Analysis of the relationship between economic factors and CO_2 emission per capita

8 November 2019

EPA1315 - Data Analytics and Visualization (2019 - 2020 Q1)

Faculty of Technology, Policy and Management

Delft University of Technology

Group 6

Pham Dang Khoa (4815033)

Nico van Eeden (4980417)

Sharath Chandar (5140633)

Table of contents

Acknowledgement

The work in this project was distributed as follows.

Section	Person in charge
1. Problem Understanding	Pham Dang Khoa Sharath Chandar Nico van Eeden
2. Data Understanding 2.1 Collection of Initial Data 2.2 Description of Data 2.3 Verification of Data Quality	Nico van Eeden
2.4 Data Exploration 2.4.1 CO2 emission per capita of the world 2.4.2 CO2 emissions per capita by region 2.4.3 Distribution of CO2 emission per capita 2.4.4 Correlation analysis	Pham Dang Khoa
2.4.5 Productivity vs CO2 emission per capita.	Nico van Eeden
2.4.6 Percentage of employment vs CO2 emission per capita	Sharath Chandar
3. Data Preparation	Nico van Eeden & Pham Dang Khoa (R code); Sharath Chandar (writing)
4. Modelling	Pham Dang Khoa

1. Problem Understanding

1.1 Sources of CO2 emissions

Carbon dioxide is emitted into the atmosphere either within a natural balanced cycle or from human activities. Natural sources of CO2 emissions include the oceans, soil, flora and fauna and the volcanoes. Human sources of CO2 emissions include manufacturing, deforestation as well as the burning of fossil fuels such as coal, oil and natural gas. Emissions from human sources of pollution are lower compared to the natural sources. However, with the increased human activities from the industrial revolution period, it has disturbed the stable balance of our planet (Smith, 1999). Therefore, it is very imperative to be aware and act upon the problems associated with increased CO2 emissions. Nearly 90% of all carbon dioxide emissions generated by human stem from the combustion of fossil fuels such as coal, natural gas and oil. Deforestation and land use make up 9%, while industrial activities such as cement manufacturing account for the remaining 4% (Che-project.eu, 2019).

1.2 Post-industrial period CO2 emissions

The increase in atmospheric CO2 concentrations during and after the industrial revolution period is represented in Figure 1. It shows that human-related emissions increase the greenhouse gas concentrations in the atmosphere in a few spans of decades and disturb the ecological balance dramatically (Climatecentral.org, 2019). There are other important greenhouse gases besides CO2, although CO2 is the most dominant.

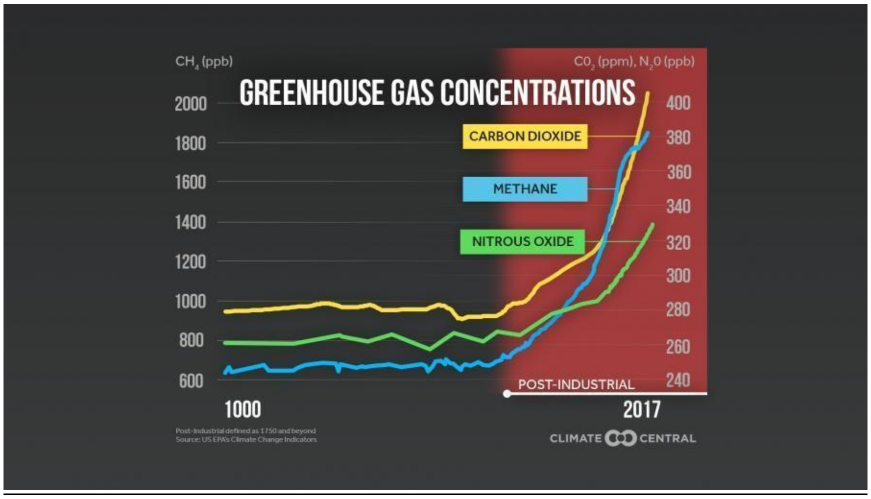


Figure 1: CO2 Emission Trend (Climatecentral.org, 2019)

1.3 2018 CO2 emissions

The global CO2 emissions are still rising significantly, which is indicated by the emissions from fossil fuel energy sources as shown in Figure 2. This increase in the CO2 emission is worrying even with the rise in the renewables share of energy and with the global call for reduction in carbon emissions (Levin, 2019).

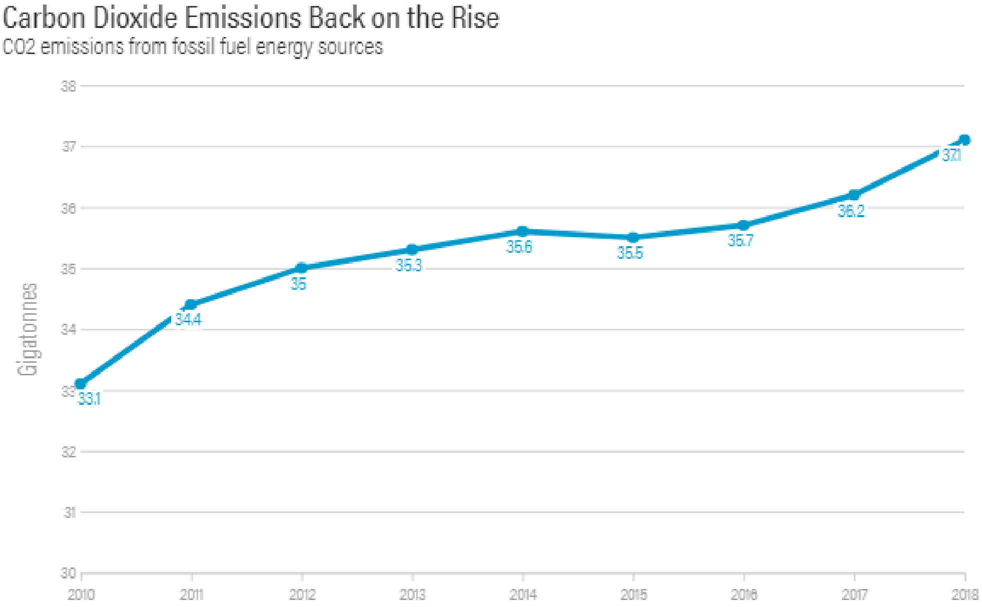


Figure 2: CO2 Emission Trend (Levin, 2019)

In figure 2, the total carbon dioxide emissions from fossil fuels was 36.2 gigatonnes of CO2 in 2017, and it rose to a new record of 37.1 gigatonnes CO2 in 2018. There are a few major factors for this increase. First, the world's major emitters are not doing enough to mitigate emissions together. Second, there is still a rise in natural gas and oil use (Levin, 2019).

196 countries came forward at the Paris climate conference which was held December 2015. They signed the Paris agreement to avoid the effects of climate change by limiting the warming of the planet to 1.5°C and reducing carbon emissions by each country. However, the rise in increased global emissions by 1.6% in 2017 and 2.7% in 2018 are clear indicators that the conditions of the agreement will be very difficult to meet (Lux, 2019). It will be very ideal if the developed countries lead the way in reducing CO2 emissions, but the trend of mitigation is not keeping up with the pace required.

1.4 CO_2 emissions by country

The global distribution of CO2 emission by country is represented in Figure 3. This chart is presented in the report to have an understanding about the top countries in the world which is responsible for CO2 emissions. In terms of absolute emissions, the heavy hitters are immediately obvious. Large economies such as China, the United States, and India alone account for almost half the world's emissions. It is even clearer that just a handful of countries are responsible for most emissions (Ghosh, 2019).

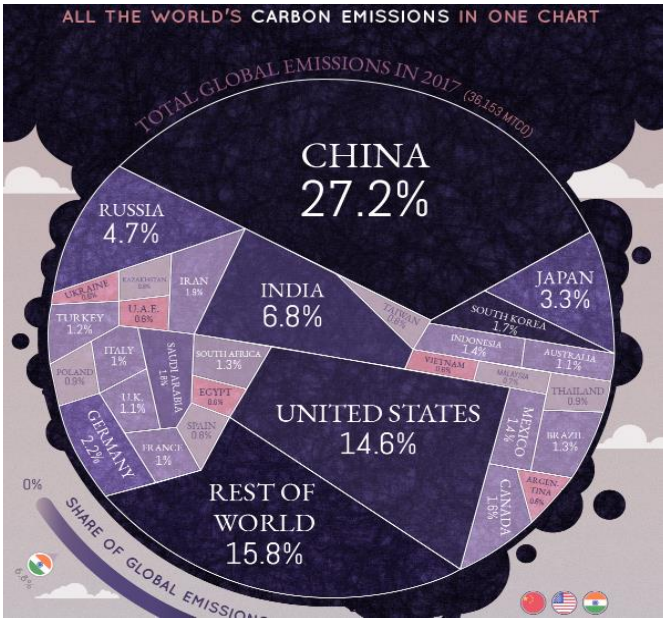


Figure 3: CO2 Emission country trend (Ghosh, 2019)

1.5 CO2 emissions per capita by country

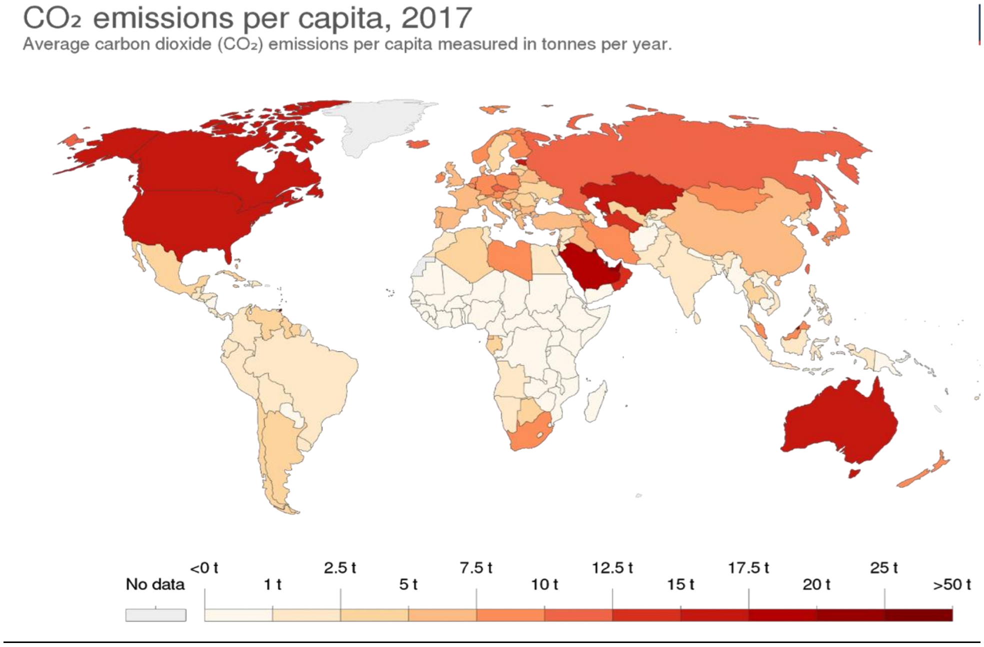

CO₂ emissions per capita, 2017
Average carbon dioxide (CO₂) emissions per capita measured in tonnes per year.
No data
<0 t
1 t
2.5 t
5 t
7.5 t
10 t
12.5 t
15 t
17.5 t
20 t
25 t
>50 t

Figure 4: CO2 Emission per capita over 2017 (Ritchie, 2019)

The contribution of the average citizen of each country is calculated by dividing its total emissions by its population. There are very large inequalities in per capita emissions across the world (see Figure 4).

The oil-producing Middle East countries are some of the largest per-capita CO2 emitters. The United States, Australia and Canada are also leading the list. Australia has an average footprint of 17 tonnes per capita, followed by the US at 16.2 tonnes, and Canada at 15.6 tonnes. This is more than 3 times higher than the global average, which in 2017 was 4.8 tonnes per person (Ritchie, 2019).

The interesting trend here is that even with the obvious strong relationship between income and per capita CO2 emissions, we cannot blindly assume that the countries with high standard of living will have a high carbon footprint. This becomes clear that Europe with similar standard of living as Australia and Canada have managed to keep their per capita CO2 Emission much lower. For example, in 2017 emissions in Portugal are 5.3 tonnes; 5.5t in France; and 5.8t per person in the UK, which is not far from the global average.

Another interesting trend is the choice of energy plays a key role. For example, in the UK, Portugal and France, a much higher share of electricity is produced from nuclear and renewable sources, so they have considerably lower carbon footprint. The poorest countries in the world still have very low per capita CO2 emissions. For example, for Sub-Saharan Africa, the average footprint is around 0.1 tonnes per year. That's more than 160 times lower than the USA, Australia and Canada (Ritchie, 2019). Thus, even though prosperity is the primary driver of CO2 emissions, policy and technological choices make a clear difference.

1.6 Consequences of increased CO2 emissions

The two main effects of an increase in CO2 emissions is the increase in CO2 concentration in the atmosphere and the increased absorption of carbon dioxide by the plants and oceans (Jain,2018). These effects are manifested in the following ways:

Global warming: The general rise in global temperature resulting from the increased CO2 emissions is a major cause of the changes in the geoclimatic conditions. Melting of polar ice caps, warming up of the oceans and increased evaporation are some of the large-scale changes. These changes have a cascading effect and lead to further global impact like the rise in sea levels hence a cause of concern to all low lying and coastal areas. Furthermore, a lot of species will be affected negatively by climate change hence a negative impact on biodiversity. Coupled to an increase in the occurrence of droughts and fresh water scarcity, an increase in the variability of the weather pattern and extreme weather phenomena makes global warming an issue of major concern (Jain, 2018).

Ocean acidification: The oceans are a large reservoir of carbon and with the atmosphere a constant carbon exchange is always in progress. At any given time before the industrial revolution, the oceans held about 38000 Gt compared to the 700 Gt in the atmosphere and about 2000 Gt in the terrestrial biosphere (Royalsociety.org, 2019). However, the recent rise in CO2 emissions has led to an increase in the absorption of CO2 by the oceans. The excess CO2 is converted to carbonic acid, reduces the pH value of the ocean and has adverse effects on the marine biology (Jain, 2018).

1.7 Objectives and methodology

The primary objective of this report is to find the relationships (impact) between the economic indicators (apart from GDP) and the CO2 emissions per capita of a country. We focused on the indicators that are related to the economy of a country. These indicators belong to the following groups:

- Economic Policy & Debt
- Social Protection & Labor: Economic activity

We use the latest data for CO2 emission per capita (year 2014). A quick correlation study will be made between the indicators selected and CO2 emission per capita. The high correlations (indicators) will be selected for further modelling. Kendall method was chosen instead of Spearman method because Spearman is sensitive to inconsistency in the data.

In the Modelling Part, our model will be checked for convergence to ensure:

- Trace plot of 2 different chains (with different automatically generated initial values) is consistent
- Density plot of the obtained coefficients show no abnormal patterns
- Gelman plot show high similarity between 2 chains
- There is low autocorrelation (approximately 0)
- Effective sizes are high

Next, the coefficients obtained by our Bayesian model will be cross-checked against the coefficients from a simple linear model. Finally, posteriors predicted by our model will be compared against actual values (from WDI data) to see whether there is any significant bias.

1.8 Research question

In the global environment, a country's economy seems to be positively correlated to its CO2 emissions per capita. Developed nations have produced much higher CO2 emissions per capita. Thus, through our report we would like to ask and answer:

How do economic factors of a country correlate with its CO2 emission per capita?

2. Data Understanding

2.1 Collection of Initial Data

We collected data from the World Development Indicators database (Datacatalog.worldbank.org, 2019). This is a free database consisting of development indicators comparable for countries all over the world. More specific, we downloaded and used the following files: WDIData.csv, WDICountry.csv and WDIVar.csv.

2.2 Description of Data

The WDIData.csv file includes 1599 development indicators accounted for the years 1961-2018 over 263 different country names. The WDICountry.csv file has 217 countries with a unique country code. We will use these country codes later on the indicator datafile. The WDIVar.csv file provides further clarification on the indicators and is used to identify the correct economic and environmental indicators of interest.

Next, we tried to explore which economic indicators (apart from GDP) influence the CO2 emission per capita of a country. Therefore, we focused on the indicators that are related to the economy of a country. These indicators belong to the following groups:

- Economic Policy & Debt
- Social Protection & Labor: Economic activity

We narrowed the list down to 14 indicators that have high potential of affecting the CO2 emission per capita.

