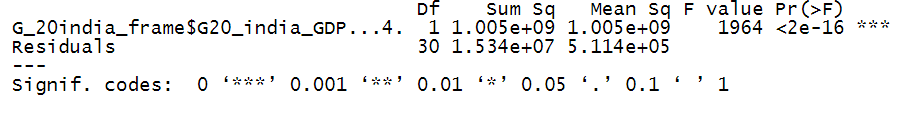
For our purposes, we want to know if the linear model is appropriate for the country of India. Therefore, our null hypothesis is the following:

**H0 = Y variable (CO2 emissions) can’t be expressed as a linear function of GDP per Capita**

While the alternate hypothesis is given by:

**HA = Y variable(C02 emissions) can be expressed a linear function of GDP per Capita.**

This hypothesis can, therefore, be tested by a **Fobs** test, which is given by an F distribution. The F-distribution is always positive (goes from 0 onwards), this test is the ratio of two chi squared distributions. The P value obtained from the model previously mentioned above can be seen below:



The P values is **2e-16** which is less than 0.05(default value of α), which means that we reject the null hypothesis and accept the Alternate hypothesis. Therefore, the **C02 emissions can be expressed as a linear function of GDP per Capita of that country**.

*11. Inferences about A and B:*

One can also make inferences about the intercept(A) and slope(B)of the model. For the intercept, the hypothesis can be given by:

**H0: A=0 (the mean value of Intercept A (CO2 emissions) is equal to 0 when GDP per capita = 0)**

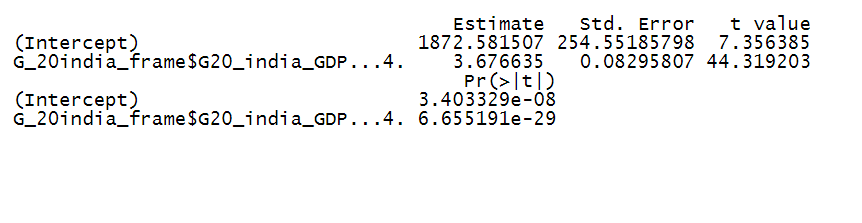
**HA: A ≠ 0(the mean value of Intercept A (CO2 emissions) is greater than 0 when GDP per capita = 0)**

Therefore, given the fact that the P value for A is **3.403329e-08**, then we reject the null hypothesis. **This means that the mean value of Intercept A is greater than 0 when GDP per capita=0.**

Also, for the slope (B), the hypothesis can be given by:

**H0: B = 0( CO2 cannot be expressed as a linear function of GDP per capita)**

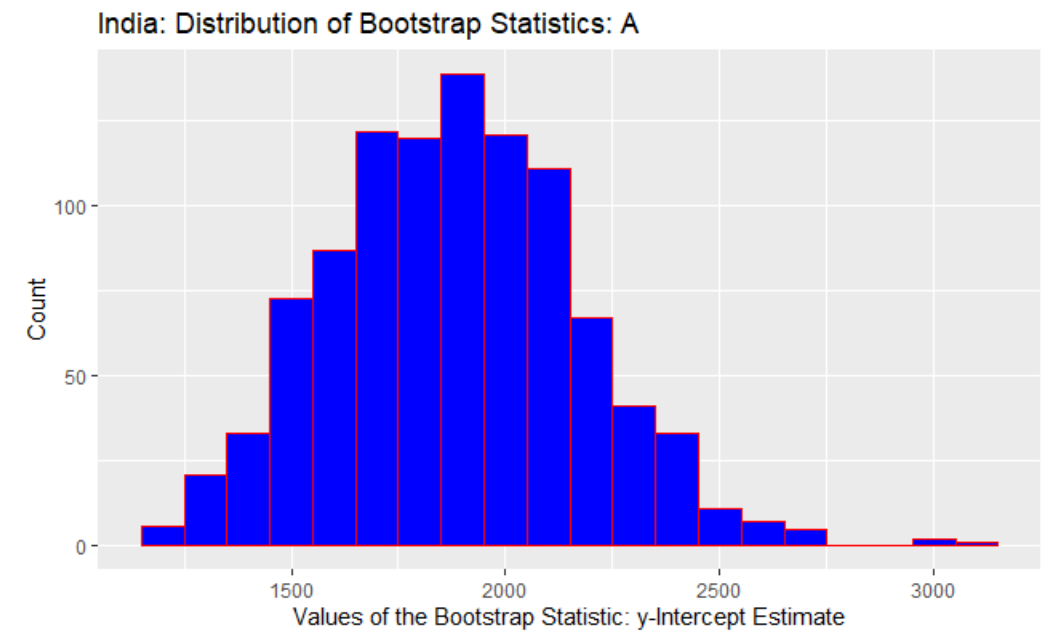
**HA: B ≠ 0(CO2 can be expressed as a linear function of GDP per capita)**

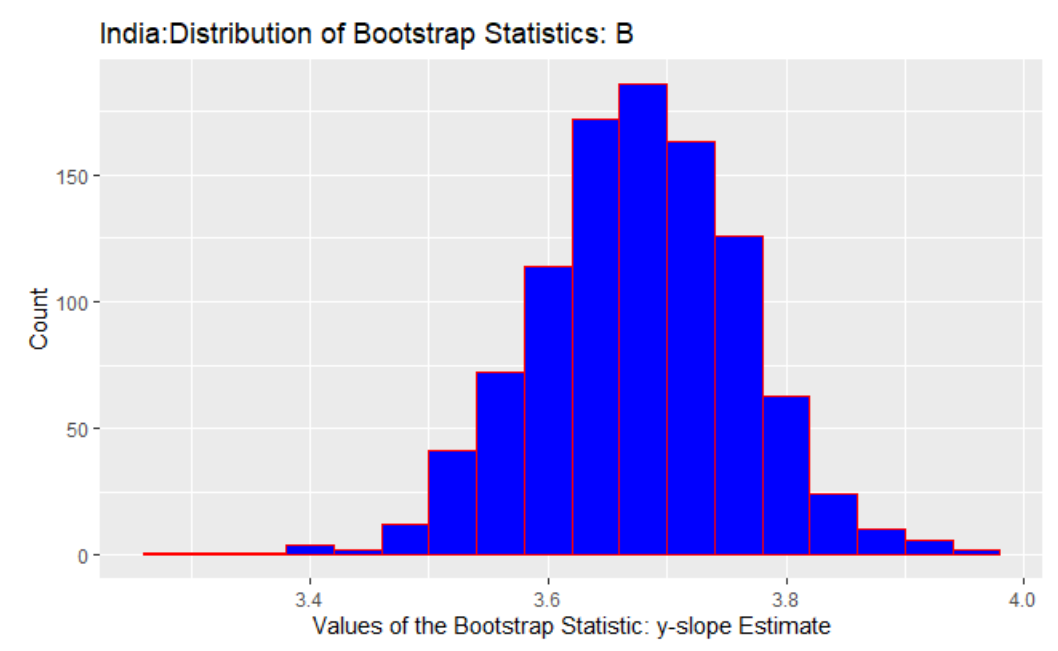


Lastly, the P value obtained for the B slope or rate of change is given by 6.65519e-29, which means that we reject the null hypothesis. **This mean that the slope(B) can be expressed, as a linear function of GDP per capita.**

