Who is Responsible for Global Warming?

Owen Xu Li

2024-04-18

Contents

1 Introduction

In recent decades, Global Warming has been a key issue for scientists and researchers to understand and find solutions to minimize its effects. However, it has become hard to find the source of the problem, since developed economies are controlling the extraction and processing of natural resources in developing countries. Therefore, the blame often goes to developing countries, and developed countries continue to exploit and mismanage natural resources, leading to record-high CO_2 emissions. In this report, I intend to find the correlation between the development of a country and its emissions by year, using GDP per capita in USD as a measure of wealth, to try to capture which economies hold the most responsibility over the stark increase in average global temperatures as a result of higher greenhouse gas emissions.

Is there enough evidence to conclude developed countries contribute more to Global Warming than developing countries?

I will utilize the following data sets (hyperlinked)

1. CO2 Emissions in Kilotons by Country
2. World GDP and GDP per capita per country

These data sets come from Kaggle. The first data set includes the CO_2 (Carbon Dioxide) emissions per country in kilotons from 1960 to 2019. The second data set is an overview of the world's nations' GDP and GDP per capita; since GDP per capita is often used as a measure of a country's wealth, I intend to use it to differentiate less developed economies from developed economies. Since we know wealthier countries are more consuming-oriented economies, I predict that developed economies will have higher CO_2 emissions. One of the assumptions is that developed economies have a GDP per capita higher than 15,000 USD, according to Investopedia.

```
emissions <- read.csv('data/emissions.csv')

gdp <- read.csv('data/gdp.csv')
```

The results of this short study show that there is convincing evidence that developed and developing countries have different mean CO_2 emissions in Kilotons. Additionally, the linear regression shows that the total GDP of a country is a good predictor of a country's CO_2 emissions, where every increase in GDP will increase CO_2 emissions.

2 Data Visualization

In this case, I will focus up until the most recent year in which the CO_2 emissions by country were recorded - 2019. Note that there are 239 unique observations for 'country_name' in 2019, but there are only 195 countries in the world. This is because some regions of the world, generalized economic regions, and territories of other nations have been inputted as countries in the CO_2 emissions and GDP data sets. Each country will have multiple entries, each entry corresponding to a unique year, and I will proceed to graph all the entries in one graph for easier visualization and comparison.

What kind of data are the data sets?

```
str(emissions)
```

```
'data.frame':   13953 obs. of  4 variables:
 $ country_code: chr  "ABW" "ABW" "ABW" "ABW" ...
 $ country_name: chr  "Aruba" "Aruba" "Aruba" "Aruba" ...
 $ year        : int  1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 ...
 $ value       : num  11093 11577 12713 12178 11841 ...
```

```
heademissions <- head(emissions, 10)

knitr::kable(heademissions,
             caption = 'CO2 emissions in kilotons per country from 1960 to 2019',
             col.names = c('Country Code',
                           'Country Name',
                           'Year',
                           'Emissions in kt'),
             digits = 3)
```

Table 1: CO2 emissions in kilotons per country from 1960 to 2019

Country Code	Country Name	Year	Emissions in kt
ABW	Aruba	1960	11092.675
ABW	Aruba	1961	11576.719
ABW	Aruba	1962	12713.489
ABW	Aruba	1963	12178.107
ABW	Aruba	1964	11840.743
ABW	Aruba	1965	10623.299
ABW	Aruba	1966	9933.903
ABW	Aruba	1967	12236.779
ABW	Aruba	1968	11378.701
ABW	Aruba	1969	14891.687

```
headgdp <- head(gdp, 10)

knitr::kable(headgdp,
             caption = 'GDP and GDP per capita per country from 1960 to 2021',
             col.names = c('Country Name', 'Country Code',
                           'Year',
                           'GDP in US Dollars',
                           'GDP per Capita in US Dollars'),
             digits = 3)
```

Table 2: GDP and GDP per capita per country from 1960 to 2021

Country Name	Country Code	Year	GDP in US Dollars	GDP per Capita in US Dollars
Aruba	ABW	1960	NA	NA
Africa Eastern and Southern	AFE	1960	21290586003	162.726
Afghanistan	AFG	1960	537777811	59.773
Africa Western and Central	AFW	1960	10404135069	107.931
Angola	AGO	1960	NA	NA
Albania	ALB	1960	NA	NA
Andorra	AND	1960	NA	NA
Arab World	ARB	1960	NA	NA
United Arab Emirates	ARE	1960	NA	NA
Argentina	ARG	1960	NA	NA