Indicator Code	**Indicator Name**
BX.KLT.DINV.CD.WD	Foreign direct investment, net inflows (BoP, current US$)
BX.TRF.PWKR.DT.GD.ZS	Personal remittances, received (% of GDP)
DT.ODA.ODAT.CD	Net official development assistance received (current US$)
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NV.AGR.EMPL.KD	Agriculture, value added per worker (constant 2010 US$)
NV.IND.EMPL.KD	Industry, value added per worker (constant 2010 US$)

NV.IND.MANF.CD	Manufacturing, value added (current US$)
NV.MNF.TECH.ZS.UN	Medium and high-tech industry (% manufacturing value added)
NV.SRV.EMPL.KD	Services, value added per worker (constant 2010 US$)
NY.ADJ.SVNX.GN.ZS	Adjusted net savings, excluding particulate emission damage (% of GNI)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)
SL.ISV.IFRM.ZS	Informal employment (% of total non-agricultural employment)
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)

2.3 Verification of Data Quality

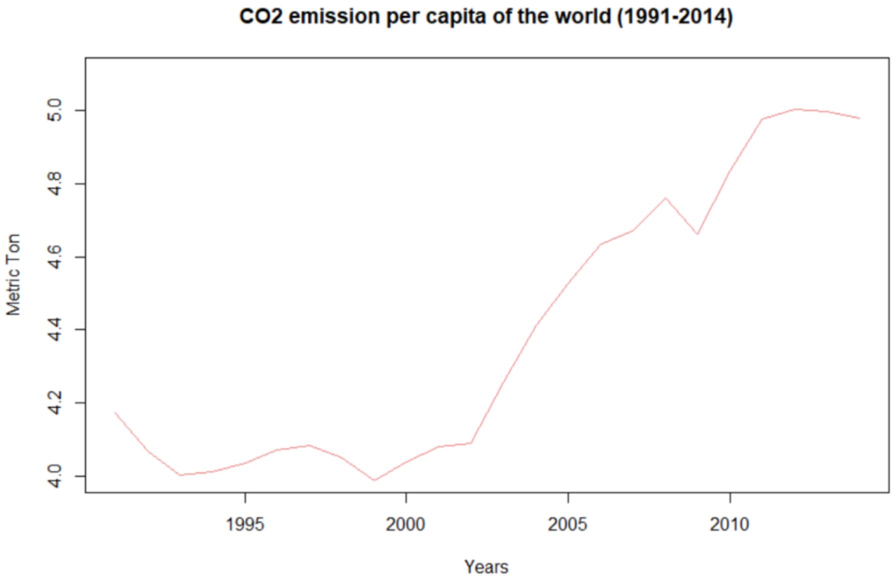
Much of the earlier data is missing (NA). However, for more recent years, starting from the 90s more data has generally become present in the columns. Also with the most recent years, some data of interest was missing. For example, for *world atmospheric CO2 emission/capita*, no numbers exist for the period before 1990 and 2015-2018.

2.4 Data Exploration

2.4.1 CO2 emission per capita of the world

The graph below shows the CO2 per capita (world average) during 1991-2014. There was a big rise from around the year 2002 to 2008, which might stem from the fast development of technology and the increase in trade among the countries. Around 2008, there was a small dip, which could be due to the global financial crisis and its effects.

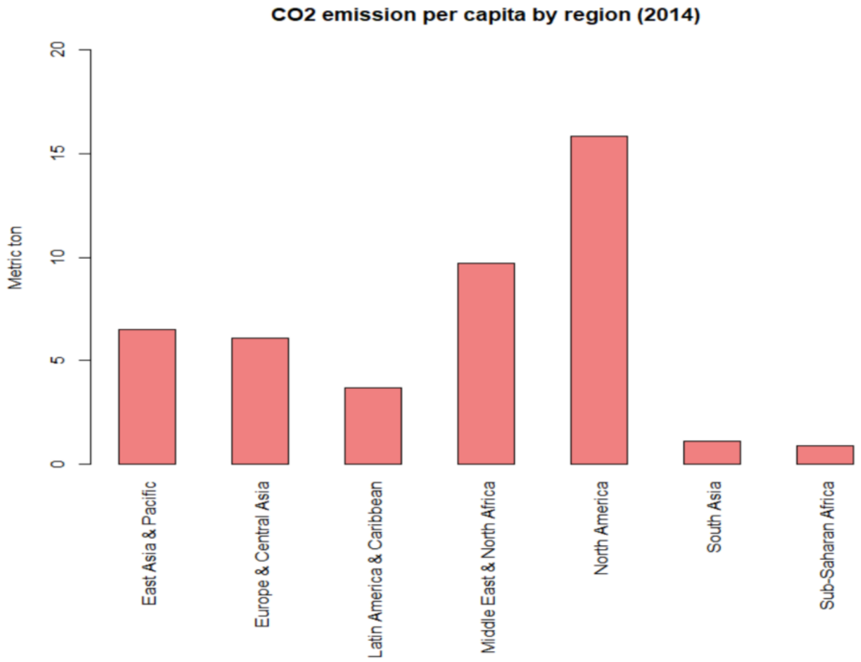
The recovery period starting from around 2009 saw a sharp increase. The world average of CO2 per capita was around 4.9 metric ton around the time of Paris agreement.



Graph 2.4.1: CO2 emissions per capita of the world for the years 1991-2014.

2.4.2 CO2 emissions per capita by region

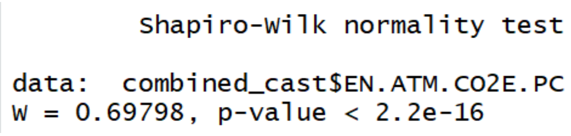
The graph shows the CO2 emissions per capita by regions. The North Americas, with its high standard of living and increased industrialization, has the highest per-capita emissions. The middle east countries also have very high emissions per capita because their economy is heavily dependent on oil production. Africa, with its low economic development, doesn't contribute much to the CO2 emissions.



Graph 2.4.2: CO2 emissions per-capita by regions

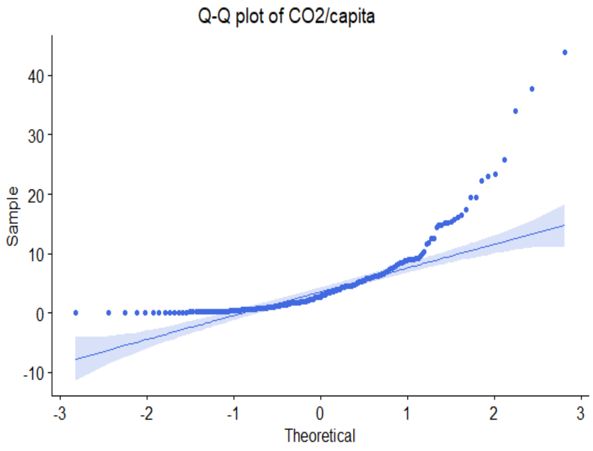
2.4.3 Distribution of CO2 emission per capita

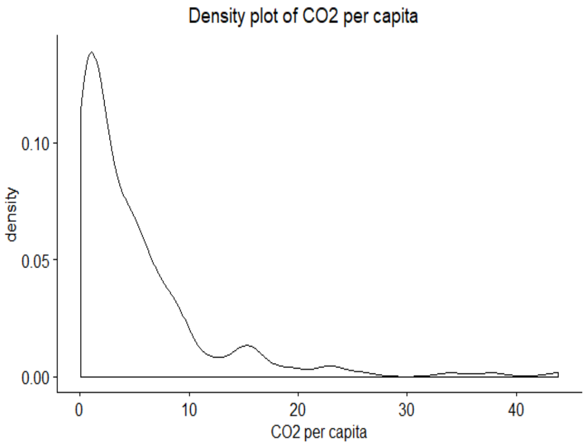
In this section, we explored whether the CO2 emission per capita follow a normal distribution. The latest data (year 2014, 204 countries) was chosen for analysis. A quick Shapiro-Wilk test provided the result:



```
	Shapiro-Wilk normality test

data:  combined_cast$EN.ATM.CO2E.PC
W = 0.69798, p-value < 2.2e-16
```

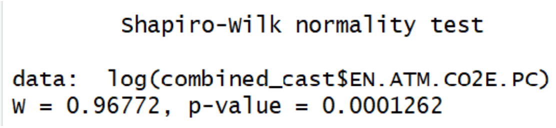
The null hypothesis of Shapiro_wilk test is that the data is normally distributed. The very small p-value (less than 0.05) rejects the null hypothesis, meaning that our data is significantly different from a normal distribution. However, it is known that Shapiro_wilk test is sensitive to extreme values. Some outliers in our CO2 emission data may affect the result. Therefore, we further checked some graphs of our CO2 emission per capita.





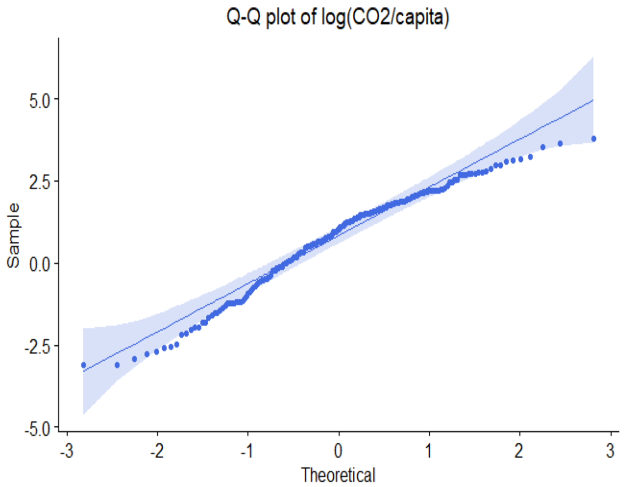
The Q-Q plot showed that our data deviated from the normal distribution (there were significant deviation at both ends of the reference line). Besides, the density plot revealed that the data is skewed towards the left with very long tail on the right. It could be concluded that CO2 emission per capita is not normally distributed. Because a small number of data points were much larger than the bulk of our data, it is advisable to examine the logarithm of our CO2 emission per capita.

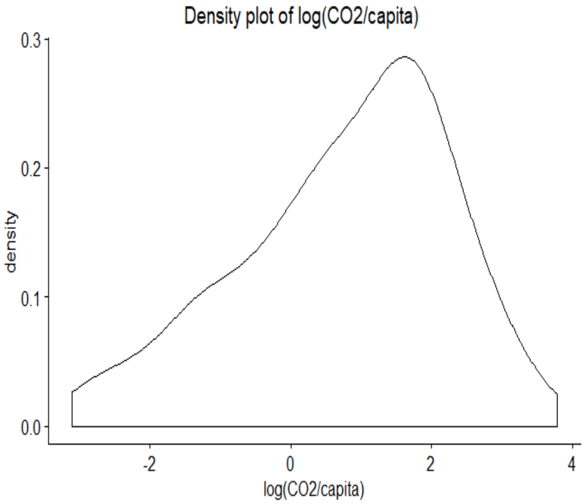
The analysis was repeated, but with the natural logarithm (ln) of CO2 emission per capita.



```
	Shapiro-Wilk normality test

data:  log(combined_cast$EN.ATM.CO2E.PC)
W = 0.96772, p-value = 0.0001262
```

The p-value improves even though it is still less than 0.05. Due to the high sensitivity of this test, we investigated the Q-Q and density plot of logarithm of CO2/capita.

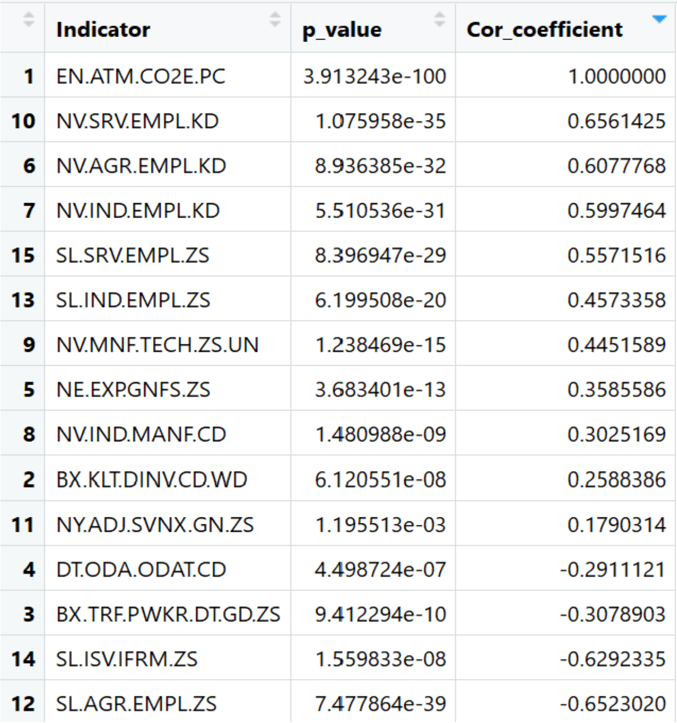




Compared to the previous graphs, the points in this Q-Q plot fell along the reference line more closely, and the density plot resembled a normal distribution more than the case of unmodified CO2 per capita although there was still some deviation. In brief, the distribution of logarithm might not be perfect, but it is more similar to a normal distribution than the pure CO2 emission data. As a result, the logarithm transformation would be more suitable for our model in part 4 of this report.

2.4.4 Correlation analysis

Because the latest data for CO2 emission per capita is in 2014, we also chose 2014 data of other indicators to conduct a quick correlation study. The correlation analysis in R quantified the relationship between the above indicators and CO2 emission per capita (EN.ATM.CO2E.PC). Kendall method was chosen instead of Spearman method because Spearman is sensitive to inconsistency in the data. In this section, we just want to obtain a quick look at how strong the correlation is. The table below shows the correlation coefficients (and associated p-values) of the chosen indicators with CO2 emission per capita (the first row is correlation of CO2 emission per capita with itself, thus tau =1). The 14 indicators mentioned in Section 2.2 was taken for the study.



	Indicator	p_value	Cor_coefficient
1	EN.ATM.CO2E.PC	3.913243e-100	1.0000000
10	NV.SRV.EMPL.KD	1.075958e-35	0.6561425
6	NV.AGR.EMPL.KD	8.936385e-32	0.6077768
7	NV.IND.EMPL.KD	5.510536e-31	0.5997464
15	SL.SRV.EMPL.ZS	8.396947e-29	0.5571516
13	SL.IND.EMPL.ZS	6.199508e-20	0.4573358
9	NV.MNF.TECH.ZS.UN	1.238469e-15	0.4451589
5	NE.EXP.GNFS.ZS	3.683401e-13	0.3585586
8	NV.IND.MANF.CD	1.480988e-09	0.3025169
2	BX.KLT.DINV.CD.WD	6.120551e-08	0.2588386
11	NY.ADJ.SVNX.GN.ZS	1.195513e-03	0.1790314
4	DT.ODA.ODAT.CD	4.498724e-07	-0.2911121
3	BX.TRF.PWKR.DT.GD.ZS	9.412294e-10	-0.3078903
14	SL.ISV.IFRM.ZS	1.559833e-08	-0.6292335
12	SL.AGR.EMPL.ZS	7.477864e-39	-0.6523020

The above results table revealed fascinating results. The below table shows indicators that have high correlation with CO2 emission per capita:

Indicator Code	Indicator Name
NV.AGR.EMPL.KD	Agriculture, value added per worker (constant 2010 US$)
NV.IND.EMPL.KD	Industry, value added per worker (constant 2010 US$)
NV.SRV.EMPL.KD	Services, value added per worker (constant 2010 US$)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)

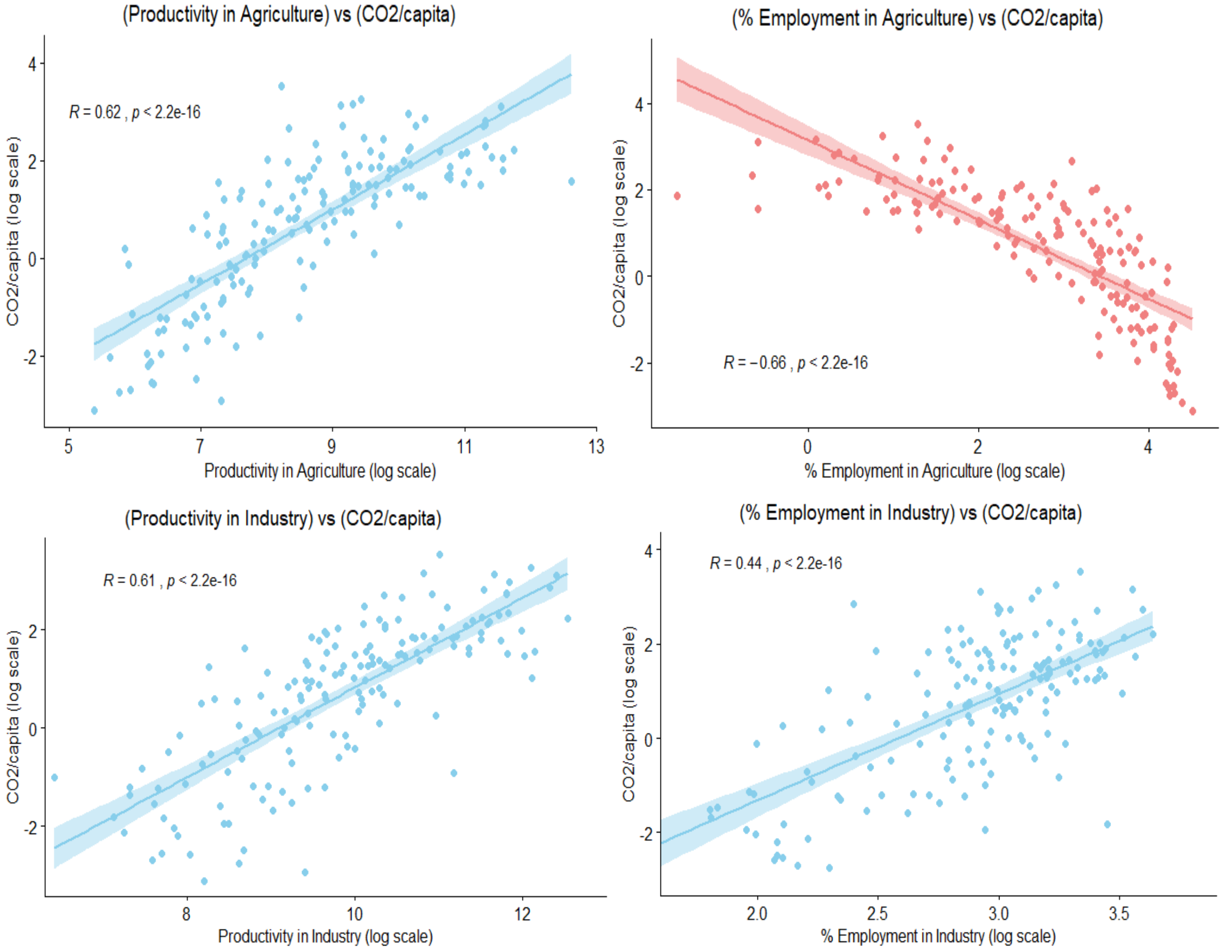
Further we divide these indicators into two groups