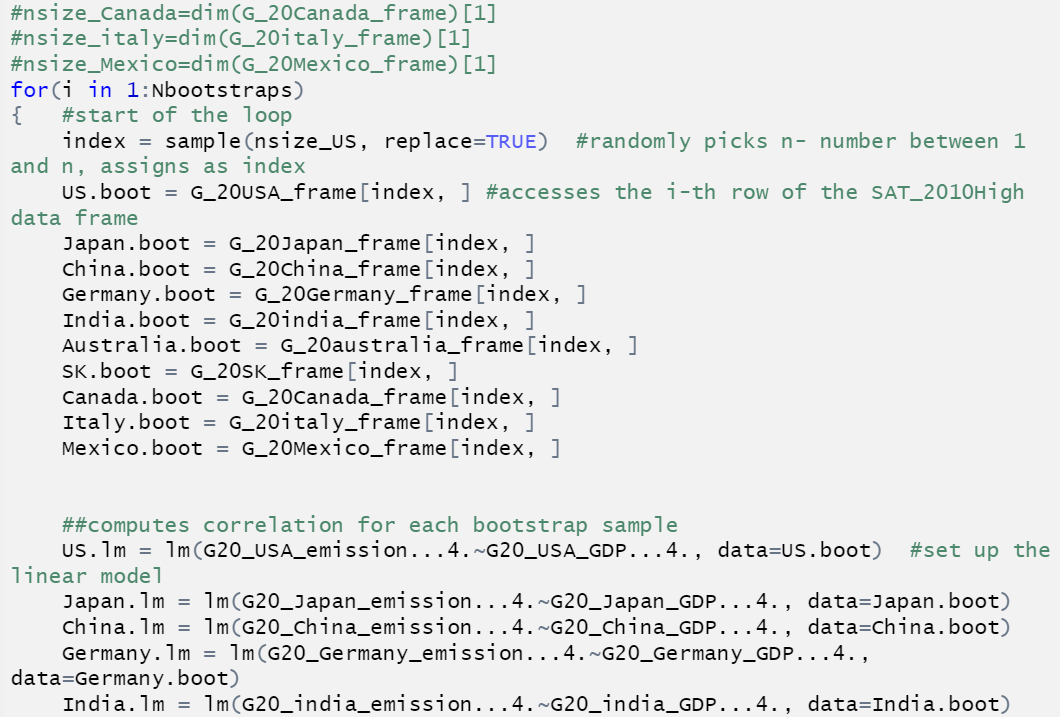
* 1. *Bootstrap Method:*

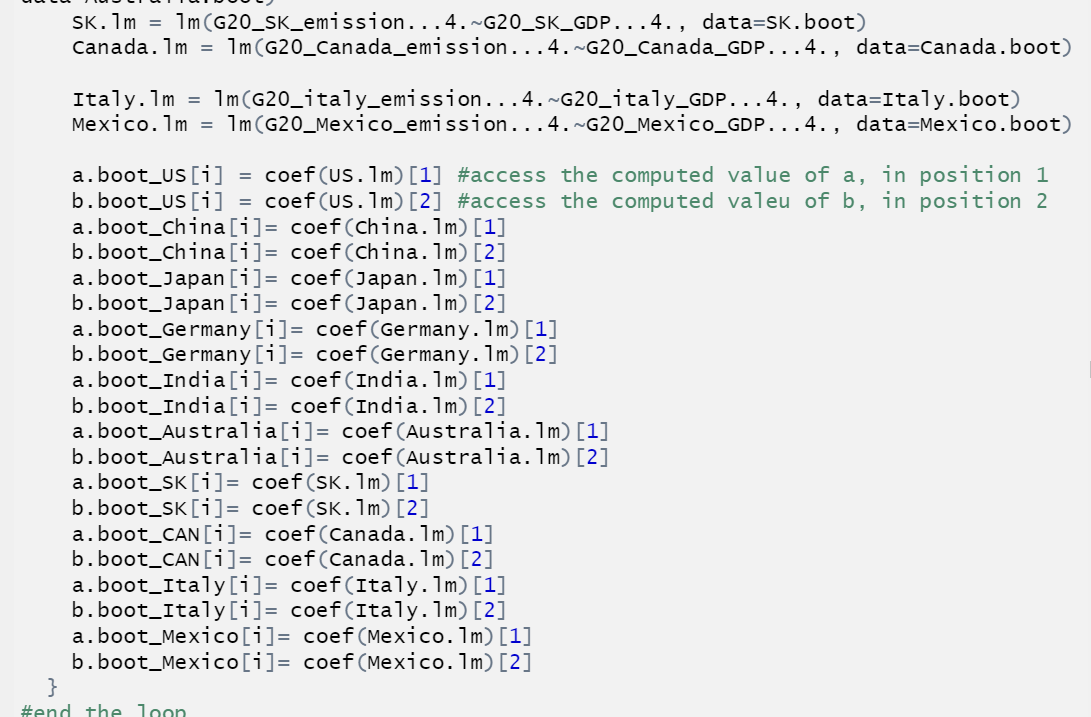
We can obtain a linear model from the data set using a parametric test; however, one can also obtain the linear model using the Bootstrap method. The bootstrap method basically uses the Law of large numbers as a bases, which states that after a large number of trials, the trials should be close to the expected value. The distribution of the A (intercept) and B(slope) can be seen below:

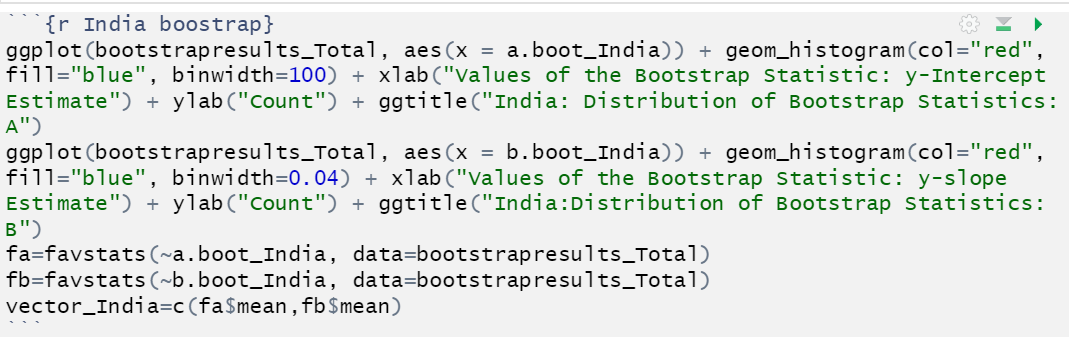




The code which produced these distributions, can be seen below:







From bootstrapping, we can obtain the model equation, which is given by:

**CO2i = 1881.036484 + 3.676151\*GDP per Capitai + ei where i = years 1985,…2015**

*Extrapolation and Comparison of both methods:*

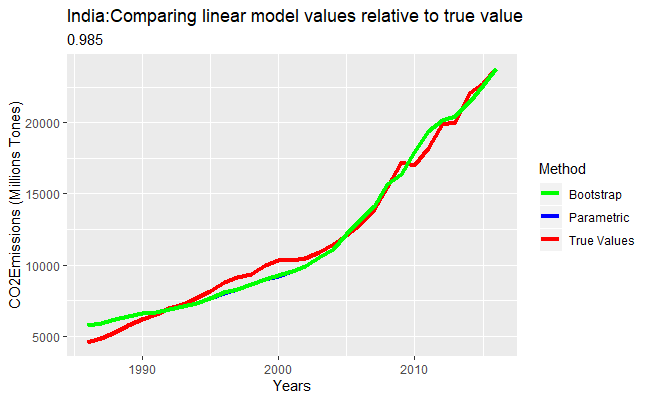
The equation obtain from the parametric method is given by :

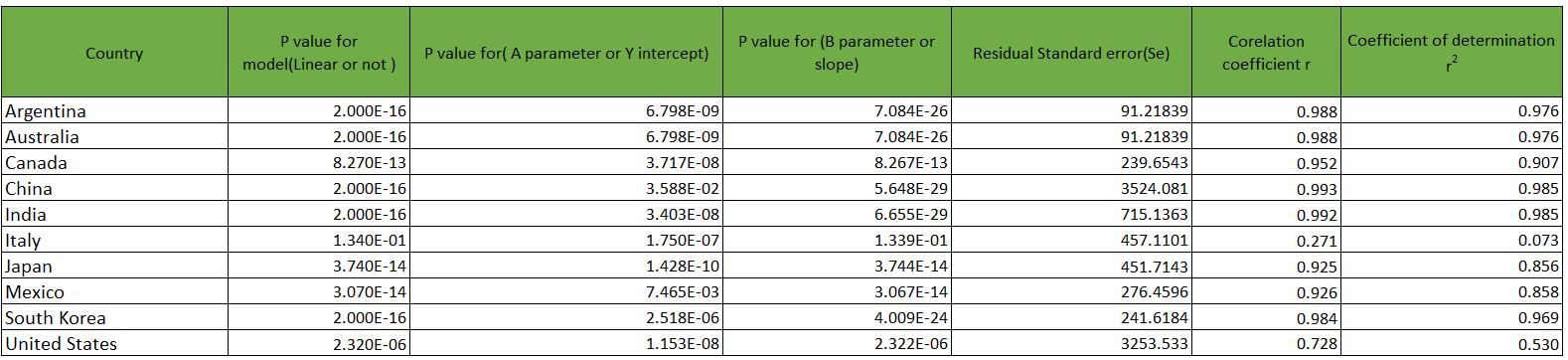
**CO2i= 1872.581507+3.676635\*GDP per Capitai + ei where i = years 1985,…2015**

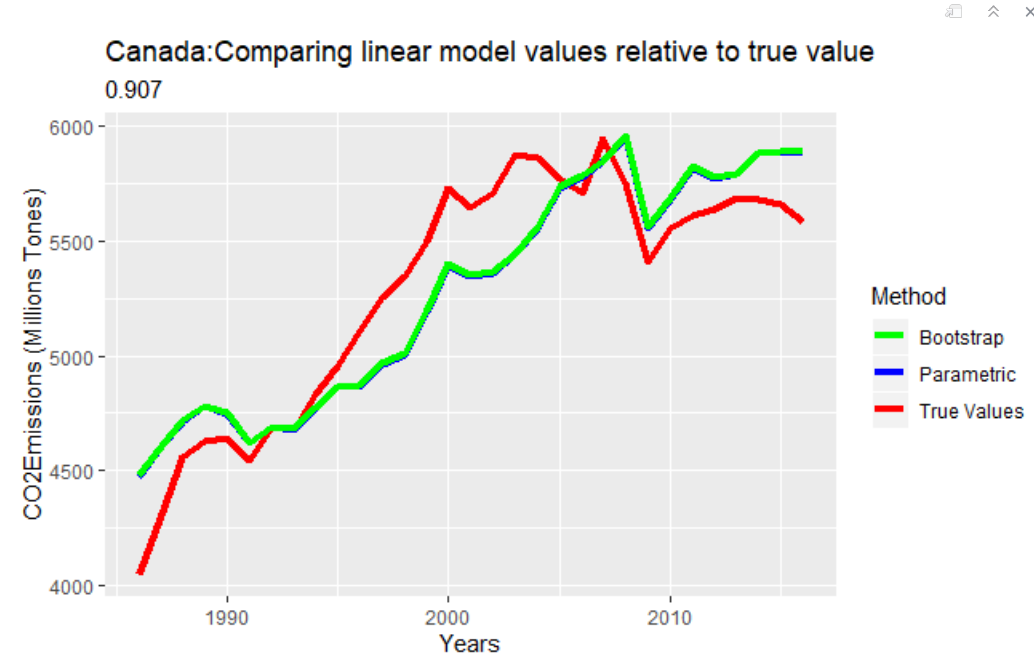
While the equation obtained from the bootstrap method is given by:

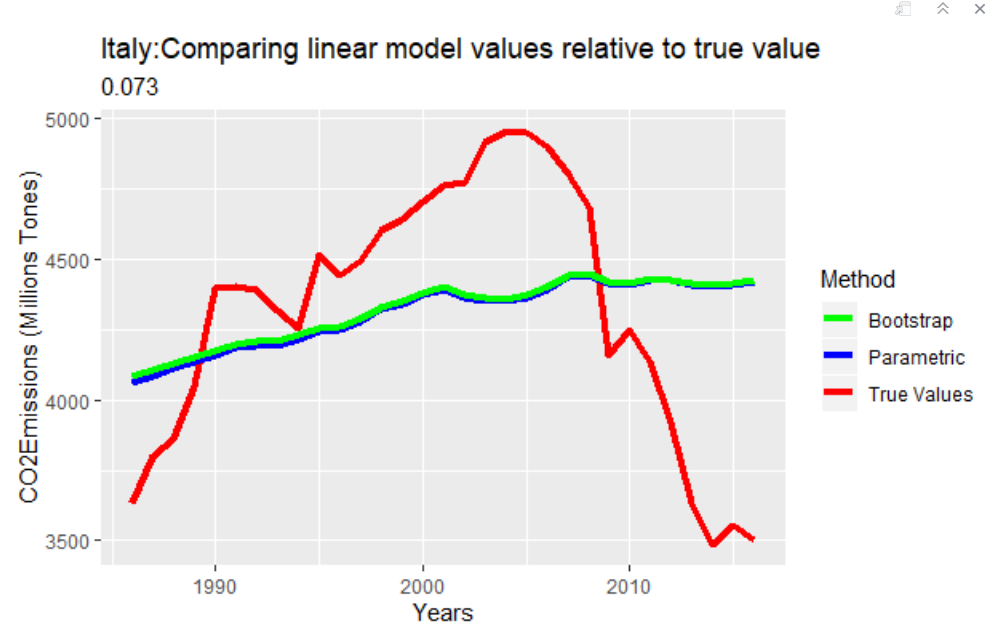
**CO2i= 1881.036484+3.676151\*GDP per Capitai + ei where i = years 1985,…2015**

The comparison of both equations can be seen below relative to true values:

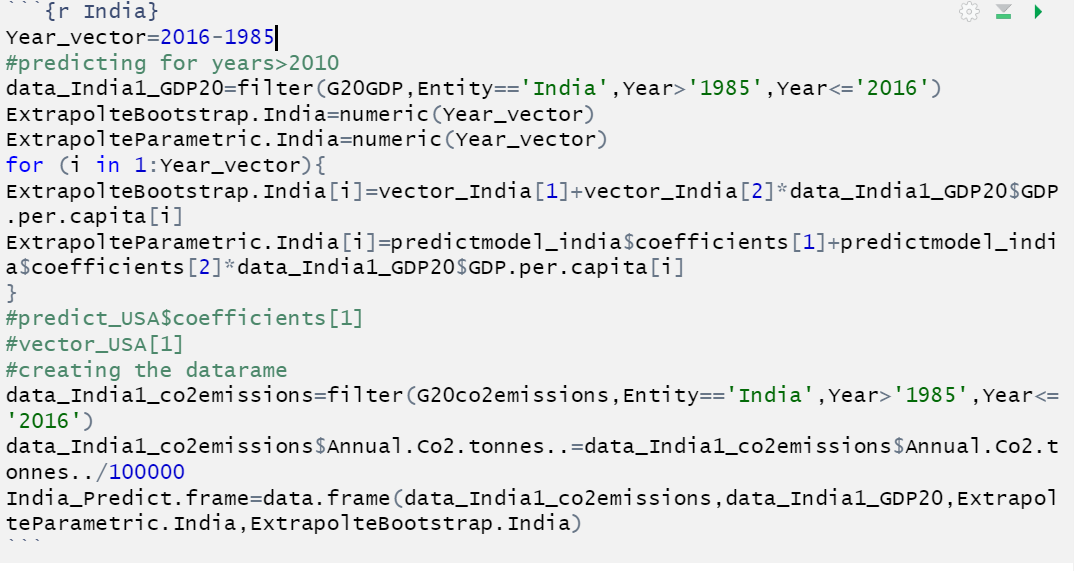


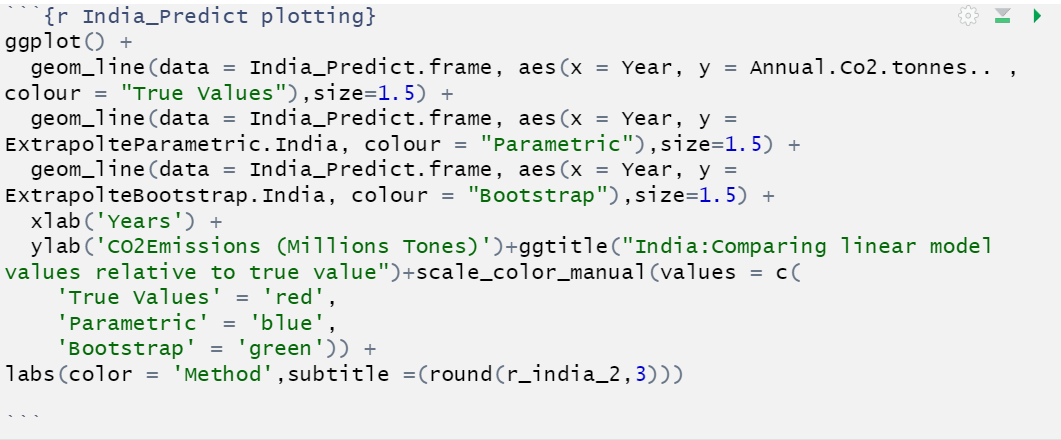






The picture observed above is given by the following code:

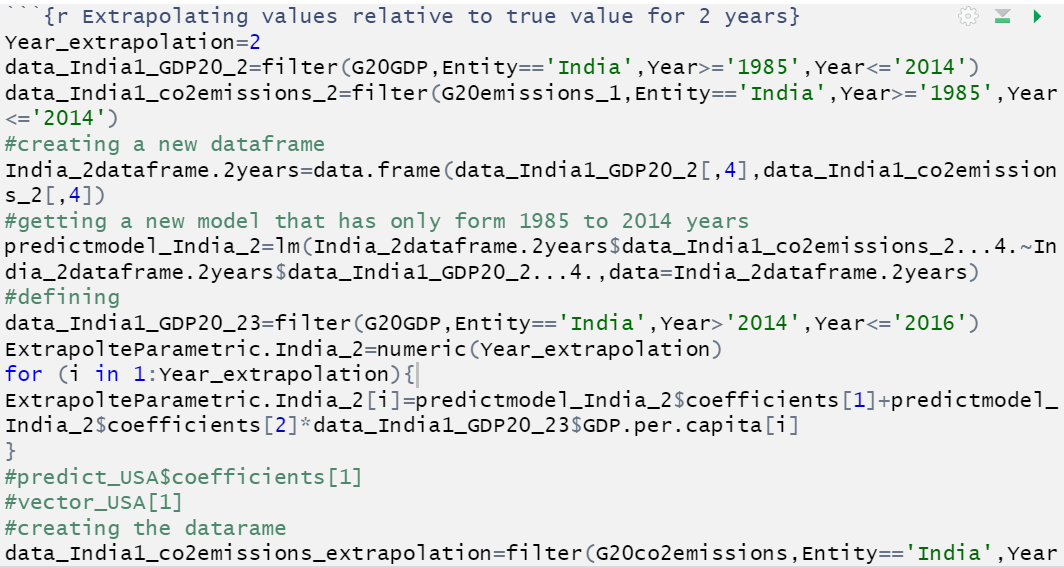


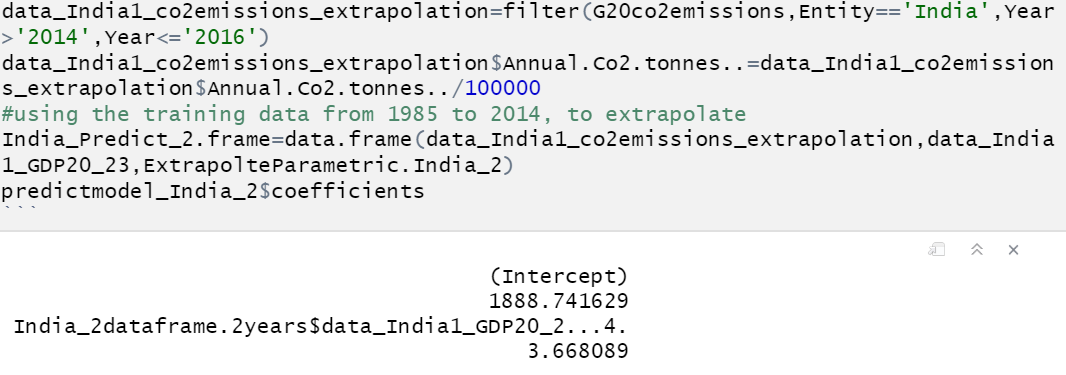


Now, in order to extrapolate the values. We use the ‘” training” data points that go from 1985 to 2014 to obtain a linear model. The linear model is then given by:

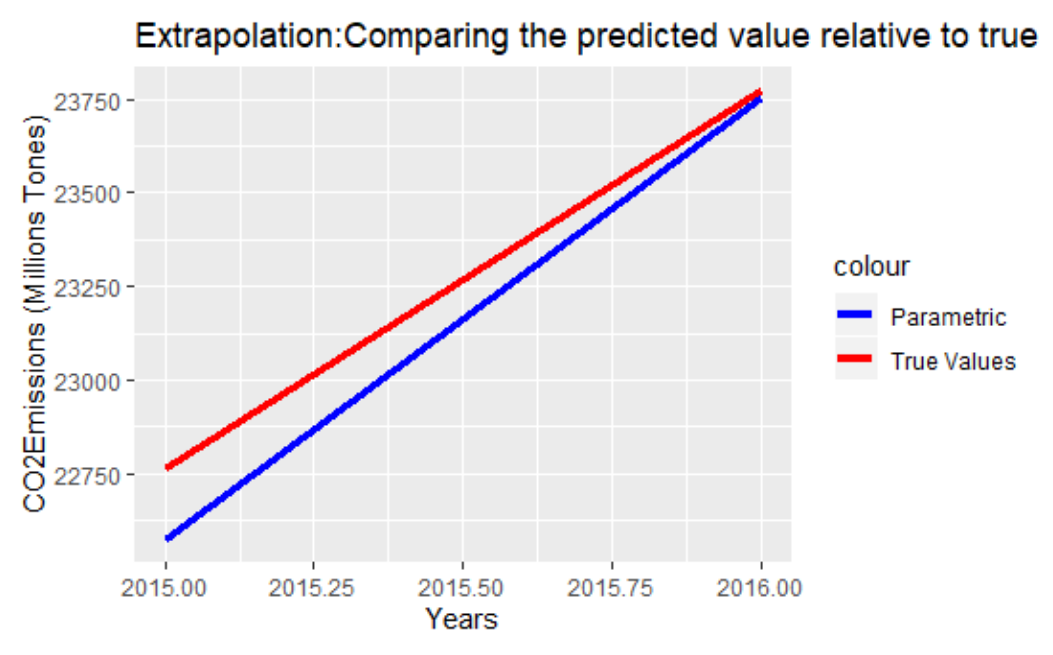
**CO2i= 1888.741629+3.668089\*GDP per Capitai +ei where i = years 1985,…2015**

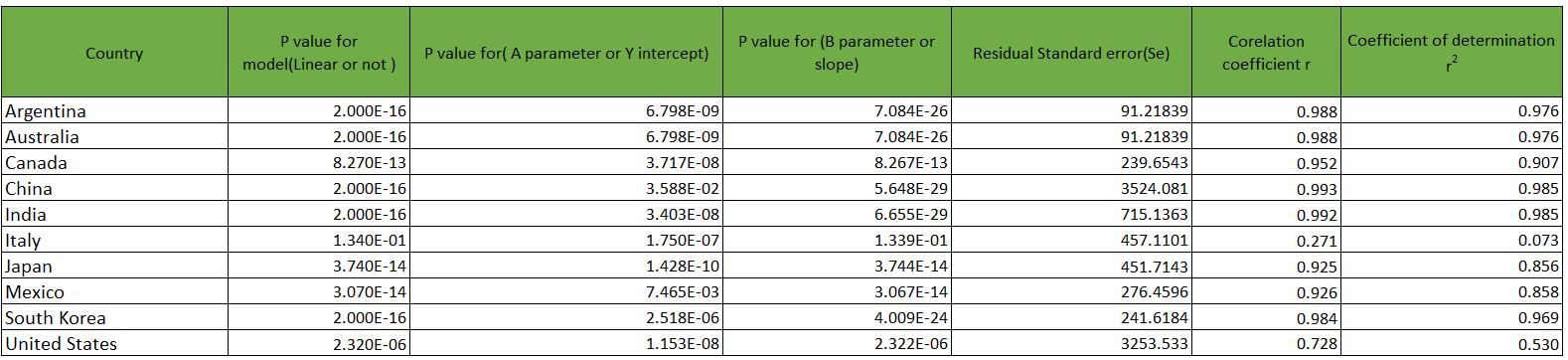
This equation can be obtained from the code given below:

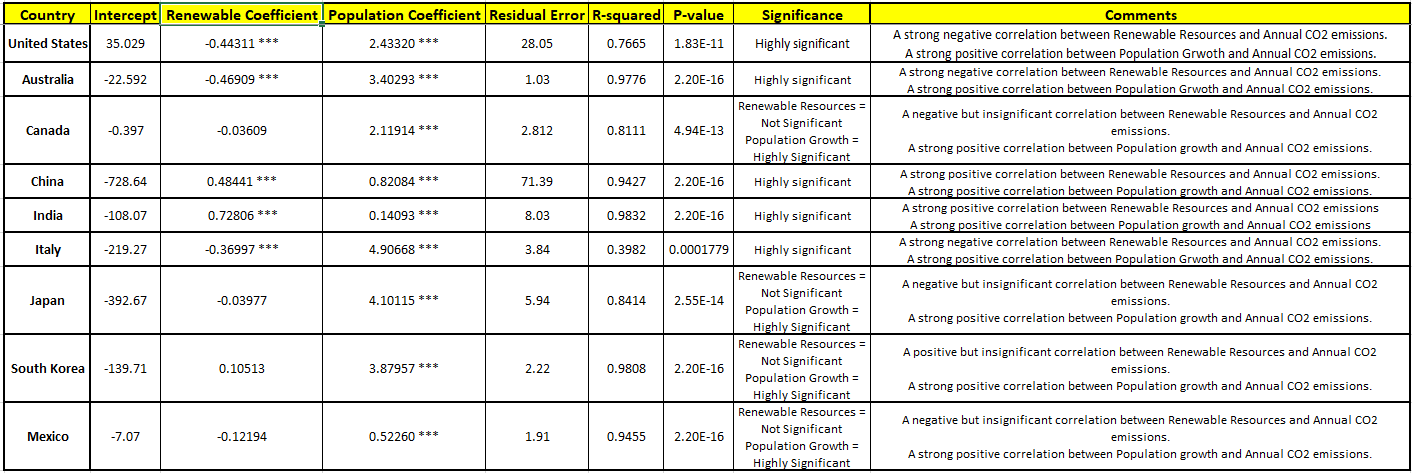




Then, we compare the true dataset values for 2015 and 2916 using the training dataset model. A picture can be observed below:







**Statistical Modeling from Multivariate data**

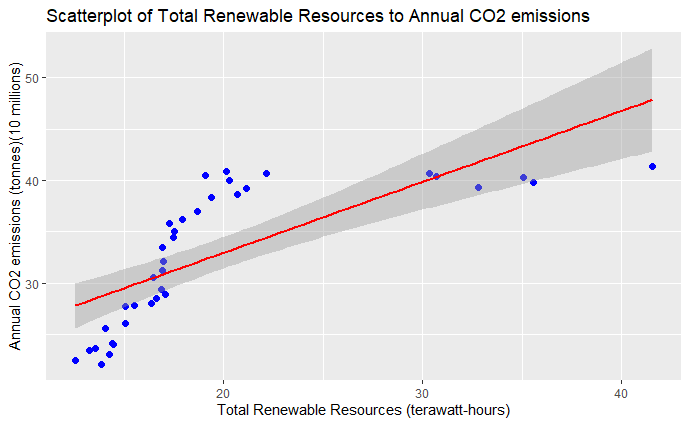
Numerous statistical modeling and investigations encompass multivariate data. Multivariate data is the observed values on three or more distinct population variables pertaining to each individual unit in the population of interest. From a notation point of view, the three variables of interest are represented by;

*X1i - the observed value of Variable X1 from population element i, i=1,2,..,n.*

*X2i - the observed value of Variable X2 from population element i, i=1,2,..,n.*

*Yi - the observed value of Variable Y from population element i, i=1,2,..,n.*

Our motivation for this study is to establish if there is a linear relationship between the X-variable **(Renewable resources)** and the Y-variable **(Annual CO2 emissions country wise)** and if the relationship is strong. We are assuming based on logic that as the renewable resources’ usage increases, the CO2 emissions should decrease. Since we want to reduce our omitted variable bias and the autocorrelation between our x-variable and y-variable, we introduce another x-variable which holds the population count for our countries of interest. Population is not correlated with the renewable resources; however, it explains a lot for Annual CO2 emissions— as the population increases, CO2 emissions increase. In order to visually observe the relationship between the Total Renewable Resources and Annual CO2 emissions, we create a scatterplot and observe that it in fact shows a positive relationship, i.e. as the total renewable resources’ usage increases, the annual CO2 emissions increase. As mentioned earlier, focusing on Australia, this is interesting as when we regress the Annual CO2 emissions on the total renewable energy consumption as bivariate data, we get a significant positive correlation of 0.6917. However, after adding the population variable in our multivariable regression, we observe a negative significant correlation of -0.46909 for Renewable Energy and a positive significant correlation of 3.40293 for Population. To check if there is a linear relationship between the Total Renewable resources and the Annual Co2 emissions, we create a scatterplot and observe that there is a positive linear relationship.



**Modeling the Relationship between Renewable Resources and Annual Co2 Emissions**

Once we determine the linear relationship between Renewable Resources and Annual Co2 emissions, we construct a statistical model that predicts the future response of the variables—depending on the increase/decrease—based on what has happened in the past. Thus, we will attempt to do a simple linear regression to build a statistical prediction model. The variable that is being predicted is called the **response variable**, denoted by Y— in our case (Annual Co2 Emissions.) The variable being used as the basis for the prediction is called the **explanatory variable**, denoted by *X1* and *X2*— in our case (Total Renewable Resources and Population count respectively.) Our study attempts to use The Probabilistic Model, which attempts to express the response variable as a linear function of the of the explanatory variables;

*Yi= A + (B \* X1i) + (C \* X2i) + ei for i=1,2,..,n*

Basing our model on the Probabilistic Model;

*Annual Co2 Emissionsi = A + (B\*Total Renewable Eneryi) + (C\*Populationi) + ei*

Where **A** is the intercept term, **B** is the correlation coefficient for Total Renewable Energy, and **C** is the correlation coefficient for Population count. In our case, estimating the model for Australia gives us the following:

|  |  |  |
| --- | --- | --- |
| **(Intercept)** | **Total Renewable** | **Population** |
| -22.5920499 | -0.4690864 | 3.4029264 |

The values of A, B, and C are **A** = -22.5920499, **B** = -0.4690864 , and **C** = 3.4029264. From this, we get an estimate of our model;

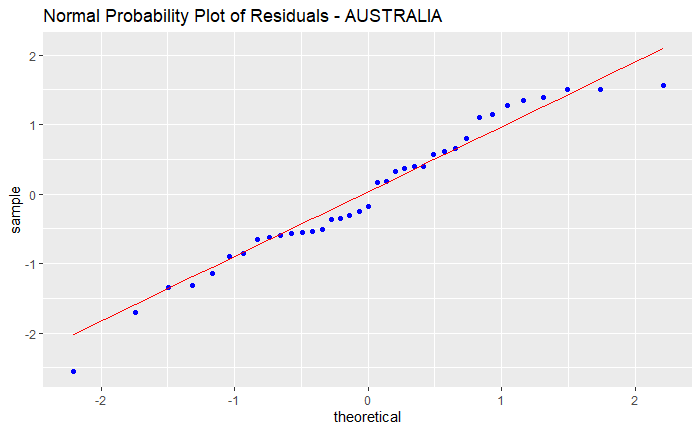
*Annual Co2 Emissionsi = -22.5920499 + (-0.4690864 \* Total Renewable Eneryi) + (3.4029264 \* Populationi) + ei*

**Conditions of the Model**

Our model is built upon two conditions;