We see that the CO_2 emissions data set contain 13953 observations of 4 variables, namely the country code and name, the year of the recorded observation, and the value of the observation. Similarly, the GDP data set is composed of 16492 observations of 5 variables, namely the country code and name, the year of the recorded observation, and the GDP and GDP per capita of each country in US dollars in the year of the observation. In my case, I will categorize the countries into developed and developing economies based on whether their GDP per capita exceeds 15,000 USD.

Since the emissions data is continuous and there is high variation between countries, it's likely a good idea to log transform the data to reduce the variance and make the data more interpretable. I will also create histograms of the data to see the distribution of the data.

We first combine the two data sets into one that matches GDP, GDP per capita, and CO_2 emissions by country per year.

```
x <- gdp %>%
  rename(country_name = Country.Name, country_code = Country.Code)

y <- emissions

# to merge the data frames by country code, name, and year:
complete <- merge(x, y, by = intersect(names(x), names(y))) %>%
  mutate(status = if_else(GDP_per_capita_USD > 15000, 'Developed', 'Developing'),
         lgemissions = log(value),
         lggdp = log(GDP_USD)) %>%
  filter(!is.na(GDP_per_capita_USD),
         !is.na(GDP_USD),
         !is.na(value),
         is.finite(lgemissions),
         is.finite(lggdp)) %>%
  select(country_code,
         country_name,
         year,
         lggdp,
         GDP_per_capita_USD,
         lgemissions,
         status)

# to add the region to which each country belongs to:
complete <- complete %>%
```

```
  mutate(wregion = countrycode(country_name,
                               origin = 'country.name',
                               destination = 'region')) %>%
  filter(!is.na(wregion)) %>%
  select(country_name,
         country_code,
         wregion,
         year,
         lggdp,
         GDP_per_capita_USD,
         lgemissions,
         status)

knitr::kable(head(complete, 10),
             caption = 'GDP and CO2 Emissions by Country',
             col.names = c('Country',
                           'Code',
                           'Region',
                           'Year',
                           'Log GDP (USD)',
                           'GDPpc (USD)',
                           'Log CO2 Emissions (kt)',
                           'Development Status'),
             digits = 3)
```