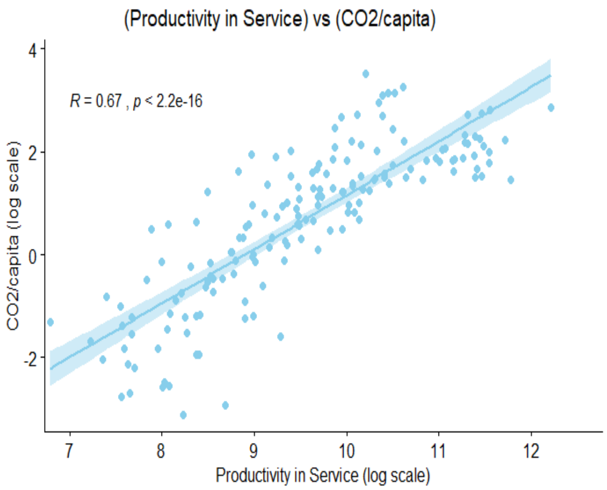
Value added per worker (productivity):

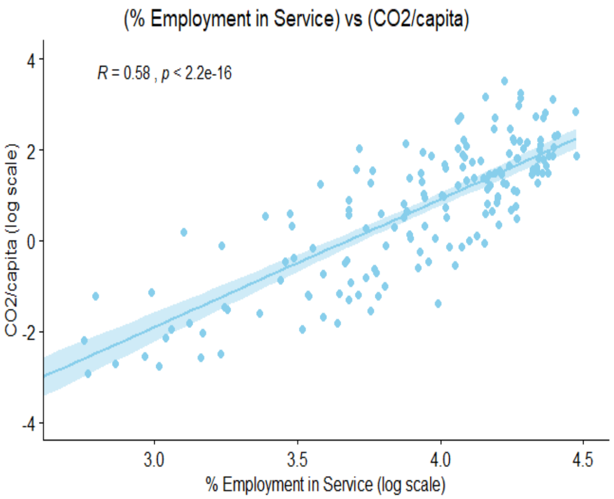
Agriculture, value added per worker (constant 2010 US$)

Industry, value added per worker (constant 2010 US$)
Services, value added per worker (constant 2010 US$)
Percentage of employment in each sector:
Employment in agriculture (% of total employment) (modeled ILO estimate)
Employment in industry (% of total employment) (modeled ILO estimate)
Employment in services (% of total employment) (modeled ILO estimate)

Out of 6 indicators, 5 have positive correlation while SL.AGR.EMPL.ZS (% employment in agriculture) has negative correlation with CO2 emissions per capita. The following graphs displayed simple linear regressions of these indicators (blue means positive, red means negative).

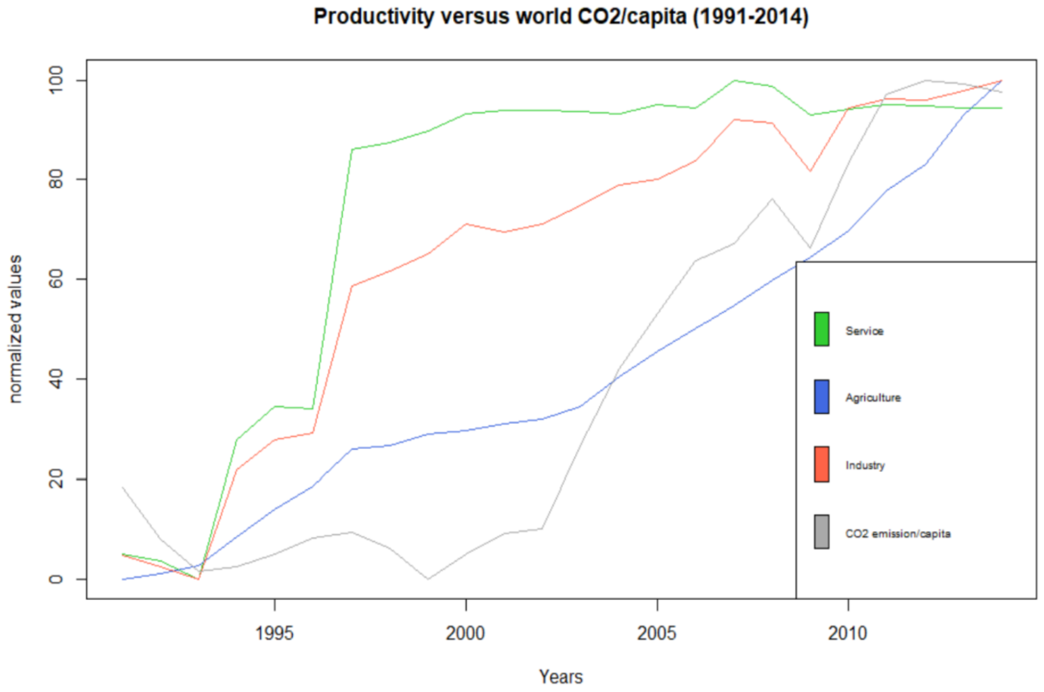






2.4.5 Productivity vs CO2 emission per capita.

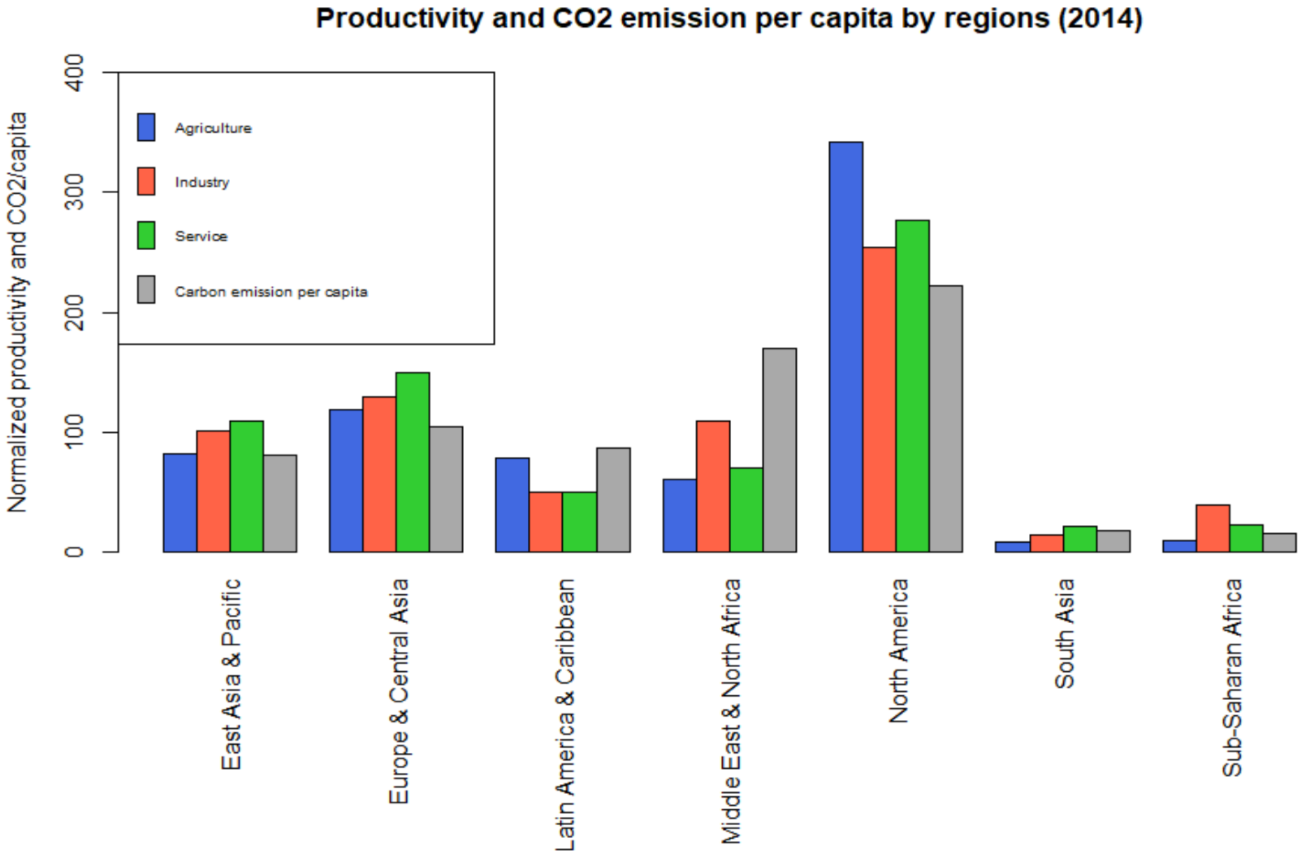
The correlation between productivity and CO2 can be graphically represented over the years. As mentioned above, only for the years 1991-2014 data existed for average world CO2 emission per capita and productivity expressed in net value added per worker (Service, Industry and Agriculture). Graph 2.4.5.1 below shows a normalized time-series over the period 1991-2014.



Graph 2.4.5.1: Productivity per sector (net value added per worker) versus world average CO2/capita for the years 1991-2014.

According to the graph, during the financial crisis around 2008, there was a slight drop in productivity for the sectors Industry and Service. Similarly, the CO2 emission/capita dropped considerably during that year. However, the productivity per worker for agriculture has seen a continuous rise over the years.

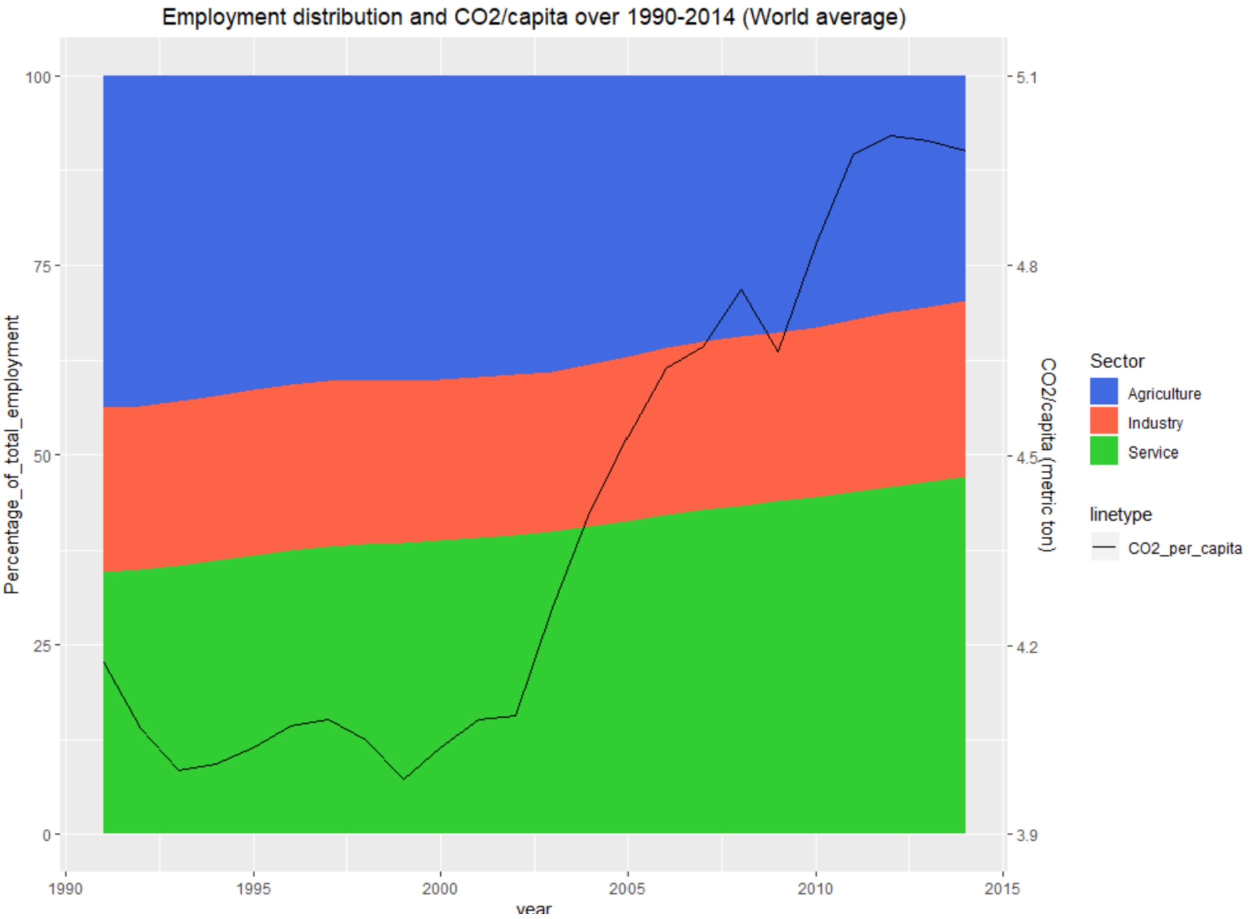
Furthermore, it is interesting to plot the productivity for these sectors versus CO2/capita for different world regions. As such, a further insight can be gained on these relationships as well as their distribution around the world. The year chosen is 2014 (i.e. the latest year for which this data has been available) and the result is shown below in graph 2.4.5.2.



Graph 2.4.5.2: Different normalized sector productivities versus CO2 emissions per capita for the world regions as of 2014.

As can be seen from the graph, the region of North America had the largest CO2/emission per capita and the highest productivity for all sectors. Meanwhile, for the Middle East region, it is very clear that the fossil fuel industry contributes to the high CO2 emission even though their productivity is relatively low. The African and South Asian countries rank low in both per-capita emissions and sector productivities, which correlates to the lower economic development in these regions.

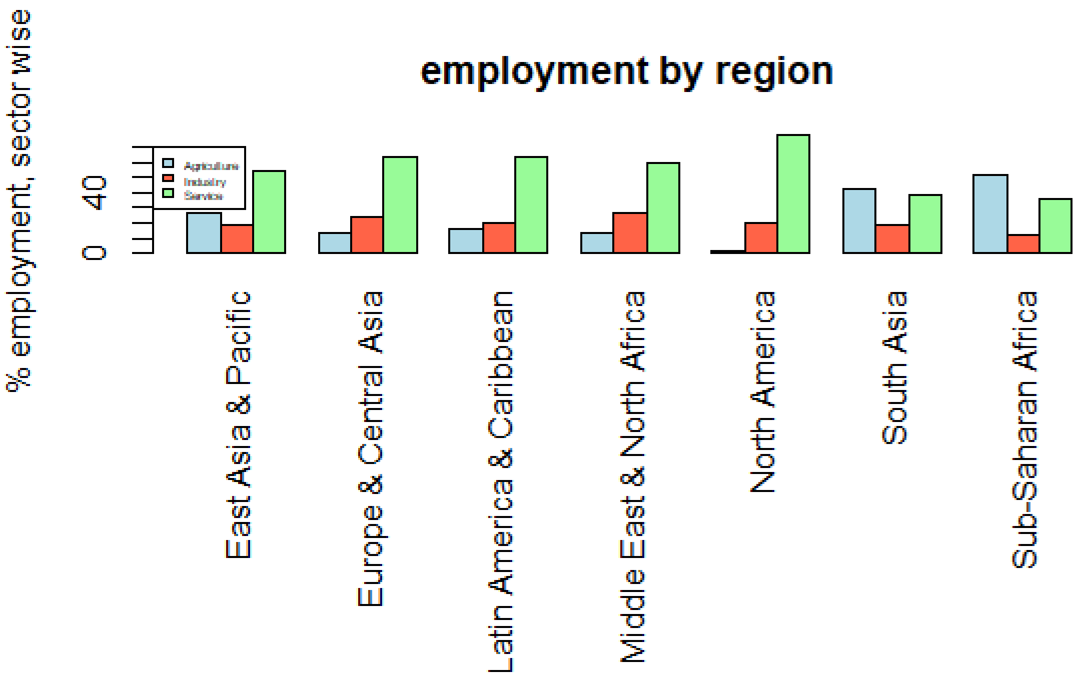
2.4.6 Percentage of employment vs CO2 emission per capita



Graph 2.4.6.1: Employment distribution and CO2 per capita over 1990-2014

The Graph represents the percentage of people employed in different sectors and the CO2 emissions per capita. The service industry has risen steadily over the years, suggesting more people have moved to work in the service industry. This rise correlates with the increase of CO2 emissions per capita over the period. This trend is consistent with our correlation analysis (section 2.4.4) that the % employment in the service industry is positively correlated to the CO2 emission per capita.