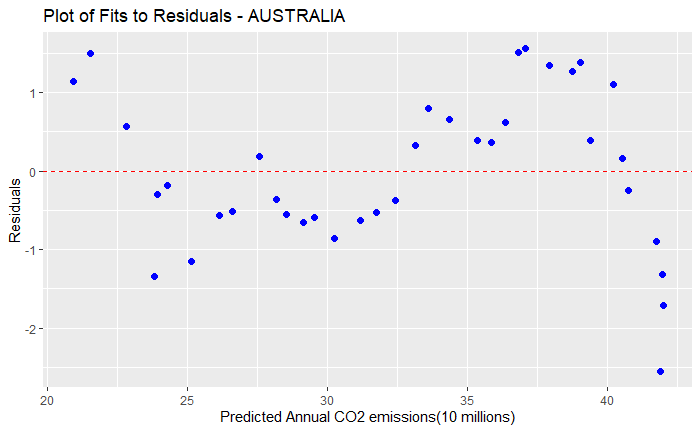
1. The Annual Co2 emissions (Response variable) are Normally distributed with the mean and standard deviation—**(Normality of the Residuals)**
2. For each distinct value of Renewable Resources and Population count (explanatory variables), the Annual Co2 emissions (Response variable) have the same standard deviation—**(Homoscedasticity)**

To check the first condition, i.e. Condition of the Normality of the Residuals, we see if the residuals are normally distributed by creating a Normal Probability Plot and plotting the residuals.



Looking at our plot, we can confirm that the first condition is met and that the residuals are normally distributed.

The second condition, i.e. Homoscedasticity, can be checked by plotting the predicted fitted values with the residuals.



Looking at our plot, we can confirm that the second condition is almost met.

**Regression Summary**

In order to look at the variation in the response variable after its linear dependency on the explanatory variable has been considered, we calculate the standard deviation of the regression. Standard deviation is more commonly known as the Residual Standard error, which describes the difference in the standard deviations of observed values and predicted values. Similarly, to determine how well our statistical model mimics the true values, we calculate the coefficient of determination, also known as r-squared. The r-squared value describes the accurateness of our model. Both the statistics can be computed using the regression summary. The multivariate regression for Australia is summarized below;

Residuals:

Min 1Q Median 3Q Max

-2.5516 -0.5950 -0.1793 0.6615 1.5609

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | -22.59205 | 1.69715 | -13.312 | 4.83e-15 \*\*\* |
| Total Renewable | -0.46909 | 0.05117 | -9.168 | 1.03e-10 \*\*\* |
| Population | 3.40293 | 0.13225 | 25.732 | < 2e-16 \*\*\* |

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

Residual standard error: 1.027 on 34 degrees of freedom

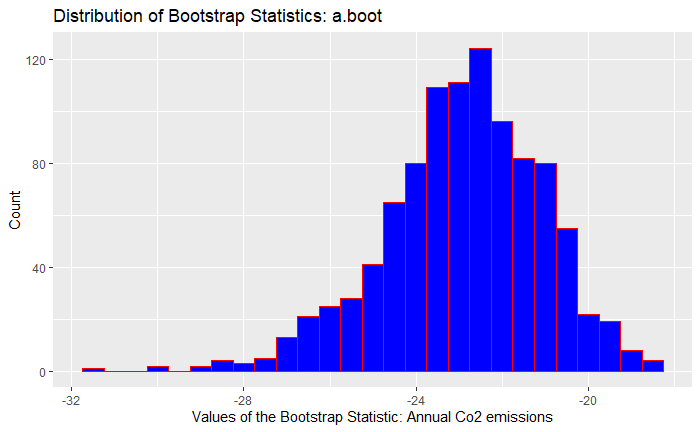
Multiple R-squared: 0.9776, Adjusted R-squared: 0.9763

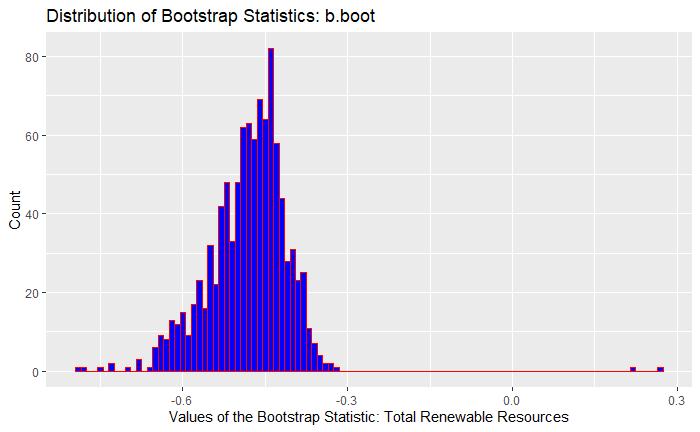
F-statistic: 741.4 on 2 and 34 DF, p-value: < 2.2e-16

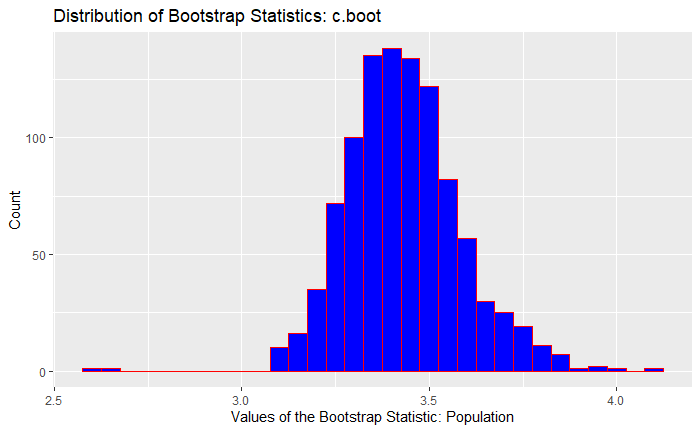
Our results show that a **1** point increase in the usage of Renewable resources decreases the Annual Co2 emissions for Australia by **0.47** points**.** Similarly, a **1** point increase in the population of Australia will result in the Annual Co2 emissions increasing by **3.4** points. A low Residual standard error of **1.027** and a high r-squared value of **0.9776** confirms the accurateness of our model and suggests that the predicted values observed from our model are close to the true values.

In order to reaffirm our model, we perform a bootstrap method to compare the parametric results and the bootstrap results. Below is the inspection of the bootstrap method.