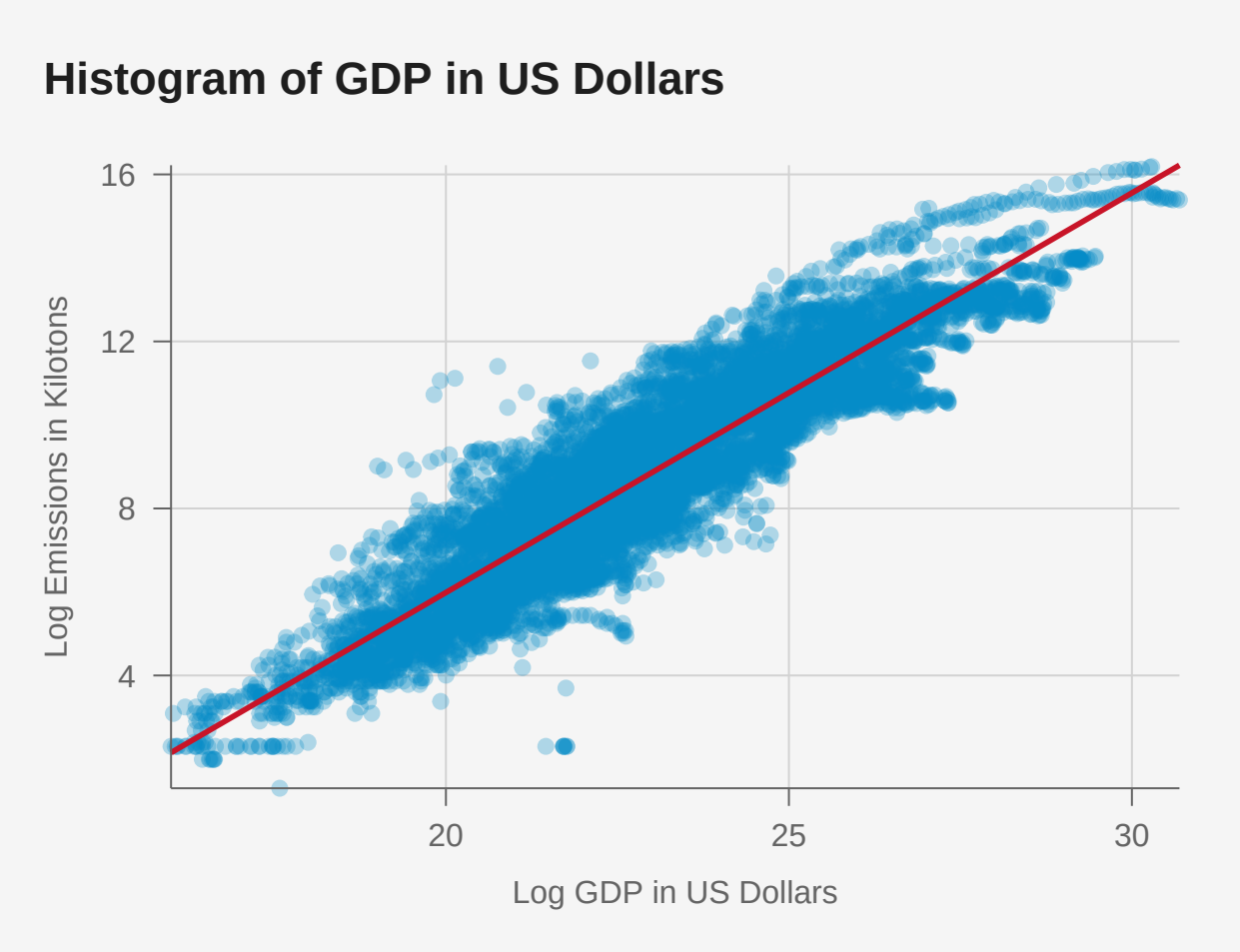
Table 3: GDP and CO2 Emissions by Country

Country	Code	Region	Year	Log GDP (USD)	GDPpc (USD)	Log CO2 Emissions (kt)	Development Status
Afghanistan	AFG	South Asia	1960	20.103	59.773	6.027	Developing
Afghanistan	AFG	South Asia	1961	20.123	59.861	6.197	Developing
Afghanistan	AFG	South Asia	1962	20.119	58.458	6.536	Developing
Afghanistan	AFG	South Asia	1963	20.437	78.706	6.562	Developing
Afghanistan	AFG	South Asia	1964	20.500	82.095	6.733	Developing
Afghanistan	AFG	South Asia	1965	20.730	101.108	6.916	Developing
Afghanistan	AFG	South Asia	1966	21.060	137.594	6.996	Developing
Afghanistan	AFG	South Asia	1967	21.238	160.898	7.157	Developing
Afghanistan	AFG	South Asia	1968	21.041	129.108	7.111	Developing
Afghanistan	AFG	South Asia	1969	21.066	129.330	6.848	Developing

Now we check the assumptions for a linear regression model: 1. The dependent and independent variable are linearly related. 2. The residuals are normally distributed, or the dependent variable is normally distributed