In contrast, the percentage of people employed in the agriculture industry is steadily decreasing. The correlation analysis shows that the % employment in the agriculture industry is negatively correlated to the CO2 emission. Finally, the percentage of employment in industry sector has been relatively stable over the years.



Graph 2.4.6.2: Employment distribution by region

The above graph shows the % of employment (sector wise) in different regions. North America has the highest percentage of service employment and the lowest percentage of agriculture employment. In contrast, South Asia and Sub-Saharan Africa have low percentage of service and high percentage of agricultural employment. Comparing with CO2 per capita of these region (section 2.4.2), it is clear that when there are more people working in the service sector and fewer people working in agriculture, the CO2 per capita will become be higher.

3. Data Preparation

We analyzed data for all countries (163 Countries after removing all the NA) in one particular year 2014. The reason why 2014 was selected was mentioned in section 2.1 & 2.2 and highlighted when needed in the sections below.

This multi-country in one year (2014) approach was chosen because we wanted to compare and visualize the economies of different countries during a particular year, and then show the relationship between different important independent variables (explained below) and its influence on CO2 emissions.

We argue that it will be biased for global decision making if we present the visualization as one country for different years. CO2 emissions is a global problem, but different countries have different emission patterns (depending on technology advancement, regulation, etc.). We want to understand the effects on the global scale and also understand the indicators which plays a major role in CO2 emissions.

3.1 Data Selection

Among the 14 socio-economic indicators investigated for correlation (section 2.4.3), we chose 6 indicators which show high correlation:

Indicator Code	Indicator Name
NV.AGR.EMPL.KD	Agriculture, value added per worker (constant 2010 US$)
NV.IND.EMPL.KD	Industry, value added per worker (constant 2010 US$)
NV.SRV.EMPL.KD	Services, value added per worker (constant 2010 US$)
SL.AGR.EMPL.ZS	Employment in agriculture (% of total employment) (modeled ILO estimate)
SL.IND.EMPL.ZS	Employment in industry (% of total employment) (modeled ILO estimate)
SL.SRV.EMPL.ZS	Employment in services (% of total employment) (modeled ILO estimate)

Another indicator, SL.ISV.IFRM.ZS (% of informal employment), ranked high among the other groups . However, a brief check of the dataset indicates that only 25 countries have data for the percentage of informal employment. Due to the lack of data, we decided to exclude informal employment from our analysis.

CO2 emission per capita is the dependent variable, while the other 6 indicators are independent variables in this model. The causal diagram (section 4.2) resembles a collider, which will be chosen to model the relationship between economic factors and CO2 emission per capita.

3.2 Data Cleaning

The data was obtained from the World bank databank , so most of it was consistent and clearly identified. We did not have many issues with data cleaning and preparation.

The World bank data is available for all countries and geographical regions in the world. Since we are interested in studying our problem on country, region and world level, we subset only the required data to plot the graph and left the other data. For example, we took the data based on countries and did not use data for groups of countries such as Arab World.

This project aims to identify the indicators that have high potential of affecting the CO2 emission per capita. Since the latest data for CO2 emission per capita is in 2014, we also chose 2014 data for other indicators. The data that we used to plot the graphs consists of 163 countries after removing all the NA, and we are interested to explore for the year 2014 .

3.3 Integrate and Format data

It was a straightforward process to integrate and format the data, since the data was listed per country in alphabetical order. We proceeded to prepare the dataset for modelling. A quick check showed that all of our data were in numeric values, which was ready for the next step.

4. Model

4.1 Test Design

First, our model will be checked for convergence to ensure:

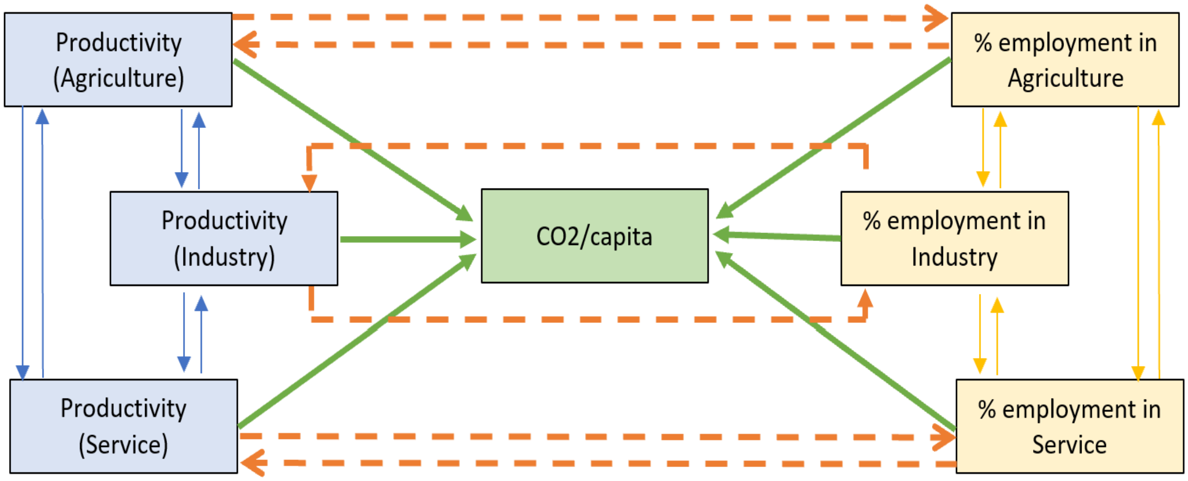
- Trace plot of 2 different chains (with different automatically generated initial values) is consistent
- Density plot of the obtained coefficients show no abnormal patterns
- Gelman plot show high similarity between 2 chains
- There is low autocorrelation (approximately 0)
- Effective sizes are high

Next, the coefficients obtained by our Bayesian model will be cross-checked against the coefficients from a simple linear model.

Finally, posteriors predicted by our model will be compared against actual values (from WDI data) to see whether there is any significant bias.

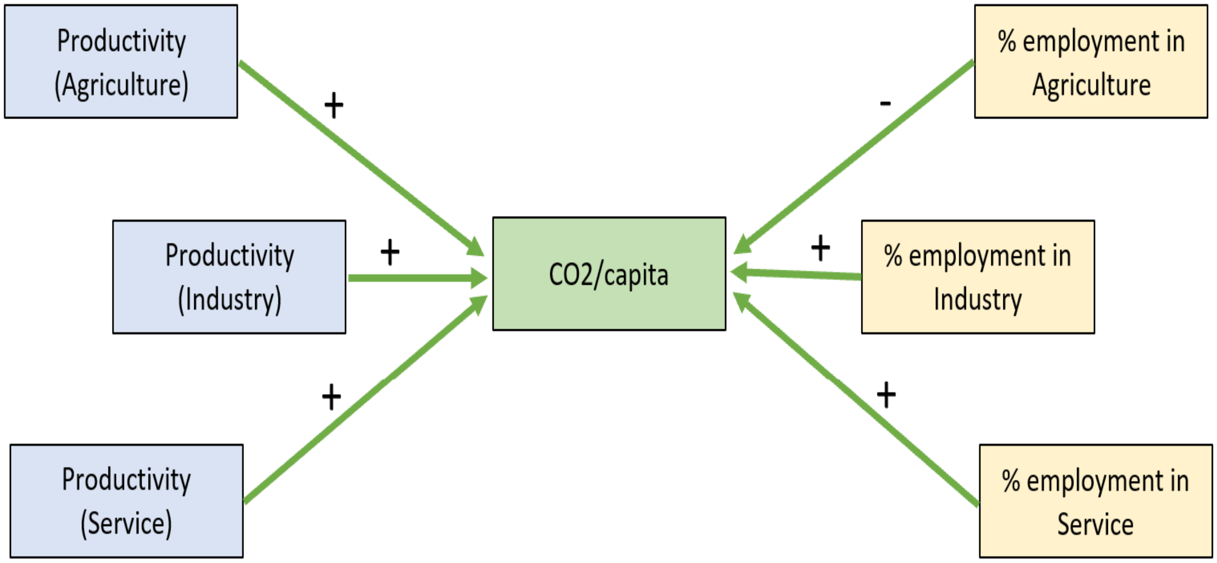
4.2 Model Selection

The interdependencies among the chosen indicators can be visualized by a causal diagram:



The CO2 emission per capita is put at the middle. The productivity group is put on the left (blue boxes) and the employment percentage group is put on the right (yellow boxes). All 3 productivity and 3 employment percentage indicators affect the CO2 emission (green arrows). For example, higher productivity in industry means that the economy is more developed and more technologically advanced, thus requiring more energy supply and generating more CO2 as a result. Within each group (productivity and employment percentage), the indicators for different sectors influence each other (blue and yellow arrows). For instance, a higher number of people working in industry and service means that fewer number work in agriculture. Higher productivity in industry (more advanced manufacturing and technology) is likely to cause higher productivity in agriculture (more use of machinery) and service (more efficient work procedures). Finally, an indicator in one group can affect the corresponding indicator in the other group (orange arrows). To illustrate, higher productivity in agriculture means that more machinery and less human labor is required to do the job, resulting in lower percentage of employment in agriculture.

Due to the time and resource constraints of this project, it is impossible to model the above complicated diagram. Therefore, the causal diagram was simplified to show only the effect of 6 indicators on CO2 emissions, becoming a directed causal graph (shown below). Percentage of employment in agriculture is negatively correlated with CO2 per capita, whereas the other 5 indicators are positively correlated (which was proved in section 2.4.2 correlation analysis).



There are 3 main types of directed acyclic graphs: chain, collider and fork. Our simplified diagram resembles a collider, which will be chosen to model the relationship between economic factors and CO2 emission per capita. CO2 emission per capita is the dependent variable, while the other 6 indicators are independent variables in this model.

4.3 Model #1 Parameter Settings & Model Description

As explained in section 2.4, the logarithm transformation of CO2 emission is more similar to a normal distribution. Therefore, the variables are transformed as follows:

y <- log(CO2 per capita)
x1 <- log(Productivity in Agriculture)
x2 <- log(Productivity in Industry)
x3 <- log(Productivity in Service)
x4 <- log(Employment percentage in Agriculture)
x5 <- log(Employment percentage in Industry)
x6 <- log(Employment percentage in Service)

It is assumed that our dependent variable follows a normal distribution.
The mean value of the distribution has linear relationships with regards to 6 independent variables (beta0 is the intercept while beta1 to beta 6 is the corresponding slope of 6 independent variables).

Likelihood:

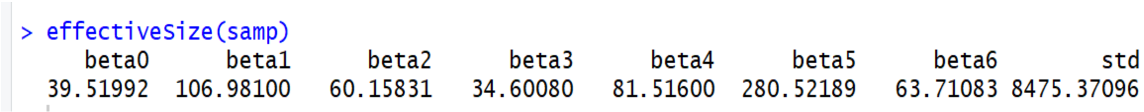
```
mu <- beta0 + beta1*x1 + beta2*x2 + beta3*x3 + beta4*x4 + beta5*x5 + beta6*x6
  y   ~ dnorm(mu[i],inv_var)
```

Priors:
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
beta3 ~ dnorm(0,0.0001)
beta4 ~ dnorm(0,0.0001)
beta5 ~ dnorm(0,0.0001)
beta6 ~ dnorm(0,0.0001)
inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)
Other parameters for the first run (they will be adjusted after assessment):
Number of chains: 2
Burn-in: 10,000
Number of iterations: 20,000
Thinning interval: 1

4.4 Model #1 Discussion and Assessment

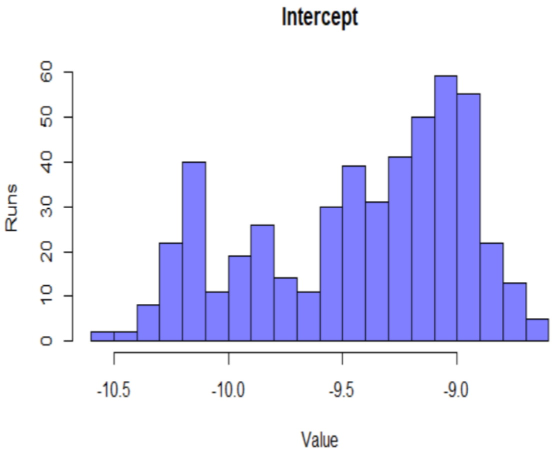
The model #1 causes a few abnormal patterns:

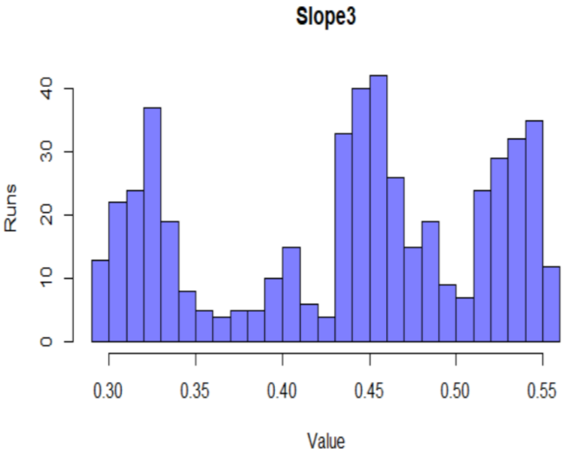
1. It produced very low effective size for all betas:



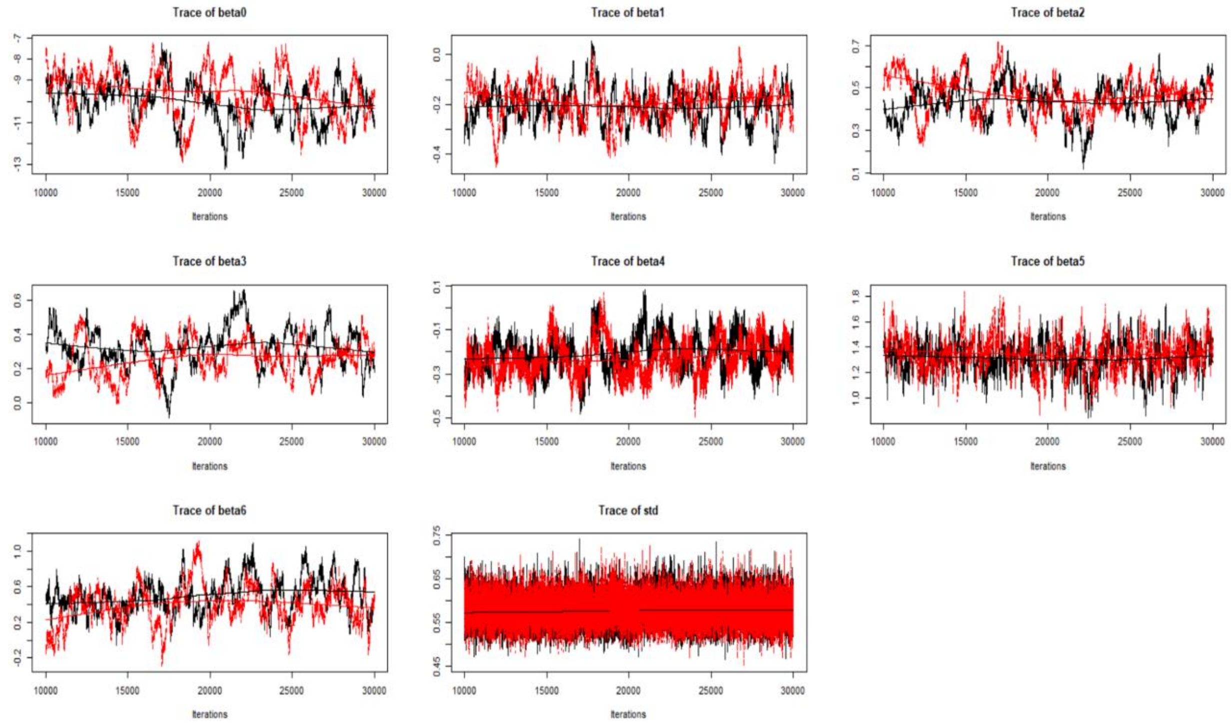
```
> effectiveSize(samp)
     beta0      beta1      beta2      beta3      beta4      beta5      beta6        std
  39.51992  106.98100   60.15831   34.60080   81.51600  280.52189   63.71083 8475.37096
```

