Statistics Project

Puru Vaish

October 30, 2020

1 Introduction and Overview

In this Project we analyse the Data Set for Consumption CO2 Emissions [1]. Since this data set is only the value of Consumption CO2 Emissions, this data has been merged with a few others to increase the number of dimensions in the data. The data sets that have been merged include: The Population of each country each Year[3], The classification of each Country in terms of GDP by World Bank[4] and The total GDP per country per year by World Bank[2]. I chose this data set for my project since I was very interested in finding the relation between Consumption CO2 Emissions and the GDP of a country. Since the data from the sources need to reshaped there are some effects summarised in Appendix B

2 Quality of Data

2.1 In terms of missing data and Range

Most of the data is of great quality and with very few missing values `NA` or `".."` for each particular data set. In the Consumption CO2 emissions data set only 117 countries existed, this is because not all countries register an accurate value for this and also the values only existed for 1990 to 2017. For this reason we truncate all data sets from 1990 to 2017 to make the analysis consistent.

2.2 In terms of Country Names

However there were some issues with non-standardised names of the country, as most did not give the ISO names of the country to use as key, and due to small different names, sound countries were lost during analysis, in the end, only 109 countries were part of the analysis, losing 8 country data in the process of merging data.

3 Research Questions

1. Is the Percent Change in the Consumption CO2 Emissions in Low Income Level Countries greater than High Income Level Countries with a significance level of 5%?
   This question explores the idea if Low Income Level countries over the period of 1990 to 2017, have had a greater increase in Percent Change than High Income Level Countries.

2. Is the Consumption CO2 Emissions per Capita of a High Income level Country greater than that of a Low Income level Country at a significance of 5%?This question is contrast from the one above, since percent increase is not the same as emitting more, and we want to know more about if a countries income level has any relation to Consumption CO2 Emissions per capita.

3. Is the Consumption CO2 Emissions in 2017 greater than that of 2016 at a significance of 5%?
   This question is contrast from the one above, since percent increase is not the same as emitting more, and we want to know more about if a countries have had greater Consumption CO2 emissions in the last years. Here we compare the last two years of the data set 2017 and 2016.

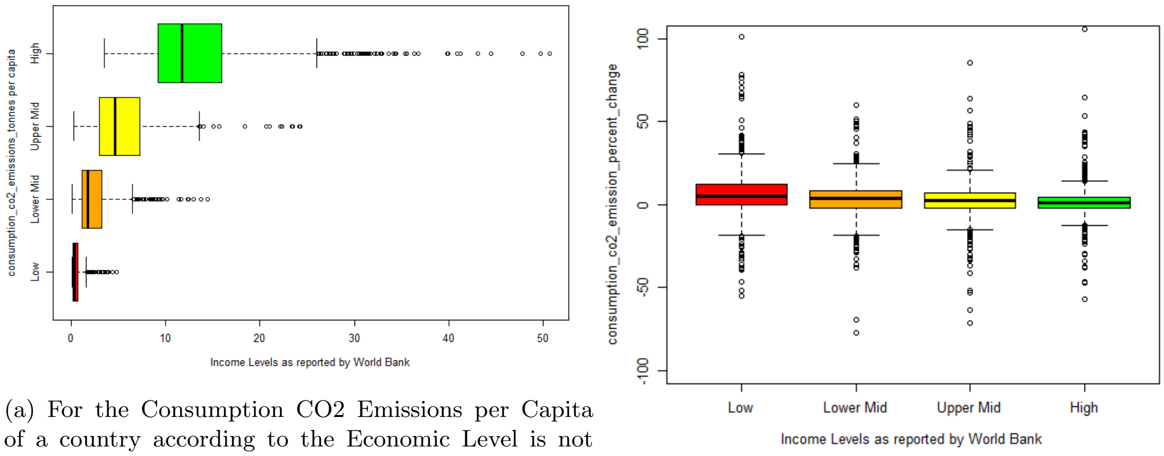
4. Is the Total increase in Consumption CO2 Emissions in the period of 1990 to 2017 for every country each year, greater than 0 at a significance level of 1%
   This Question explores the fact if Countries in the period of 1990 to 2017 have in general shown a trend of increasing their Consumption CO2 Emissions every year. A value of 1% is chosen to give

the sense of certainty to those who deny the fact that countries in general have not been increasing their Consummation CO2 emissions.

5. Given a population percentage change of a Country is it possible to create a linear model for Consumption CO2 Emissions percent change? Is the same possible for GDP percent change? And is the same possible for a combination of the two.