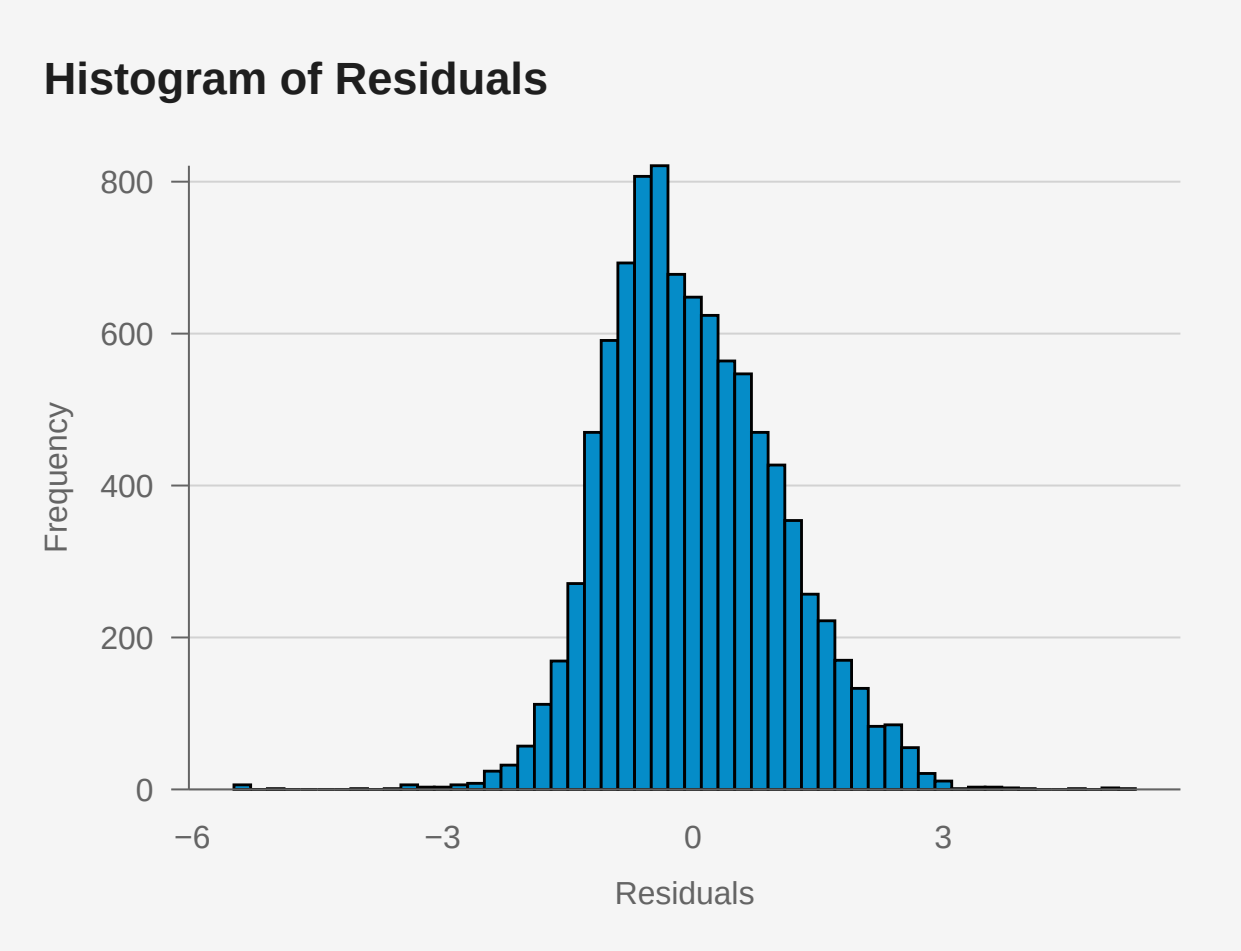
for a given value of the independent variable. 3. The variance of the residuals is constant for all values of the independent variable.

```
complete <- complete %>% mutate(bins = cut(lggdp, breaks = 10))

# write.csv(complete, 'data/complete.csv')

completegg <- ggplot(complete, aes(x = lggdp, y = lgemissions)) +
              geom_point(alpha = 0.3, color = pubblue) +
              # facet_wrap(facets = vars(bins), scales = "free") +
              geom_smooth(method = 'lm', se = FALSE, color = pubred) +
              labs(x = 'Log GDP in US Dollars',
                   y = 'Log Emissions in Kilotons',
                   title = 'Histogram of GDP in US Dollars',
                   col = NULL)

completegg %>% pub(type = "scatter")
```




```
lm <- lm(lgemissions ~ lggdp, data = complete)
resid <- data.frame(x = residuals(lm))
residgg <- ggplot(resid, aes(x = x)) +
```

```
            geom_histogram(binwidth = 0.2, fill = pubblue, color = "black") +
            labs(x = 'Residuals',
                 y = 'Frequency',
                 title = 'Histogram of Residuals',
                 col = NULL)

residgg %>% pub(type = "hist")
```