2. Some betas have Uneven distribution:

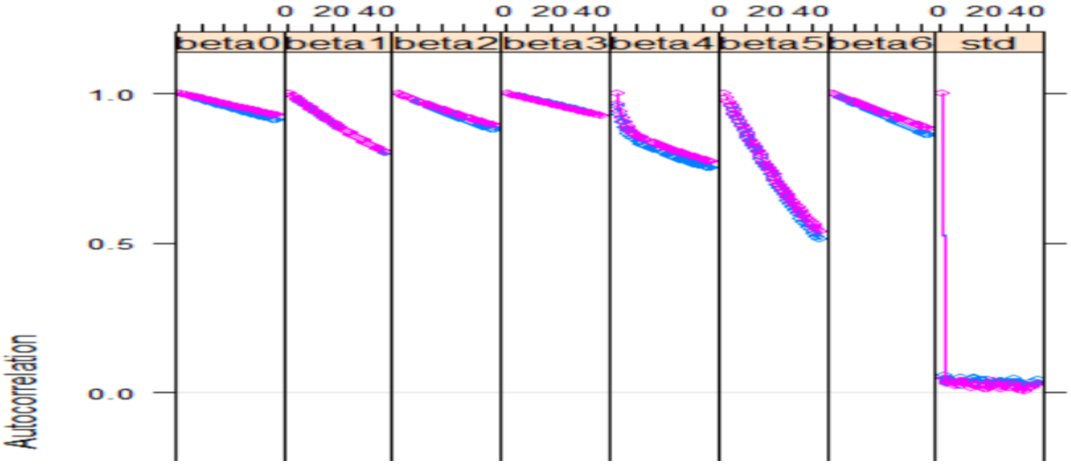




3. Trace of betas for 2 chains did not converge:

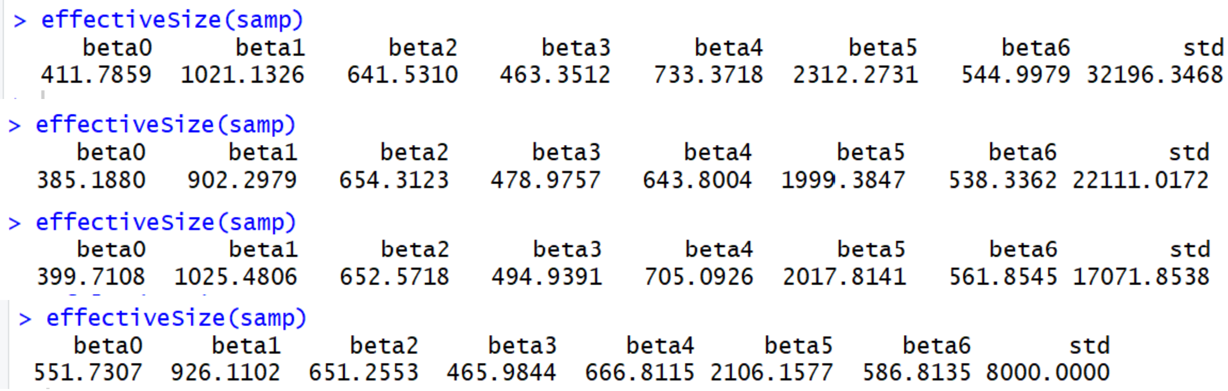


4. There is strong autocorrelation for all betas in the model, as shown in acf graph:



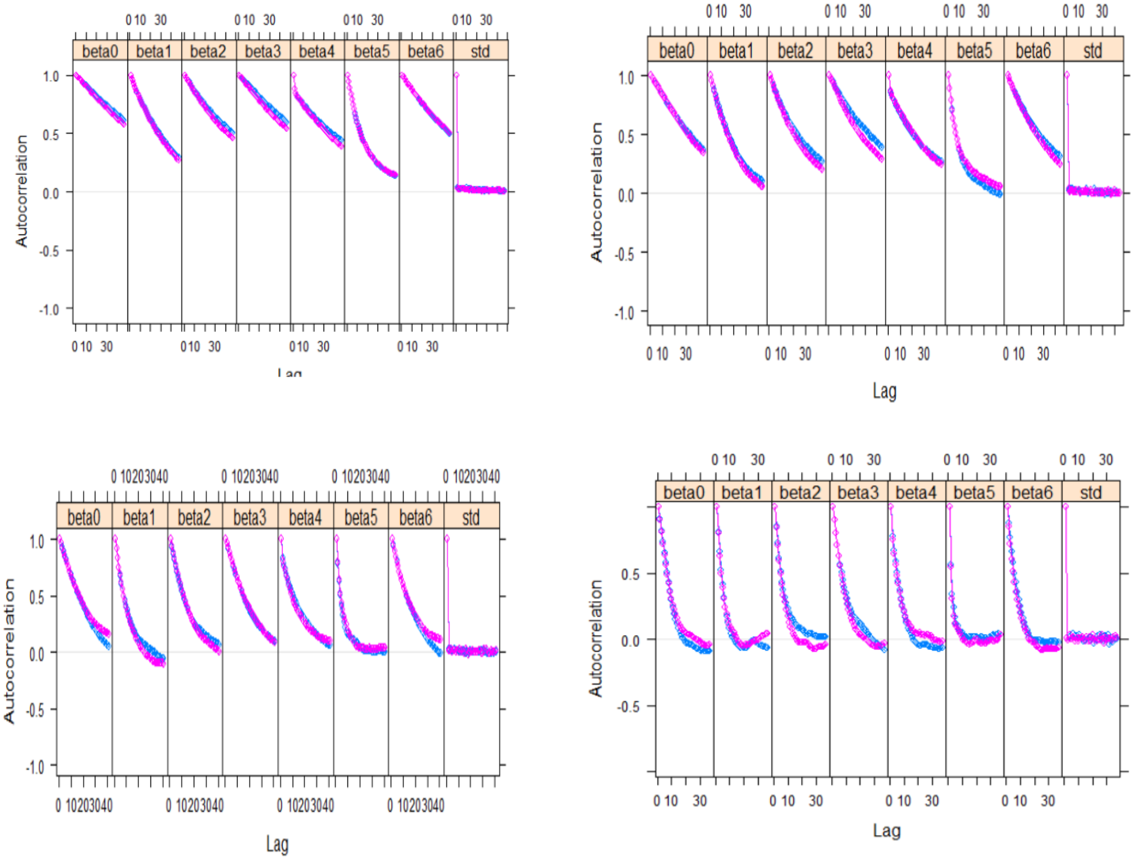
Subsequently, the numbers of burn-in and iterations were increased to 100,000 and 200,000 respectively. The thinning interval was also raised to 5, then 10, 20 and finally 50. However, the model still showed low effective sizes and autocorrelation.

Effective size for thinning interval 5, 10, 20 and 50 (from top to bottom):



```
> effectiveSize(samp)
    beta0      beta1      beta2      beta3      beta4      beta5      beta6        std
 411.7859  1021.1326   641.5310   463.3512   733.3718  2312.2731   544.9979 32196.3468

> effectiveSize(samp)
    beta0      beta1      beta2      beta3      beta4      beta5      beta6        std
 385.1880   902.2979   654.3123   478.9757   643.8004  1999.3847   538.3362 22111.0172

> effectiveSize(samp)
    beta0      beta1      beta2      beta3      beta4      beta5      beta6        std
 399.7108  1025.4806   652.5718   494.9391   705.0926  2017.8141   561.8545 17071.8538

> effectiveSize(samp)
    beta0     beta1     beta2     beta3     beta4     beta5     beta6       std
 551.7307  926.1102  651.2553  465.9844  666.8115 2106.1577  586.8135 8000.0000
```

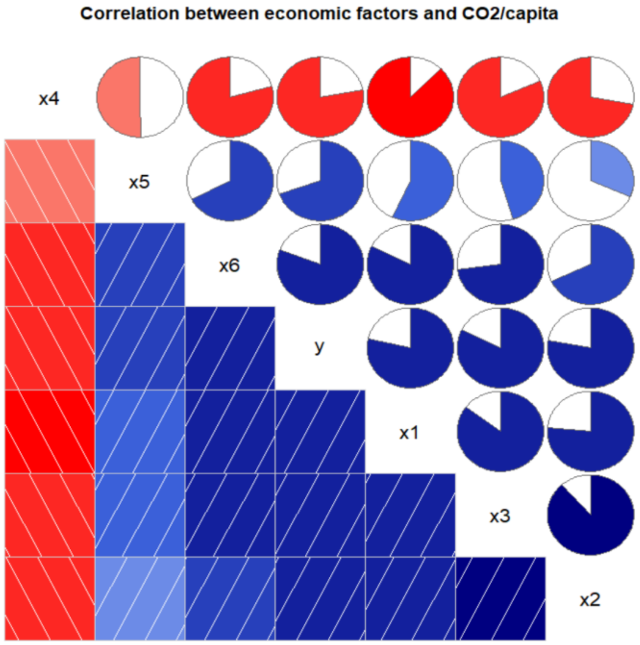
Autocorrelation for thinning interval 5 (top left), 10 (top right), 20 (bottom left) and 50 (bottom right):



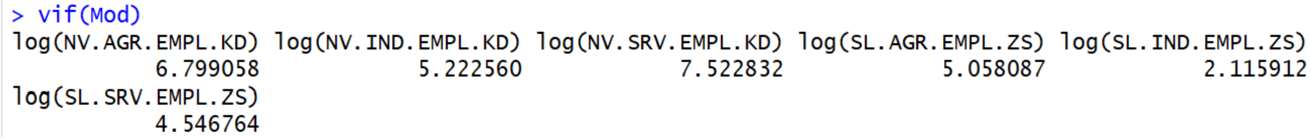
4.5 Multicollinearity among Independent variables

It is observed that our 6 independent variables correlate with each other (multicollinearity), further analysis was conducted:

1. The correlation matrix plot demonstrated high correlation among the indicators. Blue means positive correlation, and red means negative correlation. The darker the color (and the larger the shaded region of the pie), the stronger the correlation (y is CO2 per capita, x1 to x6 are the independent variables)



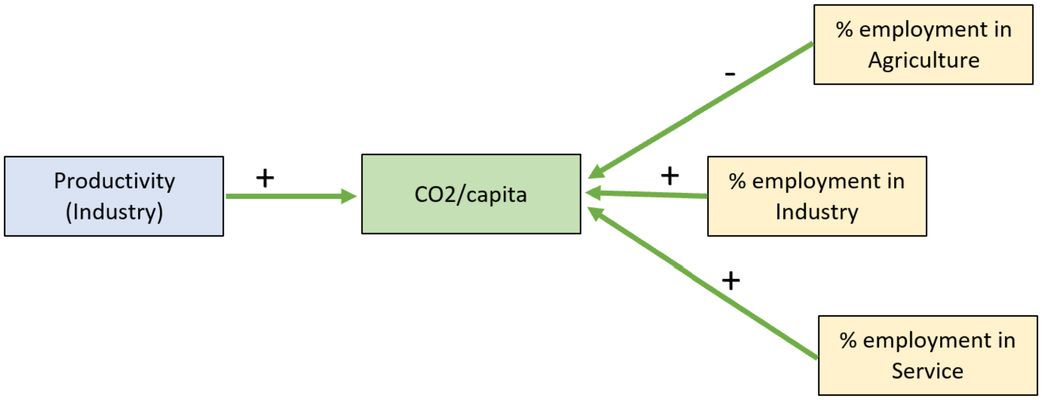
2. VIF test showed multicollinearity among the independent variables:



```
> vif(Mod)
log(NV.AGR.EMPL.KD) log(NV.IND.EMPL.KD) log(NV.SRV.EMPL.KD) log(SL.AGR.EMPL.ZS) log(SL.IND.EMPL.ZS)
           6.799058            5.222560            7.522832            5.058087            2.115912
log(SL.SRV.EMPL.ZS)
           4.546764
```

4.6 Model #2 Parameter Settings & Model Description

To improve the model, we tried removing 2 independent variables with the highest VIF: Productivity in Agriculture NV.AGR.EMPL.KD (x1) and Productivity in Service NV.SRV.EMPL.KD (x3). This leads us to Model #2.



Likelihood:

mu <- beta0 + beta2*x2 + beta4*x4 + beta5*x5 + beta6*x6
  y   ~ dnorm(mu[i],inv_var)

Priors:

beta0 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
beta4 ~ dnorm(0,0.0001)
beta5 ~ dnorm(0,0.0001)
beta6 ~ dnorm(0,0.0001)
inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)

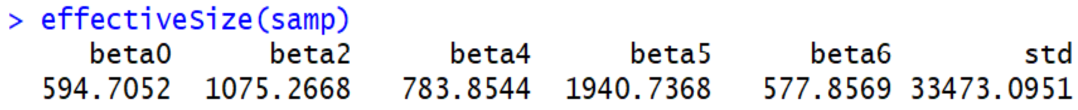
Others:

Number of chains: 2
Burn-in: 100,000
Number of iterations: 200,000
Thinning interval: 10

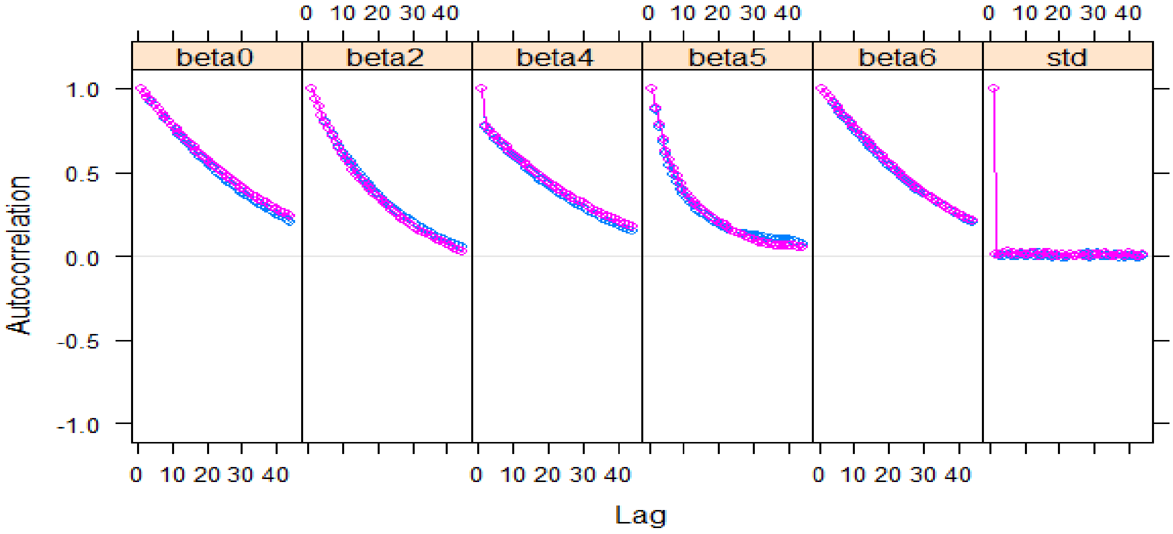
4.7 Model #2 Discussion and Assessment

Some of the noted observations after running the model 2

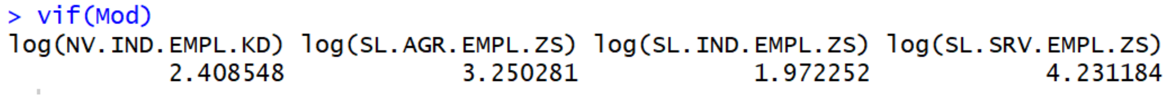
1. Again, there are low effective sizes and strong autocorrelation:



```
> effectiveSize(samp)
     beta0      beta2      beta4      beta5      beta6        std 
  594.7052  1075.2668   783.8544  1940.7368   577.8569 33473.0951 
```



2. VIF was checked again:

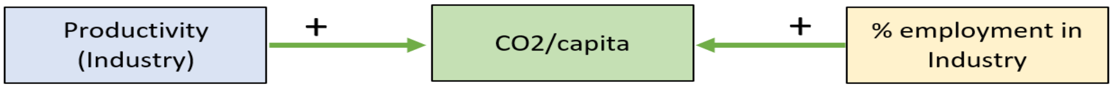


```
> vif(Mod)
log(NV.IND.EMPL.KD) log(SL.AGR.EMPL.ZS) log(SL.IND.EMPL.ZS) log(SL.SRV.EMPL.ZS)
           2.408548            3.250281            1.972252            4.231184
```

Two variables with the highest VIF (>2.5) were further removed: Percentage of employment in Agriculture SL.AGR.EMPL.ZS (x4) and Percentage of employment in Service SL.SRV.EMPL.ZS (x6).

We are trying to select the variables which will produce a perfect model converge and this leads us to try the Model 3.