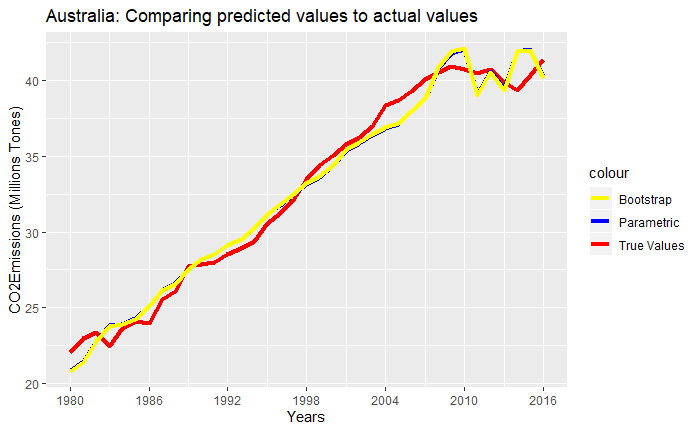








The bootstrapping method proves to be very useful in providing a normal distribution. When resampling 1000 times, following the law of large numbers, which states that by performing the same experiment many times, the average of the results obtained from the trials should be closer to the mean expected value. Looking at the bootstrap inspection above, we can observe that the results obtained from the bootstrap method are quite close to the parametric method, which in turn reaffirms the accurateness of our model. Finally, we compare the bootstrap method to the parametric model and graph them against the true values to get a better sense of the fit of our model.



Looking at the comparison graph, we can see that our bootstrapping results are almost the same as our parametric results. Furthermore, both our bootstrap results and parametric results are very close to the actual true values.

**Regressions for all the Countries**





**United States**

Our results show a strong negative correlation between Renewable Energy usage and Co2 emissions. A 1 point increase in Renewable Energy usage decreases the Annual Co2 emissions by 0.44. On the other hand, we see a strong positive correlation between Population growth and Annual Co2 emissions. A 1 point increase in population would increase the Annual Co2 emissions by 2.433 points.

**Australia**

We find evidence that higher Renewable Energy use yields negative effects on the Annual Co2 emissions. The results suggest that a 1 point increase in Renewable Energy usage would decrease the Annual Co2 emissions by 0.469 points. Population growth has a statistically significant positive impact on Annual Co2 emissions. 1 point increase in the population would lead to a 3.4 point increase in the Annual Co2 emissions.

**Canada**

Our results suggest a negative but insignificant relationship between Renewable energy use and Annual Co2 emissions. It shows that a 1 point increase in Renewable energy usage would decrease the Annual Co2 emissions by 0.036 points. While the results for Population show a positive significant relationship with Annual Co2 emissions. If the population grows by 1 point, the Annual Co2 emissions will increase by 2.119 points

**China**

Our findings suggest a strong odd significant relationship between Renewable energy usage and Annual Co2 emissions, with a positive direction. A 1 point increase in the Renewable energy usage leads to an increase in the Annual Co2 emissions by 0.484 points. We can interpret that even though China has increased its usage of Renewable energy, its dependence on Fossil fuels is still high, thus still using more fossil fuels to meet its energy demand. Our results suggest a strong positive correlation between Population and Annual Co2 usage. If the population increases by 1 point, the Annual Co2 emissions increase by 0.82 point.

**India**

Our findings suggest an odd significant relationship between Renewable energy usage and Annual Co2 emissions, with a positive direction. A 1 point increase in the Renewable energy usage leads to an increase in the Annual Co2 emissions by 0.728 points. Similar to China, we can interpret that even though India has increased its usage of Renewable energy, its dependence on Fossil fuels is still high, given its integration in the global market, thus still using more fossil fuels to meet its energy demand. The study finds a positive significant correlation between population and Co2 emissions. 1 point increase in the population will lead to 0.14 point increase in the Annual Co2 emissions.

**Italy**

We find a significant negative correlation between Renewable energy usage and Annual Co2 emissions for Italy. 1 point increase in the Renewable energy usage decreases the Annual Co2 emissions by 0.37 point. Our results suggest a strong significant relationship between population and Annual Co2 emissions, with a positive direction. 1 point increase in the population would increase the Co2 emissions by 4.9 points.

**Japan**

Our study suggests a negative but insignificant relationship between Renewable energy usage and Annual Co2 emissions. 1 point increase in the Renewable energy usage decreases the Annual Co2 emissions by 0.039. We find a significant strong positive correlation between Population and Annual Co2 emissions. 1 point increase in the population leads to 4.1 point increase in the Annual Co2 emissions.

**South Korea**

We find a positive but insignificant relationship between Renewable energy and annual Co2 emissions. Our study shows that 1 point increase in the Renewable energy usage leads to the Annual Co2 emissions to increase by 0.1 point. We find a strong significant relationship between Population and Annual Co2 emissions, with a positive direction. 1 point increase in the population increases the Annual Co2 emissions by 3.87 points.

**Mexico**

Our study suggests negative but insignificant relationship between Renewable energy usage and Annual Co2 emissions. We observe that 1 point increase in the Renewable energy usage decreases the Annual Co2 emissions by 0.12 point. Our results show a positive significant correlation between population and Annual Co2 emissions. 1 point increase in the population increases the Annual Co2 emissions by 0.52 point.

*Conclusion:*

Bivariate data (X variable and Y variable) can be correlated. Furthermore, one can obtain an equation (linear vs nonlinear) to predict/extrapolate the CO2 emissions as a function of GDP per capita, and also as a function of Renewable energy consumption and population. For the scope of this project, we only look at linear regression methods. The linear regression method follows the equation below:  
**CO2i = 1872.581507 + 3.676635\*GDP per Capitai + ei where i = years 1985,…2016**

While, the multilinear regression method uses the equation below:  
**CO2i = -22.5920499 + (-0.4690864 \* Total Renewable Energyi) + (3.4029264 \* Populationi) + ei where i = years 1980,…2016**

Both these models are valid and meet all the conditions normal probability distribution and homoscedasticity. Both the parametric method, and the bootstrap method resulted in similar linear model. The difference for both methods relative to true values is negligible. Lastly, predicted values and true values are close.

Overall, as countries grow economically (GDP), their energy demand expands and as a result consumption of fossil fuels increases. Consequently, more green house gases including CO2 are released. These green house gases are harmful to the humans, animals and the environment. Using more renewable energy sources decreases the dependency on fossil fuels to meet the ever-increasing energy demand and decreases the green house gas emissions. Hence, all countries should try to increase their renewable energy source percentage in their energy portfolio.

**References:**

[1] “Ambient air pollution: Health impacts”, World Health Organization.

https://www.who.int/airpollution/ambient/health-impacts/en/

[2] “Mortality from household air pollution”, World Health Organization, 2016.

https://www.who.int/gho/phe/indoor\_air\_pollution/burden/en/

[3] “Ambient Air Quality and Health”, World Health Organization, 2018.

https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] “Information about G20 countries”.

https://en.wikipedia.org/wiki/G20

[5]

Fossil Fuels consumption:

<https://ourworldindata.org/fossil-fuels>

CO2 bySource:

<https://ourworldindata.org/renewable-energy>

Annual\_CO2\_Emissions by region:

<https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>

Modern Renewable Energy Consumption:

<https://ourworldindata.org/energy-production-and-changing-energy-sources>