It seems there is a linear relationship between GDP and CO_2 emissions, as most data points are clustered around the regression line, and there don't seem to be any outliers. The residuals are also approximately normally distributed, as the histogram of residuals shows a bell curve with a slight right-skew. This means that a linear regression model seems appropriate for the data. We also check for constant variance:

```
variance <- complete %>%
            group_by(bins) %>%
            summarize(variance = var(lgemissions),
                      mean = mean(lgemissions))

knitr::kable(variance,
             caption = 'Variance per logGDP bin',
             col.names = c('Bin',
```

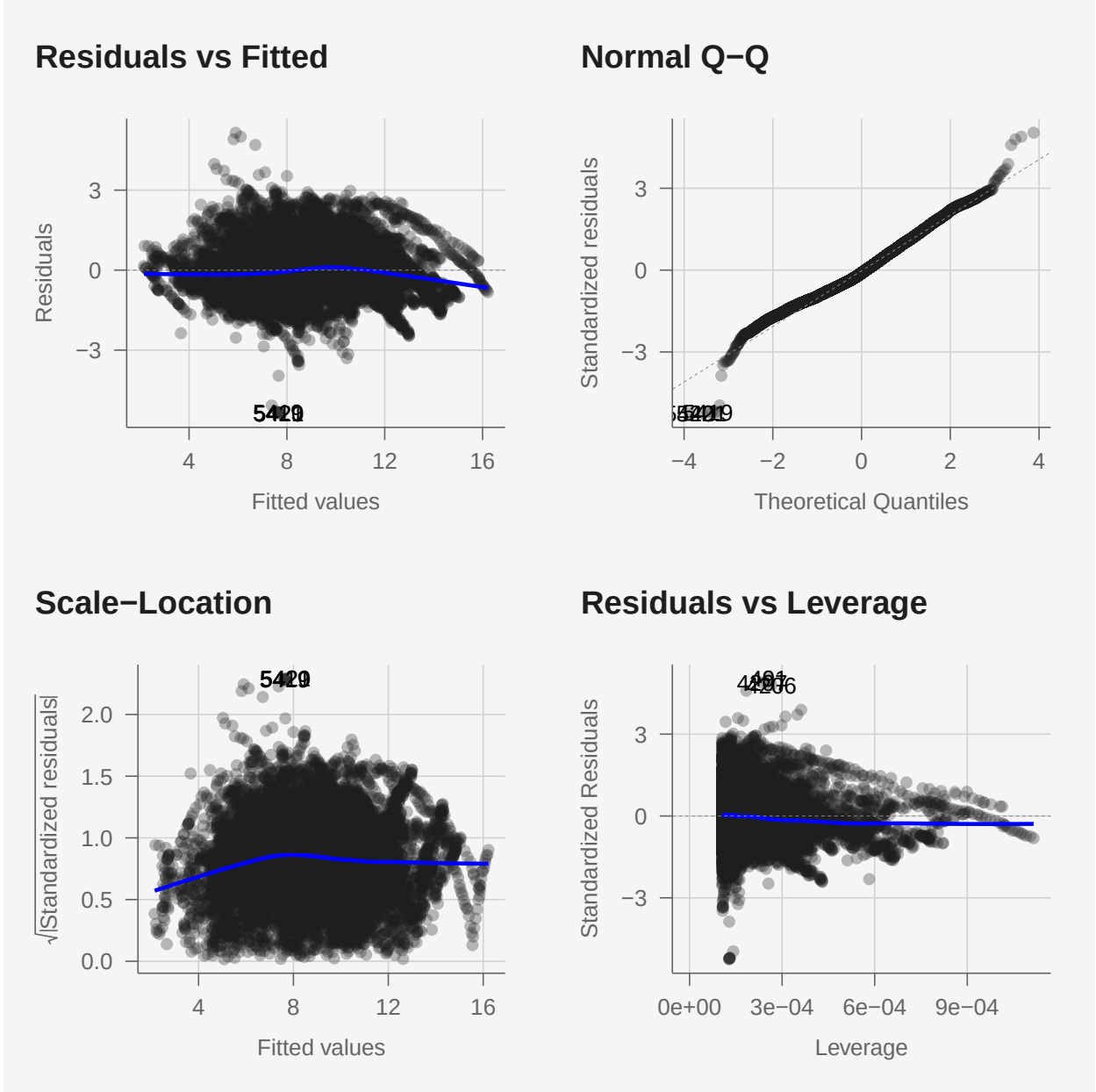
```
                        'Variance',
                        'Mean'),
          digits = 3)
```