4.8 Model #3 Parameter Settings & Model Description



Likelihood:

mu <- beta0 + beta2*x2 + beta5*x5

y ~ dnorm(mu[i],inv_var)

Priors:

beta0 ~ dnorm(0,0.0001)

beta2 ~ dnorm(0,0.0001)

beta5 ~ dnorm(0,0.0001)

inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)

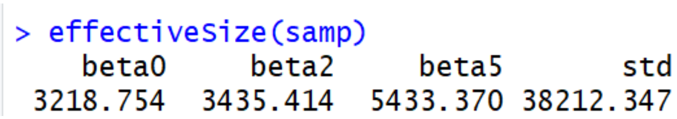
Others:

Number of chains: 2
Burn-in: 100,000
Number of iterations: 200,000
Thinning interval: 10

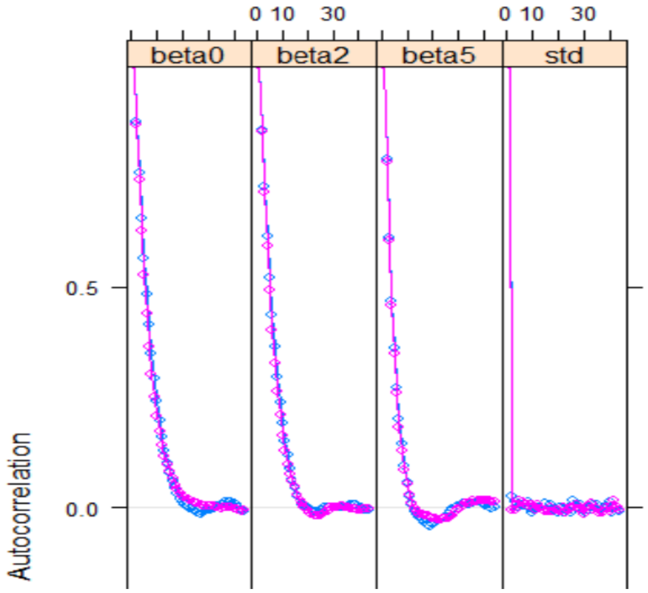
4.9 Model #3 Discussion and Assessment

The new model shows the following below observation

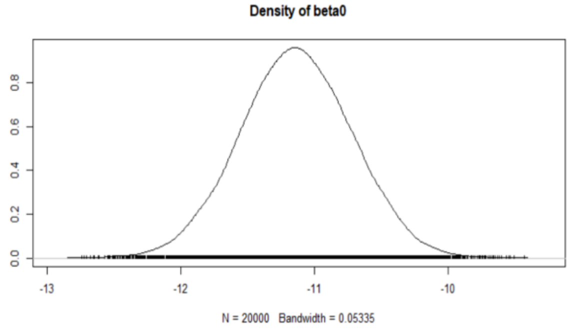
1. Effective size and acf plot improved:

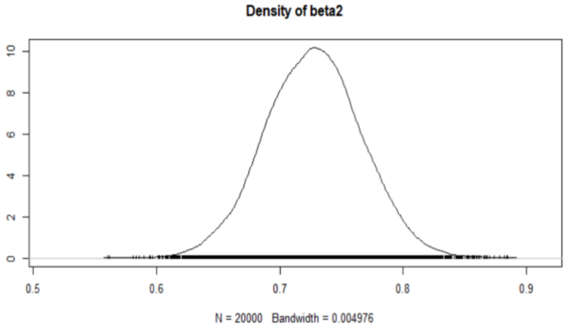


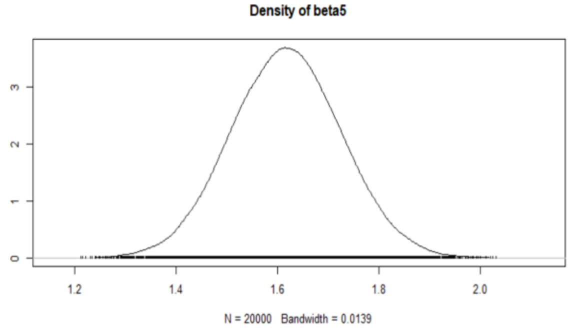
```
> effectiveSize(samp)
   beta0     beta2     beta5       std 
3218.754  3435.414  5433.370 38212.347 
```

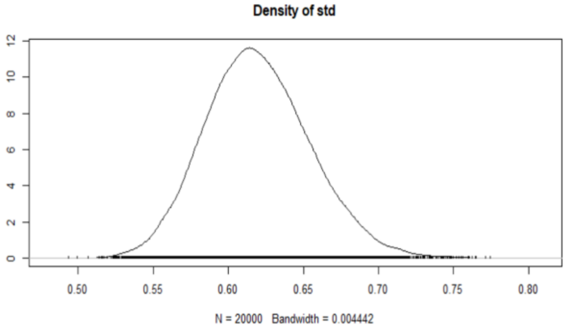


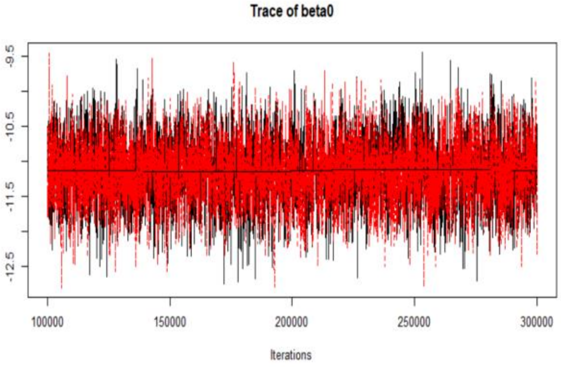
2. Density, Trace and Gelman plots are acceptable (there is no abnormal patterns):

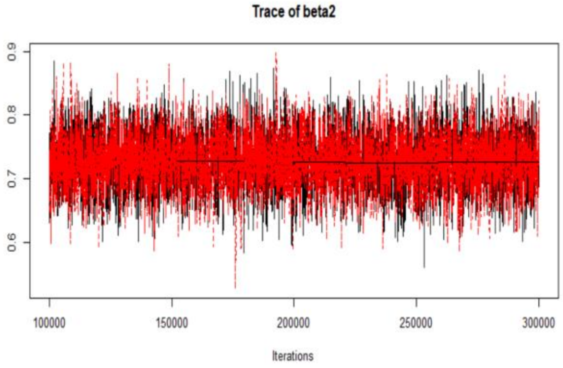

Density of beta0
0.0
0.2
0.4
0.6
0.8
-13
-12
-11
-10
N = 20000 Bandwidth = 0.05335

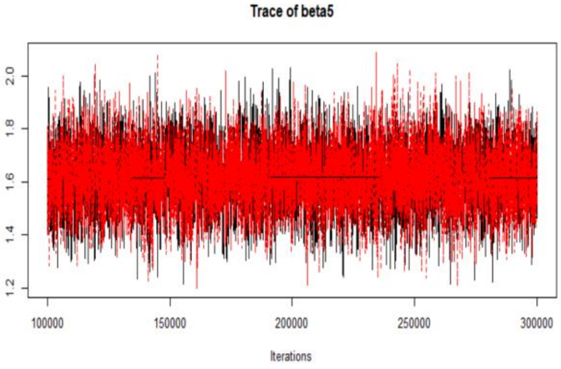

Density of beta2
0
2
4
6
8
10
0.5
0.6
0.7
0.8
0.9
N = 20000 Bandwidth = 0.004976

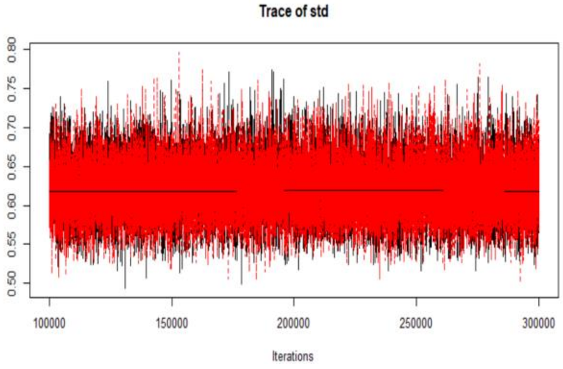

Density of beta5
0
1
2
3
1.2
1.4
1.6
1.8
2.0
N = 20000 Bandwidth = 0.0139

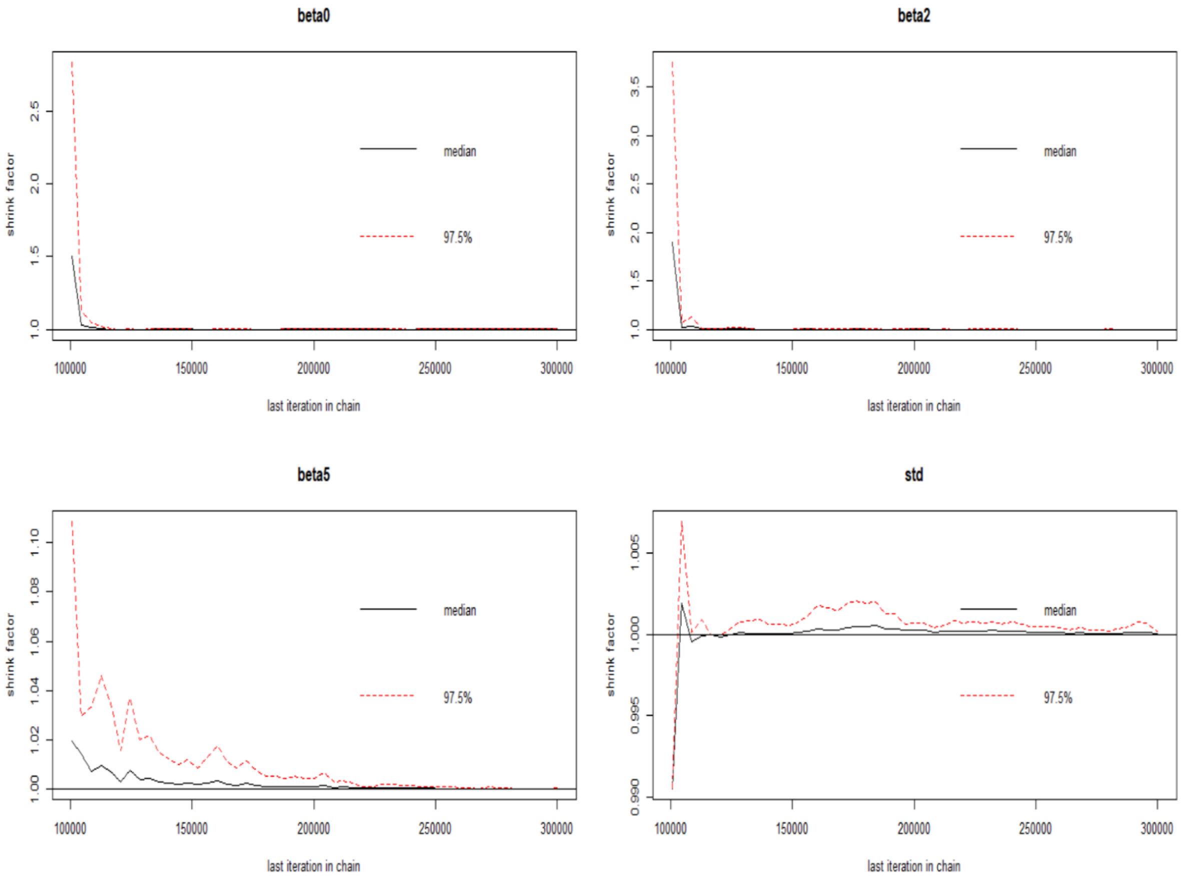

Density of std
0
2
4
6
8
10
12
0.50
0.55
0.60
0.65
0.70
0.75
0.80
N = 20000 Bandwidth = 0.004442


Trace of beta0
-9.5
-10.5
-11.5
-12.5
100000
150000
200000
250000
300000
Iterations

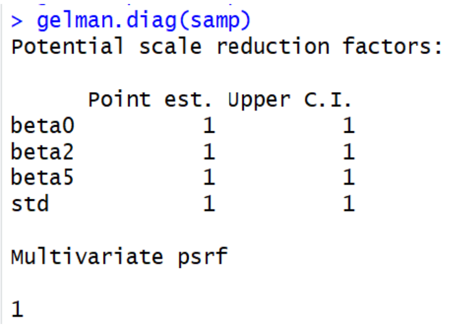

Trace of beta2
0.9
0.8
0.7
0.6
100000
150000
200000
250000
300000
Iterations


Trace of beta5
2.0
1.8
1.6
1.4
1.2
100000
150000
200000
250000
300000
Iterations


Trace of std
0.80
0.75
0.70
0.65
0.60
0.55
0.50
100000
150000
200000
250000
300000
Iterations



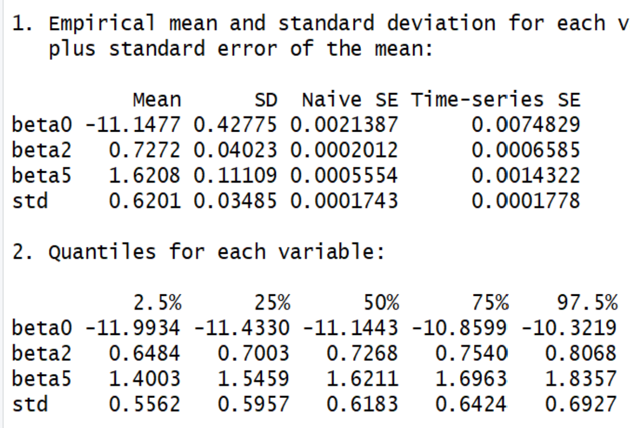
3. Gelman diagnosis showed value of 1 for all:



```
> gelman.diag(samp)
Potential scale reduction factors:

      Point est. Upper C.I.
beta0          1          1
beta2          1          1
beta5          1          1
std            1          1

Multivariate psrf

1
```

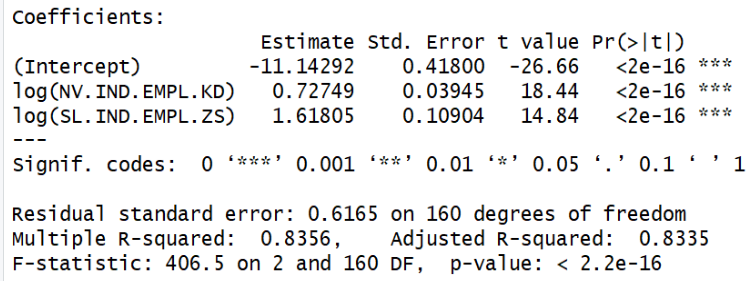
We can conclude that Model #3 converge.

4. A quick look at the obtained values of beta0, beta2 and beta5 showed the significance of these coefficients. Standard errors are low (last columns in first table). In the second table, all coefficients do not change sign from positive to negative (or vice versa) within 95% confidence interval, showing their significant contribution to predict the dependent variable. Beta2 and beta5 are positive, which is consistent with the positive correlation in our causal diagram.



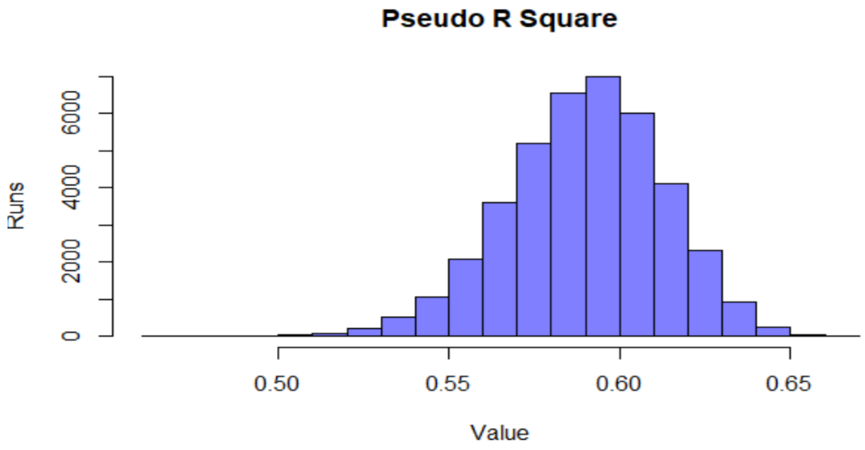
```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean      SD  Naive SE Time-series SE
beta0 -11.1477 0.42775 0.0021387      0.0074829
beta2   0.7272 0.04023 0.0002012      0.0006585
beta5   1.6208 0.11109 0.0005554      0.0014322
std     0.6201 0.03485 0.0001743      0.0001778

2. Quantiles for each variable:

          2.5%      25%      50%      75%    97.5%
beta0 -11.9934 -11.4330 -11.1443 -10.8599 -10.3219
beta2   0.6484   0.7003   0.7268   0.7540   0.8068
beta5   1.4003   1.5459   1.6211   1.6963   1.8357
std     0.5562   0.5957   0.6183   0.6424   0.6927
```

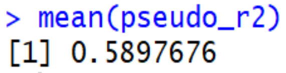
5. The coefficients obtained from our Bayesian model are close to those obtained from a simple linear model, confirming the accuracy of our model #3. The below is coefficients obtained from a linear model:



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -11.14292    0.41800  -26.66   <2e-16 ***
log(NV.IND.EMPL.KD)    0.72749    0.03945   18.44   <2e-16 ***
log(SL.IND.EMPL.ZS)    1.61805    0.10904   14.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6165 on 160 degrees of freedom
Multiple R-squared:  0.8356,	Adjusted R-squared:  0.8335 
F-statistic: 406.5 on 2 and 160 DF,  p-value: < 2.2e-16
```

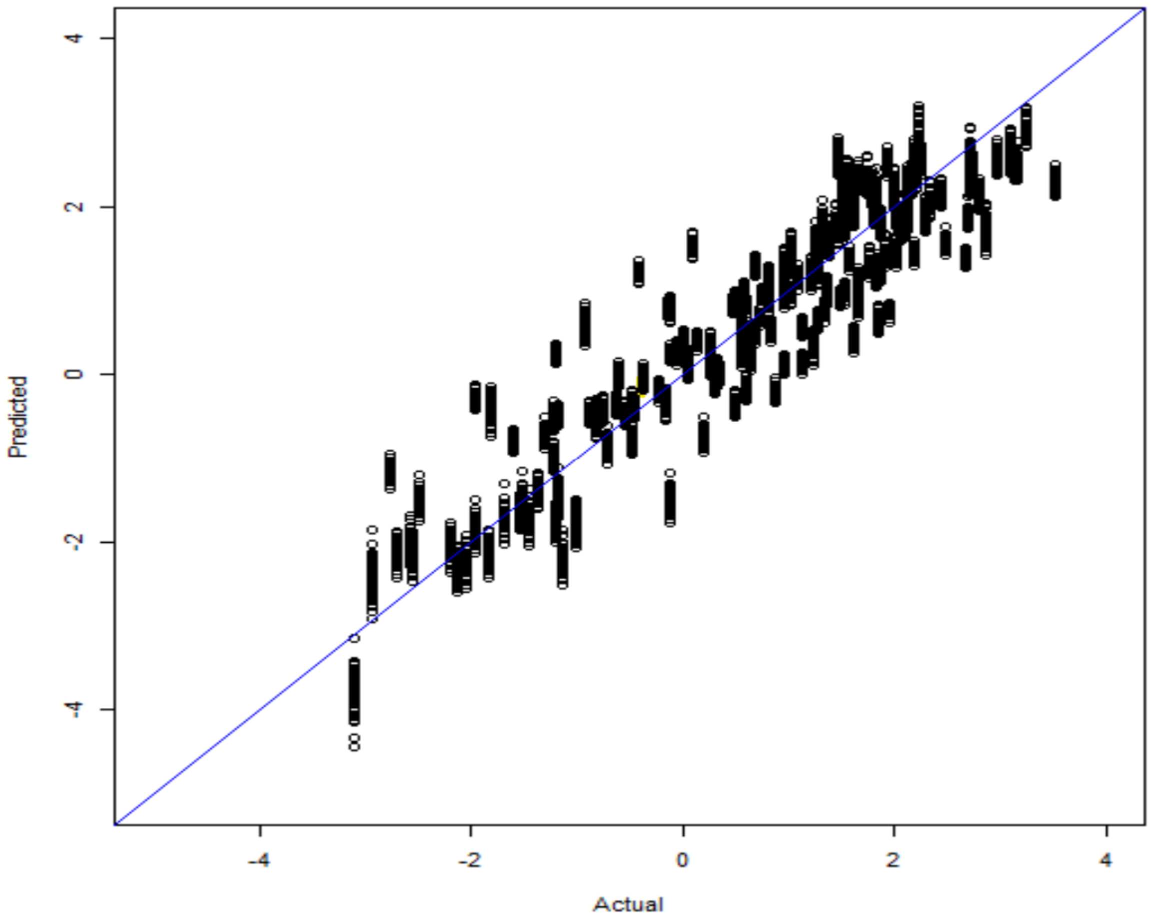
6. Together, Productivity in Industry (x2) and Percentage of employment in Industry (x5) can predict CO2 per capita (y) with R square = 0.59





```
> mean(pseudo_r2)
[1] 0.5897676
```

7. The predicted posteriors is plotted against actual values. If a data point falls exactly on the reference line (45 degree line), the predicted value is the same as the actual value (100% accuracy). All resulting points are close to the reference line, confirming high accuracy of our Bayesian model.



5. Conclusion

The objective of this report is to study which economic indicators (apart from GDP) influence the CO2 emission per capita of a country. The data was taken from the world bank database. The report focused on the indicators that are related to the economy of a country and zeroed in two groups - “Economic Policy & Debt” and “Social Protection & Labor: Economic activity”. In this report, the visualization and the modelling were based on a multi-country one-year (2014) approach because we wanted to account for the differences among the countries. We examine the relationship between different important independent variables and its influence on CO2 emissions. It may generate bias in our model prediction if we use data from only one country over several years.

The results from the data exploration indicated that logarithm transformation would be more suitable for our modeling part. The report narrowed down to 14 indicators that have high potential of affecting the CO2 emission per capita, and a correlation analysis was carried out on all these 14 indicators. The results from the correlation analysis helped us to choose 6 indicators that have high correlation with CO2 emission per capita. Out of 6 indicators, 5 have positive correlation while

SL.AGR.EMPL.ZS (% employment in agriculture) has negative correlation with CO2 emissions per capita. Furthermore, the report divided these indicators into two groups. One is Value added per worker (productivity) and the other is Percentage of employment in each sector.

The report plotted graphs based on 2 groups of indicators which are productivity and percentage of employment in different sectors. One interesting trend noted is that during 2008 around the financial crisis which caused global disturbances, there was a slight drop in productivity for the sectors Industry and Service. It is also noted that the CO2 emission/capita also dropped during that time. However, the productivity for agriculture has a continuous rise over the years. The other graphs indicated that the region of North America had the largest CO2/emission per capita, and their productivity per capita was the highest for all sectors. In contrast, the African and South Asian countries have lower emissions and productivity in different sectors, which also correlates to the lower economic development in these countries.

In terms of employment distribution, North America has the highest percentage of employment in service and the lowest in agriculture compared to other regions. As a result, they also have the highest CO2 per capita. The opposite pattern was observed for South Asia and Sub-Saharan Africa. We can say that when a country transitions from an agriculture-based to a service-based economy, their CO2 emission per capita will generally become higher.

The modelling part of this report focused on developing the causal diagram to model the relationship between economic factors and CO2 emission per capita. CO2 emission per capita was the dependent variable, while the other 6 indicators were independent variables in this model. It was assumed that our dependent variable follows a normal distribution. The mean value of the distribution has linear relationships with regards to 6 independent variables (beta0 is the intercept while beta1 to beta 6 is the corresponding slope of 6 independent variables).

Due to multicollinearity, the independent variables with the highest VIF were removed, and our final model could predict CO2 per capita based on 2 independent variables namely productivity in industry and % employment in Industry with an R square of nearly 0.6. The coefficients obtained by our Bayesian model was cross-checked against the coefficients from a simple linear model. Finally, posteriors predicted by our model was compared against actual values (from WDI data) to see whether there is any significant bias. Possible future work will be to explore more economic indicators and fit more variables to the model.

References

Che-project.eu. (2019). *Main sources of carbon dioxide emissions | CO2 Human Emissions*. [online] Available at: https://www.che-project.eu/news/main-sources-carbon-dioxide-emissions [Accessed 6 Nov. 2019].

Climatecentral.org. (2019). *Greenhouse Gas Concentrations*. [online] Available at: https://www.climatecentral.org/gallery/graphics/greenhouse-gas-concentrations [Accessed 6 Nov. 2019].

Datacatalog.worldbank.org. (2019). *World Development Indicators (WDI) | Data Catalog*. [online] Available at: https://datacatalog.worldbank.org/dataset/world-development-indicators [Accessed 6 Nov. 2019].

Ghosh, I. (2019). *All the World's Carbon Emissions in One Chart*. [online] Visual Capitalist. Available at: https://www.visualcapitalist.com/all-the-worlds-carbon-emissions-in-one-chart/ [Accessed 7 Nov. 2019].

Jain, M. (2018). Impact of Increasing CO2 Emissions on Environment. Retrieved from Science ABC: https://www.scienceabc.com/social-science/greenhouse-gas-co2-emmission-effect-environment.html[Accessed 6 Nov. 2019].

Levin, K. (2019). *New Global CO2 Emissions Numbers Are In. They're Not Good.*. [online] World Resources Institute. Available at: https://www.wri.org/blog/2018/12/new-global-co2-emissions-numbers-are-they-re-not-good [Accessed 6 Nov. 2019].

Lux, H. (2019). The UN Paris Agreement was 3 years ago; here's how countries are doing with their climate goals. Retrieved from Good: https://www.good.is/paris-agreement-progress[Accessed 6 Nov. 2019].

Ritchie, H. (2019). *Where in the world do people emit the most CO2?*. [online] Our World in Data. Available at: https://ourworldindata.org/per-capita-co2 [Accessed 6 Nov. 2019].

Royalsociety.org. (2019). *Ocean acidification due to increasing atmospheric carbon dioxide | Royal Society*. [online] Available at: https://royalsociety.org/topics-policy/publications/2005/ocean-acidification/ [Accessed 6 Nov. 2019].

Smith, H. J.(1999). *Dual modes of the carbon cycle since the Last Glacial Maximum*. Available from: https://scholar.google.com/scholar?hl=nl&as_sdt=0%2C5&q=H.+J.+Smith%2C+H.+Fischer%2C+M.+Wahlen%2C+D.+Mastroianni%2C+B.+Deck%2C+Nature+400%2C+248+%281999%29.&btnG= [Accessed 8 Nov. 2019]

Appendix

This section includes all R programming code of this project.

7.1 Data preparation

```
# Set wd
setwd("D:/Year 2 Quarter 1/EPA1315 Data Analytics/Final Project/Modelling")

# Read the data in correctly
mydata_utf8 = read.table("WDIData.csv",sep=",",fileEncoding="UTF-8-BOM",header=TRUE)
country_code = read.table("country_code.csv",sep=",",header=TRUE)

# drop off the final blank column
mydata = mydata_utf8[,1:63]

# find the right correlated variables
my_var = c("EN.ATM.CO2E.PC","BX.KLT.DINV.CD.WD","BX.TRF.PWKR.DT.GD.ZS",
         "DT.ODA.ODAT.CD","NE.EXP.GNFS.ZS","NV.AGR.EMPL.KD",

"NV.IND.EMPL.KD","NV.IND.MANF.CD","NV.MNF.TECH.ZS.UN","NV.SRV.EMPL.KD",
         "NY.ADJ.SVNX.GN.ZS","SL.AGR.EMPL.ZS","SL.IND.EMPL.ZS",
         "SL.ISV.IFRM.ZS","SL.SRV.EMPL.ZS")

# subset the data
country_data = subset(mydata,mydata$Country.Code %in% country_code$Country.Code)
main_data = subset(country_data,country_data$Indicator.Code %in% my_var_corr)
# find the columns with the year we want
main_idx <- match(c("Country.Code","Indicator.Code","X2014"), names(mydata))
# peel off only the column we want from 2014
main_data <- main_data[,main_idx]

# Name the columns
names(main_data) <- c("Country","Indicator","Value")

#install.packages('reshape')
library('reshape')

# Reshaping
combined_melt = melt(main_data, id=c("Country","Indicator","Value"))
combined_cast = cast(combined_melt, value = "Value", Country  ~ Indicator)
```

```
final_data <- combined_cast

# remove NA data:
final_data = na.omit(final_data)
```

7.2 Check normality of CO2 emission per capita

```
# To plot Q-Q plot and Density plot
library(ggpubr)
ggqqplot(log(combined_cast$EN.ATM.CO2E.PC), main="Q-Q plot of log(CO2/capita)", color
= "royal blue")
ggdensity(log(combined_cast$EN.ATM.CO2E.PC),
          main = "Density plot of log(CO2/capita)",
          xlab = "log(CO2/capita)")

# To test whether the CO2 data follows normal distribution:
shapiro.test(log(combined_cast$EN.ATM.CO2E.PC))
```

7.3 Correlation Analysis Table

```
# Create a dataframe for the result of the analysis:
correlation_table=data.frame(matrix(ncol=3,nrow=15))
names(correlation_table) <- c("Indicator","p_value","Cor_coefficient")
correlation_table$Indicator <- my_var

# Run the correlation analysis:
i=1
for (indicator in my_var) {
  result <- cor.test(final_data1 %>% pull(indicator), final_data1 %>% pull(EN.ATM.CO2E.PC),
                     use="pairwise.complete.obs",method="kendall")
  correlation_table[i,2] <-  result$p.value
  correlation_table[i,3] <-  result$estimate
  i=i+1
}
```

7.4 Correlation Analysis Graphs

```
final_data$y <- log(final_data$EN.ATM.CO2E.PC)
final_data$x1 <- log(final_data$NV.AGR.EMPL.KD)
final_data$x2 <- log(final_data$NV.IND.EMPL.KD)
final_data$x3 <- log(final_data$NV.SRV.EMPL.KD)
```

```
final_data$x4 <- log(final_data$SL.AGR.EMPL.ZS)
final_data$x5 <- log(final_data$SL.IND.EMPL.ZS)
final_data$x6 <- log(final_data$SL.SRV.EMPL.ZS)

#install.packages("ggpubr")
library("ggpubr")

#install.packages('tidyverse')
library('tidyverse')

# Try plotting an indicator against CO2/capita:
ggscatter(final_data, x = "x2", y = "y",
        add = "reg.line", conf.int = TRUE,
        cor.coef = TRUE, cor.method = "kendall", color = "sky blue",
        cor.coef.coord = c(7,3), xlim=c(6.6,12.7),
        xlab = "Productivity in Industry (log scale)", ylab="CO2/capita (log scale)",
        main = "                 (Productivity in Industry) vs (CO2/capita)")
```

7.5 Productivity vs CO2 per capita

```
# find the right variables coming from correlation analysis
my_var_corr =
c("NV.SRV.EMPL.KD","SL.AGR.EMPL.ZS","SL.ISV.IFRM.ZS","NV.AGR.EMPL.KD","NV.
IND.EMPL.KD",
           "SL.SRV.EMPL.ZS","SL.IND.EMPL.ZS","EN.ATM.CO2E.PC")

# subset the data
country_data = subset(mydata,mydata$Country.Code %in% country_code$Country.Code)
main_data = subset(country_data,country_data$Indicator.Code %in% my_var_corr)
# find the columns with the year we want
main_idx <- match(c("Country.Code","Indicator.Code","X2014"), names(mydata))
# peel off only the column we want from 2014
main_data <- main_data[,main_idx]

# Name the columns
names(main_data) <- c("Country","Indicator","Value")

#install.packages('reshape')
library('reshape')

# Reshaping
```

```
combined_melt = melt(main_data, id=c("Country","Indicator","Value"))
combined_cast = cast(combined_melt, value = "Value", Country  ~ Indicator)
final_data <- combined_cast

# Adding regions to the dataframe
country_info = read.table("WDICountry.csv",sep=",",header=TRUE)
country_code_region = country_info[,c(1,8)]
colnames(country_code_region)=c("Country","Region")
country_core_final = subset(country_code_region,country_code_region$Country %in%
country_code$Country.Code)
region_indicat_data = merge(country_core_final,final_data,id = "Country")

#Calculate mean productivity and CO2/capita
productivity <- aggregate(region_indicat_data[, 3:6], list(region_indicat_data$Region), mean,
                          na.rm=TRUE, na.action=NULL)

#normalizing (ratio)
col7=c()
for (i in 1:7){
  norm_columnval= productivity$EN.ATM.CO2E.PC[i] /
mean(productivity$EN.ATM.CO2E.PC) * 100
  col7=c(col7,norm_columnval)
}

col8=c()
for (i in 1:7){
  norm_columnval= productivity$NV.AGR.EMPL.KD[i] /
mean(productivity$NV.AGR.EMPL.KD) * 100
  col8=c(col8,norm_columnval)
}

col9=c()
for (i in 1:7){
  norm_columnval= productivity$NV.IND.EMPL.KD[i] /
mean(productivity$NV.IND.EMPL.KD) * 100
  col9=c(col9,norm_columnval)
}

col10=c()
for (i in 1:7){
  norm_columnval= productivity$NV.SRV.EMPL.KD[i] /
mean(productivity$NV.SRV.EMPL.KD) * 100
  col10=c(col10,norm_columnval)
}

test=rbind(col8,col9,col10,col7)

#Plot a grouped bar chart
par(mar=c(12,5,4.1,2.1)) # Increase the size of the margin
barplot(test,beside=T,main="Productivity and CO2 emission per capita by regions (2014)",
        ylab="Normalized productivity and CO2/capita", ylim=c(0,400),
        col=c("royalblue","tomato","limegreen","darkgrey"),
        legend = c("Agriculture","Industry","Service","Carbon emission per capita"),
        args.legend = list(x = "topleft", cex = 0.6),
        names.arg = productivity$Group.1, las = 3)

# WLD -> Timeseries "EN.ATM.CO2E.PC" and value added per worker in SER, AGR, IND

# find rownumbers for World:
"NV.SRV.EMPL.KD","NV.AGR.EMPL.KD","NV.IND.EMPL.KD" and "EN.ATM.CO2E.PC"
a=c(which(mydata$Indicator.Code == "NV.SRV.EMPL.KD" &
mydata$Country.Code=="WLD"))
b=c(which(mydata$Indicator.Code == "NV.AGR.EMPL.KD" &
mydata$Country.Code=="WLD"))
c=c(which(mydata$Indicator.Code == "NV.IND.EMPL.KD" &
mydata$Country.Code=="WLD"))
d=c(which(mydata$Indicator.Code == "EN.ATM.CO2E.PC" &
mydata$Country.Code=="WLD"))

# Making dataframe whith these variables
Df_timeseries=mydata[c(a,b,c,d),]

# Peel off years 1961-1990 and 2015-2018 (since no data exists of value added per worker and
CO2/capita)
index_start = match(c("X1991"),names(mydata))
index_end = match(c("X2014"),names(mydata))
Df_timeseries=Df_timeseries[,c(1:4,index_start:index_end)]