4 Descriptive Statistics

4.1 Box Plots

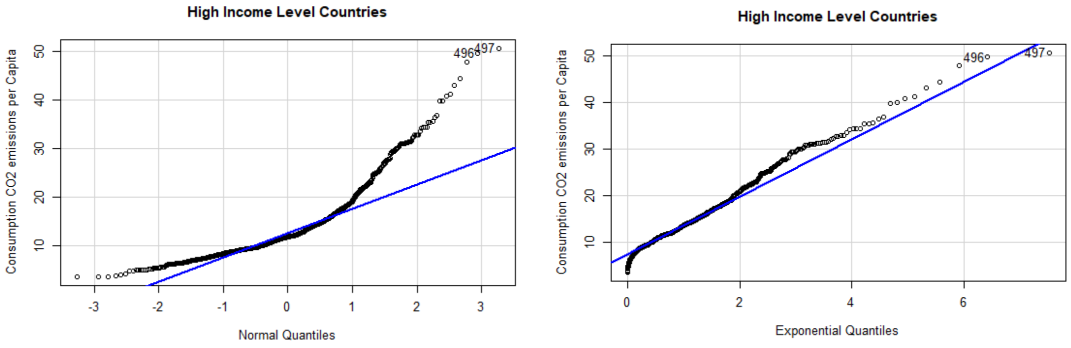


(a) For the Consumption CO2 Emissions per Capita of a country according to the Economic Level is not distributed Normally Distributed, but instead seems to be either Exponentially Distributed or Gamma Distributed. This tells us it is likely not a good idea to test them using Parametric tests. This is also somewhat expected since this attribute is almost never negative.

(b) For the Percent Change it seems plausible that the Percent change for the 4 different Economic Levels of a country is Normally Distributed, given a close to symmetric distribution.

Figure 1: Historical Data

4.2 QQ Plots



(a) Normal Quantiles

(b) Exponential Quantiles

Figure 2: We continue our debate on the distribution of Consumption CO2 Emissions per Capaita based on Income Level using QQ Plots. Only High Income Level Countries is shown here. High Income Country definitely shown the more likely distribution, being the exponential distribution as seen from the QQ with exponential quantiles compared to normal quantiles.

5 Computational Results

5.1 Research Question 1

Method From the descriptive statistics, in the box plots Figure 1b, we assume the Percent Change of the Consumption CO2 Emissions is normally distributed, we have reason to believe so from the Box Plot frawn above. We then continue with the F-Test with 1% significance level to check if the Variance are equal, if they are, we can use the Two Sample T-test with equal variance else we perform the Welch's t-test, which is the default, both with 5% significance level.

Hypothesis Test

1. The F-test gives a 99% confidence level of (1.499276, 2.15707) where the ratio of variances was 1.803869. Since the ratio is in the confidence interval, we can assume the variances to be equal. We then proceed with Two Sample t-test with equal variances assumed
2. Let X$\sim N(\mu_x, \sigma_x^2)$ and Y$\sim N(\mu_y, \sigma_y^2)$, where X represents the distribution for Low Level Income and Y for the High Level Income Countries.
3. $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x > \mu_y$
4. If H_0 then $X - Y \sim t_{n+m-2})$ where n = length(x) and m = length(y) and x and y are vectors of the observations of X and Y
5. Observation of Test Statistic = 6.4723
6. p-value: $6.428e - 11$, hence reject Null Hypothesis in favour of alternative hypothesis as p-value < 0.05

Conclusion The Low Income Level Countries have greater Percent Change in Consumption CO2 Emissions than High Level Income Countries at a significance of 5% for years 1990 through 2017.

5.2 Research Question 2

Method From the Figure 2a and Figure 2b it is clear that the distribution of Consumption CO2 Emissions per Capita for a country is not distributed normally and likely exponentially. For this reason we can not test our Research Question using a parametric test. We can not apply a Sign Test here since the data is not paired, and we also can not apply MP test since the hypothesis test is not simple, hence with my current knowledge this research question is left unanswered.

5.3 Research Question 3

Method First we check the distribution of data again. Plotting a histogram tells us that data is again non normal, and also since most countries only have positive Consumption CO2 Emissions assuming normal distribution is highly likely to be erroneous. Here we perform the Sign Test on the median of the differences. In R this implemented as a `binom.test()`.

Hypothesis Test

1. let X be the distribution of Consumption CO2 Emissions in 2016 and Y be the distirbution for 2017.
2. $S_i = sign(Y_i - X_i)$ where $i = (1, .., 109)$ where 109 is the number of countries in the data set and $S_i \sim Bernoulli(p)$
3. Test Statistic = t = $\sum_{i=1}^{109} S_i$
4. $H_0 : p = 0.5$ against $H_1 : p > 0.5$
5. If $H_0 : t \sim B(109, 0.5)$
6. Observation of Test Statistic = 83
7. p-value: $3.338e - 10$, hence reject Null Hypothesis in favour of alternative hypothesis as p-value$<$ 0.05

Conclusion We reject the Null Hypothesis that there is no difference in Consumption CO2 emissions between years 2016 and 2017 at a significance level of 0.05 for the alternative that in 2017 the emissions were greater.

5.4 Research Question 4

Method First we check if data is normal. This data is a combination of data we assume to be normally distributed, each of the High, Upper Middle, Lower Middle and Low level ncome countries show a symmetric box plot in Firgure 1b, hence we assume it is normally distributed as well. We will then apply one sample t-test.