Table 4: Variance per logGDP bin

Bin	Variance	Mean
(16,17.5]	0.388	2.962
(17.5,18.9]	0.806	4.374
(18.9,20.4]	1.238	5.693
(20.4,21.9]	1.458	6.997
(21.9,23.3]	1.308	8.495
(23.3,24.8]	1.178	10.042
(24.8,26.3]	0.871	11.412
(26.3,27.8]	1.249	12.297
(27.8,29.2]	0.661	13.487
(29.2,30.7]	0.576	15.220

Although the variance of the residuals is not constant across all values of the independent variable, the variance is not significantly different across the bins. It seems that the variance of log emissions for the wealthiest and the poorest countries differs the most from the variance of log emissions for the rest of the countries. Nevertheless, this doesn't appear to be a significant issue, and we can further check the assumptions of the linear regression model.

```
autoplot(lm, col = pubdarkgray, alpha = 0.3) + theme_pub()
```



From the residuals vs. fitted plot, we see that the residuals are randomly scattered around the center, and from the Q-Q plot, we see that the residuals are approximately normally distributed. From the Scale-Location plot, we see that the variance of the residuals is similar for all values of the independent variable. Finally, from the Residuals vs. Leverage, we see that there are no potential outliers. This means that the linear regression model is a good fit for the data.

We now plot log emissions against log GDP for all countries, colored by world region and faceted by development status.

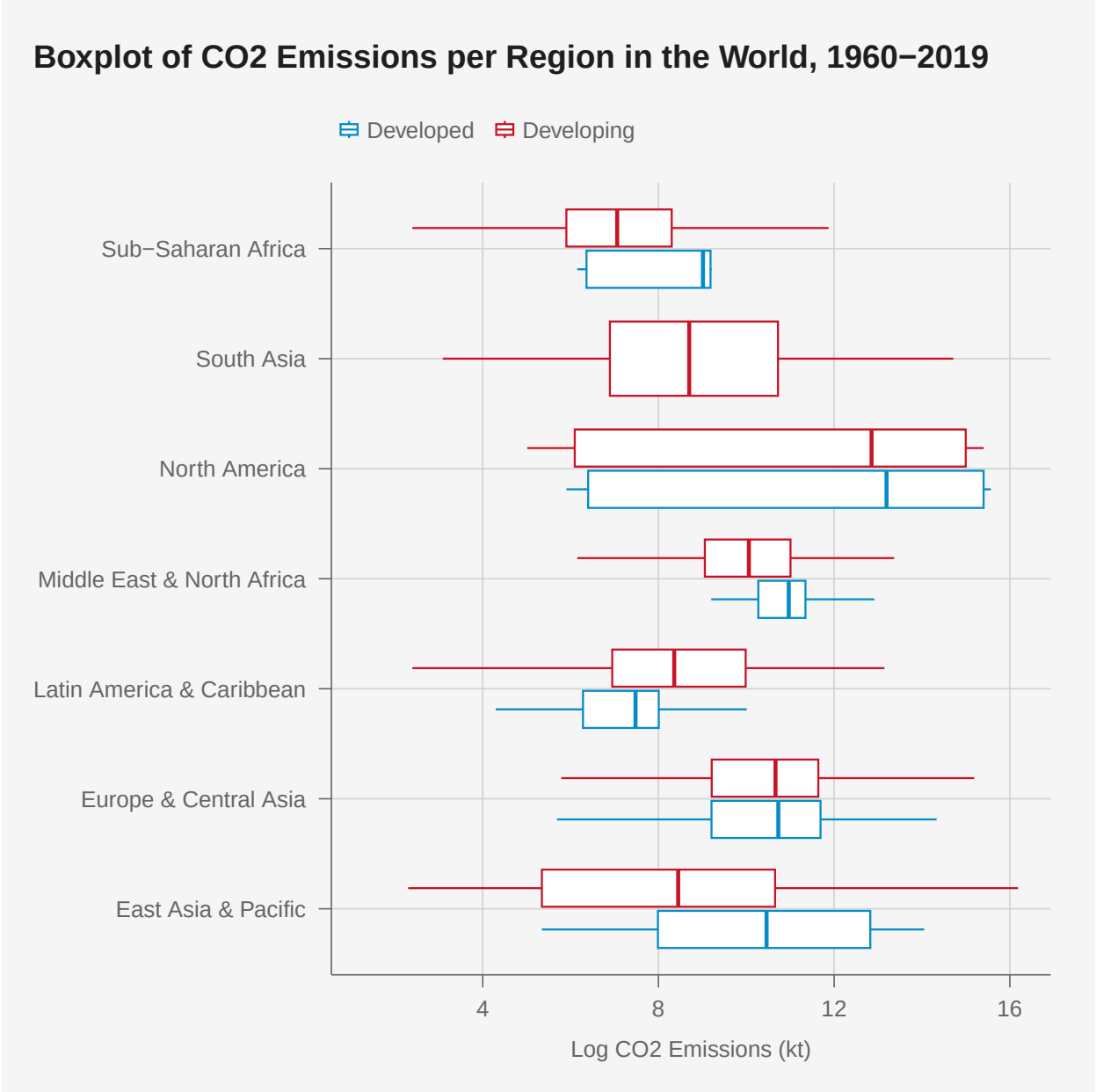
```
complete %>%
  ggplot(aes(
    x = lggdp,
    y = lgemissions,
    col = wregion
```

```
)) +
geom_point(alpha = 0.3) +
facet_wrap(facets = vars(status)) +
labs(x = 'Log GDP (USD)',
     y = 'Log CO2 emissions (kt)',
     title = 'Historical GDP against CO2 Emissions',
     col = 'World Region') +
theme_pub() +
theme(plot.title = element_text(size = 20),
      axis.title = element_text(size = 18),
      strip.text.x = element_text(size = 18),
      axis.text = element_text(size = 18),
      legend.title = element_text(size = 20),
      legend.text = element_text(size = 18),
      legend.position = c(.95, .05),
      legend.justification = c('right', 'bottom'),
      legend.margin = margin(0,0,0,0))
```