# Normalizing values in dataframe
col1=c()   # "NV.SRV.EMPL.KD"
```

```
for (i in 5:ncol(Df_timeseries)){
  norm_columnval= (Df_timeseries[1,i]-min(Df_timeseries[1,5:ncol(Df_timeseries)])) /
(max(Df_timeseries[1,5:ncol(Df_timeseries)])-min(Df_timeseries[1,5:ncol(Df_timeseries)])) *
100
  col1=c(col1,norm_columnval)
}

col2=c()   # "NV.AGR.EMPL.KD"
for (i in 5:ncol(Df_timeseries)){
  norm_columnval= (Df_timeseries[2,i]- min(Df_timeseries[2,5:ncol(Df_timeseries)])) /
(max(Df_timeseries[2,5:ncol(Df_timeseries)])-min(Df_timeseries[2,5:ncol(Df_timeseries)])) *
100
  col2=c(col2,norm_columnval)
}

col3=c()   # "NV.IND.EMPL.KD"
for (i in 5:ncol(Df_timeseries)){
  norm_columnval= (Df_timeseries[3,i]-min(Df_timeseries[3,5:ncol(Df_timeseries)])) /
(max(Df_timeseries[3,5:ncol(Df_timeseries)])-min(Df_timeseries[3,5:ncol(Df_timeseries)])) *
100
  col3=c(col3,norm_columnval)
}

col4=c()   # "EN.ATM.CO2E.PC"
for (i in 5:ncol(Df_timeseries)){
  norm_columnval= (Df_timeseries[4,i] - min(Df_timeseries[4,5:ncol(Df_timeseries)]))/
(max(Df_timeseries[4,5:ncol(Df_timeseries)])-min(Df_timeseries[4,5:ncol(Df_timeseries)]))*
100
  col4=c(col4,norm_columnval)
}

Years=c()
for (i in 1:24){
  Years=c(Years,1990+i)
}

plot_data=data.frame(Years)
plot_data["NV.SRV.EMPL.KD"]=col1
plot_data["NV.AGR.EMPL.KD"]=col2
plot_data["NV.IND.EMPL.KD"]=col3
```

```
plot_data["EN.ATM.CO2E.PC"]=col4

# making line-plot
x=c(1991:2014)
q=plot_data$NV.SRV.EMPL.KD
r=plot_data$NV.AGR.EMPL.KD
s=plot_data$NV.IND.EMPL.KD
t=plot_data$EN.ATM.CO2E.PC

# Plot
my_xlab = "Years"
my_ylab = "normalized values"
plot(x,q,xlab=my_xlab, ylab=my_ylab, main="Productivity versus world CO2/capita (1991-
2014)",type="l",col="limegreen",ylim=c(0,100))
lines(x,r,col='royalblue')
lines(x,s,col='tomato')
lines(x,t,col='darkgrey')
legend("bottomright",c("Service","Agriculture","Industry","CO2 emission/capita"),
       fill=c("limegreen","royalblue","tomato","darkgrey"),cex=0.65)
```

7.6 Employment percentage vs CO2 emission

```
# WLD -> Timeseries "EN.ATM.CO2E.PC" and % of employment in SER, AGR, IND

# find rownumbers for World:
a=c(which(mydata$Indicator.Code == "SL.AGR.EMPL.ZS" &
mydata$Country.Code=="WLD"))
b=c(which(mydata$Indicator.Code == "SL.IND.EMPL.ZS" &
mydata$Country.Code=="WLD"))
c=c(which(mydata$Indicator.Code == "SL.SRV.EMPL.ZS" &
mydata$Country.Code=="WLD"))
d=c(which(mydata$Indicator.Code == "EN.ATM.CO2E.PC" &
mydata$Country.Code=="WLD"))

# Making dataframe whith these variables
Df_timeseries=mydata[c(a,b,c,d),]

# Peel off years 1961-1990 and 2015-2018 (since no data exists of value added per worker and
CO2/capita)
index_start = match(c("X1991"),names(mydata))
```

```
index_end = match(c("X2014"),names(mydata))
Df_timeseries=Df_timeseries[,c(1:4,index_start:index_end)]
Df_timeseries2 <- as.data.frame(t(Df_timeseries[,-c(1:4)]))
names(Df_timeseries2)[1] <- "Agriculture"
names(Df_timeseries2)[2] <- "Industry"
names(Df_timeseries2)[3] <- "Service"
names(Df_timeseries2)[4] <- "CO2_per_capita"

time_df = data.frame(year = rep(1991:2014, 3),
                CO2_per_capita = rep(Df_timeseries2$CO2_per_capita, 3),
                Percentage_of_total_employment = c(Df_timeseries2$Agriculture,
                                           Df_timeseries2$Industry,
                                           Df_timeseries2$Service),
                Sector = c(rep("Agriculture", 24), rep("Industry", 24),rep('Service',24)))



library(ggplot2)
ggplot() +
  geom_area(data = time_df, aes(y = Percentage_of_total_employment, x = year, fill = Sector)) +
  geom_line(data = time_df[1:24, ], aes(y = CO2_per_capita*250/3-325, x = year,linetype =
"CO2_per_capita")) +
  scale_fill_manual(values = c("royalblue", "tomato","limegreen"))  +
  scale_y_continuous(sec.axis = sec_axis(name =  "CO2/capita (metric ton)", ~.*0.012+3.9)) +
  ggtitle("Employment distribution and CO2/capita over 1990-2014 (World average)") +
  theme(plot.title = element_text(hjust = 0.5))

# Code to plot the Employment by region
employment <- aggregate(region_indicat_data[, 6:9], list(region_indicat_data$Region), mean,
                    na.rm=TRUE, na.action=NULL)
test <- rbind(employment$SL.AGR.EMPL.ZS
,employment$SL.IND.EMPL.ZS,employment$SL.SRV.EMPL.ZS)
par(mar=c(12,5,4.1,2.1)) # Increase the size of the margin
barplot(test,beside=T,main="employment by region",
      ylab="% employment, sector wise", ylim=c(0,70),
      col=c("lightblue","tomato","palegreen"),
      legend = c("Agriculture","Industry","Service"),
      args.legend = list(x = "topleft", cex = 0.4),
      names.arg = employment$Group.1, las = 3)
```

7.7 Model #1 and Multicollinearity

```
# 1. MODEL
library(rjags)

y <- log(final_data$EN.ATM.CO2E.PC)
x1 <- log(final_data$NV.AGR.EMPL.KD)
x2 <- log(final_data$NV.IND.EMPL.KD)
x3 <- log(final_data$NV.SRV.EMPL.KD)
x4 <- log(final_data$SL.AGR.EMPL.ZS)
x5 <- log(final_data$SL.IND.EMPL.ZS)
x6 <- log(final_data$SL.SRV.EMPL.ZS)
n <- nrow(final_data)

model_string1 <- "model{

# Priors
beta0 ~ dnorm(0,0.0001)
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
beta3 ~ dnorm(0,0.0001)
beta4 ~ dnorm(0,0.0001)
beta5 ~ dnorm(0,0.0001)
beta6 ~ dnorm(0,0.0001)

inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)

# Likelihood
for(i in 1:n){
  mu[i] <- beta0 + beta1*x1[i] + beta2*x2[i]+beta3*x3[i]+beta4*x4[i]+beta5*x5[i]+beta6*x6[i]
  y[i]   ~ dnorm(mu[i],inv_var)
}
}"

model <- jags.model(textConnection(model_string1), n.chains=2,
                    data = list(x1=x1,x2=x2,x3=x3,x4=x4,x5=x5,x6=x6,y=y,n=n))
update(model, 10000, progress.bar="none"); # Burnin for 100000 samples
samp <-
coda.samples(model,variable.names=c("beta0","beta1","beta2","beta3","beta4","beta5","beta6","
std"),
                n.iter=20000, thin=1, progress.bar="text")

model_output<-as.matrix(samp)
saveRDS(samp,"modelrunfinal.RDS")
write.csv(model_output,"modelrunsfinal.csv",row.names = FALSE)

summary(samp)
effectiveSize(samp)
acfplot(samp)

# 2. HISTOGRAM PARAMETERS
library(scales)
library(lattice)
# load in the coda matrix of model runs
# this consists of slope, intercept, and noise
model_runs = read.table("modelrunsfinal.csv", sep = ",", header = TRUE)

# load in the data
emission <- log(final_data$EN.ATM.CO2E.PC)

# compute the variance
sd<-sd(emission)
pseudo_r2 <- 1-model_runs[,4]/sd
# set however many sample runs you want to display
n <- 500

model_runs <- model_runs[1:n,1:8]
hist(model_runs[,1],col=alpha("blue",0.5),breaks=20,main="Intercept",xlab="Value",ylab="Runs")
hist(model_runs[,2],col=alpha("blue",0.5),breaks=20,main="Slope1",xlab="Value",ylab="Runs")
hist(model_runs[,3],col=alpha("blue",0.5),breaks=20,main="Slope2",xlab="Value",ylab="Runs")
hist(model_runs[,4],col=alpha("blue",0.5),breaks=20,main="Slope3",xlab="Value",ylab="Runs")
hist(model_runs[,5],col=alpha("blue",0.5),breaks=10,main="Slope4",xlab="Value",ylab="Runs")
hist(model_runs[,6],col=alpha("blue",0.5),breaks=10,main="Slope5",xlab="Value",ylab="Runs")
hist(model_runs[,7],col=alpha("blue",0.5),breaks=10,main="Slope6",xlab="Value",ylab="Runs"
)
hist(model_runs[,8],col=alpha("blue",0.5),breaks=16,main="Error",xlab="Value",ylab="Runs")
hist(pseudo_r2,col=alpha("blue",0.5),breaks=16,main="Pseudo R
Square",xlab="Value",ylab="Runs")
splom(model_runs)
mean(pseudo_r2)

# 3. CHECK CONVERGENCE:
samp=readRDS("modelrunfinal.rds")

summary(samp)

# sometimes the gelman plot won't fit on a screen
# we have to reduce the margins
par(mar=c(3,3,1,1))
gelman.plot(samp)
gelman.diag(samp)
plot(samp, trace=FALSE, density = TRUE)
plot(samp, trace=TRUE, density = FALSE)
acfplot(samp)

# get the effective sample size
effectiveSize(samp)

# To plot correlation among the variables
library(corrgram)
final_data1 <- final_data[c(1,9:15)]
corrgram(final_data1, order=TRUE, lower.panel=panel.shade,
        upper.panel=panel.pie, text.panel=panel.txt,
        main="Correlation between economic factors and CO2/capita")

# Check VIF
Mod <- lm(log(EN.ATM.CO2E.PC) ~ log(NV.IND.EMPL.KD)
             +log(SL.IND.EMPL.ZS), data=final_data)

#install.packages('faraway')
library(faraway)
vif(Mod)
```

7.8 Model #2

```
# 1. MODEL

library(rjags)

y <- log(final_data$EN.ATM.CO2E.PC)
x2 <- log(final_data$NV.IND.EMPL.KD)
x4 <- log(final_data$SL.AGR.EMPL.ZS)
x5 <- log(final_data$SL.IND.EMPL.ZS)
x6 <- log(final_data$SL.SRV.EMPL.ZS)
n <- nrow(final_data)


model_string1 <- "model{

# Priors
beta0 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)

beta4 ~ dnorm(0,0.0001)
beta5 ~ dnorm(0,0.0001)
beta6 ~ dnorm(0,0.0001)
inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)

# Likelihood
for(i in 1:n){
mu[i] <- beta0 + beta2*x2[i]+beta4*x4[i]+beta5*x5[i]+beta6*x6[i]
y[i]   ~ dnorm(mu[i],inv_var)
}
}"

model <- jags.model(textConnection(model_string1), n.chains=2,
                    data = list(x2=x2,x4=x4,x5=x5,x6=x6,y=y,n=n))
update(model, 100000, progress.bar="none"); # Burnin for 100000 samples
samp <- coda.samples(model,variable.names=c("beta0","beta2","beta4","beta5","beta6","std"),
                     n.iter=200000, thin=10, progress.bar="text")

model_output<-as.matrix(samp)
saveRDS(samp,"modelrunfinal.RDS")
```

```
write.csv(model_output,"modelrunsfinal.csv",row.names = FALSE)

summary(samp)
effectiveSize(samp)
acfplot(samp)


# 2. CHECK CONVERGENCE:
samp=readRDS("modelrunfinal.rds")

summary(samp)

# sometimes the gelman plot won't fit on a screen
# we have to reduce the margins
par(mar=c(3,3,1,1))
gelman.plot(samp)
gelman.diag(samp)
plot(samp, trace=FALSE, density = TRUE)
plot(samp, trace=TRUE, density = FALSE)
acfplot(samp)

# get the effective sample size
effectiveSize(samp)
```

7.9 Model #3

```
# 1. MODEL
library(rjags)

y <- log(final_data$EN.ATM.CO2E.PC)
x2 <- log(final_data$NV.IND.EMPL.KD)
x5 <- log(final_data$SL.IND.EMPL.ZS)
n <- nrow(final_data)

model_string1 <- "model{

# Priors
beta0 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
beta5 ~ dnorm(0,0.0001)
inv_var ~ dgamma(0.1,0.1)
std <- 1/sqrt(inv_var)

# Likelihood
for(i in 1:n){
mu[i] <- beta0 + beta2*x2[i]+beta5*x5[i]
y[i]   ~ dnorm(mu[i],inv_var)
}
}"

model <- jags.model(textConnection(model_string1), n.chains=2,
                    data = list(x2=x2,x5=x5,y=y,n=n))
update(model, 100000, progress.bar="none"); # Burnin for 100000 samples
samp <- coda.samples(model,variable.names=c("beta0","beta2","beta5","std"),
                     n.iter=200000, thin=10, progress.bar="text")

model_output<-as.matrix(samp)
saveRDS(samp,"modelrunfinal.RDS")
write.csv(model_output,"modelrunsfinal.csv",row.names = FALSE)

summary(samp)
effectiveSize(samp)
acfplot(samp)

# 2. HISTOGRAM PARAMETERS
library(scales)
library(lattice)
# load in the coda matrix of model runs
# this consists of slope, intercept, and noise
model_runs = read.table("modelrunsfinal.csv", sep = ",", header = TRUE)
# load in the data
emission <- log(final_data$EN.ATM.CO2E.PC)

# compute the variance
sd<-sd(emission)
pseudo_r2 <- 1-model_runs[,4]/sd
# set however many sample runs you want to display
n <- 500

model_runs <- model_runs[1:n,1:4]
```

```
hist(model_runs[,1],col=alpha("blue",0.5),breaks=20,main="Intercept",xlab="Value",ylab="Run
s")
hist(model_runs[,2],col=alpha("blue",0.5),breaks=20,main="Slope2",xlab="Value",ylab="Runs"
)
hist(model_runs[,3],col=alpha("blue",0.5),breaks=20,main="Slope5",xlab="Value",ylab="Runs"
)
hist(model_runs[,4],col=alpha("blue",0.5),breaks=20,main="Error",xlab="Value",ylab="Runs")
hist(pseudo_r2,col=alpha("blue",0.5),breaks=16,main="Pseudo R
Square",xlab="Value",ylab="Runs")
splom(model_runs)
mean(pseudo_r2)


# 3. POSTERIORS:
#the data consists of 163 countries
code <- final_data$Country
n <- length(code)
intercept <- rep(1,n)

# 80000 runs by 3 parameters
model_runsfinal = read.table("modelrunsfinal.csv", sep = ",", header = TRUE)
model_runsfinal <- as.matrix(model_runsfinal[1:7])

# 3 parameters by 163 nations
design_matrix <- t(cbind(intercept,x2,x5))

p <- model_runsfinal %*% design_matrix
posteriorsfinal <- as.data.frame(p)
names(posteriorsfinal) <- code

write.csv(posteriorsfinal,"posteriorsfinal.csv",row.names = FALSE)

# 4. CHECK CONVERGENCE:
samp=readRDS("modelrunfinal.rds")
summary(samp)

# sometimes the gelman plot won't fit on a screen
# we have to reduce the margins
par(mar=c(3,3,1,1))
gelman.plot(samp)
gelman.diag(samp)
plot(samp, trace=FALSE, density = TRUE)
plot(samp, trace=TRUE, density = FALSE)
acfplot(samp)

# get the effective sample size
effectiveSize(samp)

# 5. VALIDATE
sampmatrix = as.matrix(samp)
# take the first seven parameters, but drop off the noise!
param <-sampmatrix[sample(nrow(sampmatrix),size = 50, replace=FALSE),1:3]
# create a design matrix
n <- nrow(final_data)
# make an intercept
intercept <- rep(1,n)
# bind your data together in the order used for modelling
data <- cbind(intercept,x2,x5)
pred <- param %*% t(data)
pred <- t(pred)
# replicate the results
act <- rep(y,50)
x11(width=10, height=10, pointsize=8)
par(mar=c(5,5,1,1))
plot(act,pred,xlab="Actual",ylab="Predicted",xlim=c(-5,4),ylim=c(-5,4))
abline(0,1,col=alpha("blue",1))
```