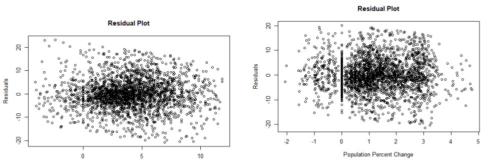
Hypothesis Test

1. Let X$\sim$ $N(\mu_x, \sigma_x^2)$ where X represents the distribution for Consumption CO2 Emissions percent change.
2. $H_0 : \mu_x = 0$ against $H_1 : \mu_x > 0$
3. If H_0 then $T \sim t_{n-1}$
4. Observation of Test Statistic = 6.4723
5. p-value: $6.428e - 11$, hence reject Null Hypothesis in favour of alternative hypothesis as p-value < 0.05

Conclusion We reject the null Hypothesis that Percent Change in the Consumption CO2 Emission is 0 for the Alternative Hypothesis that Percent Change is greater than 0 in general over the years 1990 to 2017 for every country at a significance level of 5%.

5.5 Research Question 5

Method For this analysis we only look at the historical percent changes of Consumption CO2 Emissions Percent Change, and we do with the assumption that the percentage changes in CO2 emissions in one year for a country is independent to the percent changes in the other years, similarly for percent changes in population and Total GDP percent change. Notice we are not saying here that we assume all three percent changes are independent to each other. A boxplot is used to find the range of the data which is inside the standard $1.5 * (IQR) - Q1$ and $1.5 * (IQR) + Q3$ for each of the explanatory and descriptive variables. The reason for this is that a lot of countries historically can have a lot of reasons for unnatural changes, like war and famine, and economic changes like the Stock Market Crash that might have happened in the previous year.Since we are using percent changes, these changes normalise in the next year, so we are not removing too many data points. Another example of the kind of scenario we are filtering out here is the case of hyper inflation in Zimbabwe which do not give any correct information about a normal functioning economy to get a linear model from, case such as those deserve their own model.



(a) GDP-Residuals

(b) Population Residuals

Figure 3: Residual Plots

Conclusion For GDP: the p-value for the coefficient is 0 and the Estimate of the coefficient is non 0. But the R squared is only 0.08975. Doing Shapiro Test gives the residuals is not normal.

For Population: the p-value for the coefficient is 0 as well and the estimate of the coefficient is non 0. But the R square is only 0.04774. Doing Shapiro Test gives the residuals is not normal.

For the multiple regression model: the p-value for both the gdp_percent_change and population_percent_change is also 0, but the Adjusted R-Square score is only 0.1179. Looking at the residuals plot it seems Random Chaos, but performing the Shapiro-Wilks test reveals that the residuals are not normally distributed as p-value calculate is $1.106e - 13$.

So we do conclude there is a correlation but our model does not explain most of the variation in data. The residuals are not normally distributed, but residual plot seems very chaotic, so there is a possibility of a better model with the explanatory variable being related non-linearly. Also compared to single variable linear model, the multi-variable linear model is better explain the variation.

References

[1] *Consumption CO2 Emissions (1000 metric tons)*. 2017. URL: `https://github.com/open-numbers/ddf--gapminder--co2_emission`.

[2] *Total GDP Adjusted for Inflation*. 2019. URL: `https://data.worldbank.org/indicator/NY.GDP.MKTP.KD`.

[3] *Total Population*. 2019. URL: `http://gapm.io/dpop`.

[4] *World Bank Countrty Classification by Income*. 2019. URL: `http://databank.worldbank.org/data/download/site-content/OGHIST.xls`.

Appendix A Data Reshaping

A major issue with the data that was extracted, was the shape of the data, so much so it needs to be talked about. The data for Country Classification was in a non standard xlsx file, so that to extracted separately. All data was in the shape of a time-series while we want the data to be in a data frame format to do some appropriate tests. For this reason I used a library called `tidyverse` and use a function called `melt` to change the shape. The code for which can be found in Listing **??**. Please read about the known effect that reshaping the data in such a way has for the analysis of the data at Appendix B.

Appendix B Effect of Reshaping the Data

When we reshaped the data in the way currently done, for the parts of the data analysis where we consider the percent change in data we treat the country in one year to be independent of the same country in other years. This mostly works since the percent change of a quantity is independent of the change in other years. In cases where this is not true we assume they are independent, for instance a country with set economic goals may consistently only have growth of 5% every year for 10 years, making the data dependant, but in this analysis we assume the data to be independently. For the historical Consumption CO2 Emission Box Plots, we have a similar issue since the country is same, the value of this quantity is likely to be similar, but we treat as if they are independent as we are blocking on there Income Levels which do change over the years for a single country and this was an analysis specially on that aspect, but for all intents and purposes it is assumed to be independent even though in general they are likely not.

Appendix C Calculated Fields

From the Data of the CSV files, we generate the following data:

1. population percent change
2. gdp percent change
3. consumption co2 emissions percent change
4. consumption co2 emissions per captita
5. gdp per capita

Appendix D Data Cleaning

All NA, and Inf were dropped from the data set. For WB Classification of country, ".." was used unspecified so this was also dropped. Some Inf and NA were created calculating the percent changes. This cleaning was done when needed before every particular analysis.