The plot shows that, from 1960 to 2019, both developed and developing countries have followed a similar pattern, in which the higher the total GDP, the higher their CO_2 emissions, regardless of development status or world region. This is interesting, since it is generally accepted that wealthier countries (developed) emit less CO_2 because they are able to substitute into sustainable energies and materials, but the plot shows the opposite, which could suggest wealthier countries, although more technologically advanced, consume so much that their sustainable efforts are outweighed by their consumption. It is also interesting to see how the vast majority of developed countries are in North America, Europe, Central Asia, and Latin America & Caribbean, while the vast majority of developing countries are found in Africa, the Middle East, and South

Asia.

```
complete %>%
  ggplot(aes(
    x = wregion,
    y = lgemissions,
    col = status
  )) +
  geom_boxplot(outlier.size = -1) +
  labs(
    x = NULL,
    y = 'Log CO2 Emissions (kt)',
    title = 'Boxplot of CO2 Emissions per Region in the World, 1960-2019',
    col = NULL) +
  theme_pub() +
  coord_flip()
```



Interestingly, there is a lot of overlap on the CO_2 emissions per developed and developing countries.

3 Modeling and Analysis

I will first use a Linear Regression to predict whether GDP and CO_2 emissions have a strong or weak correlation, and to test whether development status and world region are good additional predictors. Then I will use a hypothesis testing to test whether developed and developing economies pollute equally.

3.1 Least Squares Regression Line Analysis

Recall we checked the Linear Regression assumptions previously, so we now can make models to predict emissions based on GDP, development status, and world region. It is also reasonable to assume that the predictors are not strongly correlated, since the world region and development status of a country are not directly related to the GDP of a country. Additionally, instead of using every world region, we devide the world regions into two categories: Western and Non-Western, since the majority of developed countries are in the Western world (North America and Europe). Thus, we first create two indicator variables, one for the development status of a country and one for the world region of a country.

```
complete <- complete %>%
  mutate(
    developed = ifelse(status == 'Developed', 1, 0),
    western = ifelse(wregion %in% c('Europe & Central Asia',
                                    "North America"), 1, 0),
    region = ifelse(wregion %in% c('Europe & Central Asia',
                                   "North America"), 'Western', 'Non-Western'))

modeld <- lm(lgemissions ~ lggdp + western + developed, data = complete)
modele <- lm(lgemissions ~ lggdp + western, data = complete)
modelf <- lm(lgemissions ~ lggdp + developed, data = complete)
modelg <- lm(lgemissions ~ lggdp, data = complete)
modelh <- lm(lgemissions ~ developed + western, data = complete)
modeln <- lm(lgemissions ~ 1, data = complete)

stargazer(modeln, modeld, modele, modelf, modelg, modelh, type = 'text',
          column.labels = c('Intercept Only',
                            'All Predictors',
                            'logGDP and Region',
                            'logGDP and Status',
                            'logGDP Only',
                            'Status and Region'),
          omit.stat = c('f', 'ser'),
          notes.align = 'l',
          notes = "Standard errors are in parentheses.",
          notes.append = FALSE)
```

```
==========================================================================================================
                                                Dependent variable:
            ------------------------------------------------------------------------------------------
                                                      lgemissions
            Intercept Only All Predictors logGDP and Region logGDP and Status logGDP Only Status and Re
                 (1)            (2)              (3)               (4)            (5)            (6)
----------------------------------------------------------------------------------------------------------
lggdp                          0.992***        0.943***          1.013***       0.957***
```

		(0.004)	(0.004)	(0.004)	(0.004)	
western		0.491***	0.233***			2.009***
		(0.025)	(0.027)			(0.063)
developed		-1.185***		-1.037***		0.791***
		(0.030)		(0.030)		(0.074)
Constant	8.780***	-13.883***	-12.887***	-14.286***	-13.154***	8.189***
	(0.027)	(0.097)	(0.101)	(0.096)	(0.096)	(0.029)
Observations	9,445	9,445	9,445	9,445	9,445	9,445
R2	0.000	0.870	0.849	0.865	0.847	0.143
Adjusted R2	0.000	0.870	0.848	0.865	0.847	0.142

Note: Standard errors are in parentheses.

It seems the best model is (2), since all the coefficients are statistically significant, the adjusted R^2 is the highest among all the other models, which indicates it can explain more of the variation in the data, and the coefficient signs make sense. We would expect that being a developed nation is correlated with lower emissions, since developed nations often outsource their production to developing nations, and they can switch towards new technologies that reduce their impact on the environment. Model (2) shows that being a developed nation is associated with a 1.185 decrease in the log emissions, or a 118.5% decrease in emissions. Additionally, we would expect that being part of the Western world is correlated with higher emissions, since developing nations close to developed nation in the Western world are more likely to have higher emissions due to the demand for goods and services from developed nations. Model (2) shows that being part of the Western world is associated with a 0.491 increase in the log emissions, or a 49.1% increase in emissions. Lastly, we would expect that every increase in GDP will increase emissions, since GDP is a measure of consumption and production. Model (2) shows that a 1% increase in GDP will increase emissions by 0.992%.

3.2 Hypothesis Testing

We run a two-sample t-test because we want to test whether there is a difference in the true mean of CO_2 emissions for developed and developing countries, with the null hypothesis being that the difference in means is zero.

Assuming samples are independent, meaning that a country's GDP and CO_2 emissions will not affect another country's GDP and CO_2 emissions, we have to verify whether the data follows a normal distribution with QQ plots. Recalling the previous QQ plot in the document, we can confirm that the CO_2 emissions approximately follow a normal distribution. This also means that the data for each category fits the theoretical distribution for each category.

We can try using a two sample t-test with the null hypothesis H0 being that the mean CO_2 emissions of developed and developing countries is equal and alternative hypothesis Ha being otherwise. Since both groups have more than 30 countries (n > 30), then the Central Limit Theorem states that the sampling distributions of the sample means will be approximately normal, so even if the variances of the two groups are not equal we can perform a standard two sample t-test.

```
developed <- complete %>% filter(status == 'Developed')
developing <- complete %>% filter(status == 'Developing')

t.test(developed$lgemissions, developing$lgemissions, var.equal = T)
```

```
    Two Sample t-test

data:  developed$lgemissions and developing$lgemissions
t = 22.488, df = 9443, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.488542 1.772821
sample estimates:
mean of x mean of y
10.157285  8.526603
```

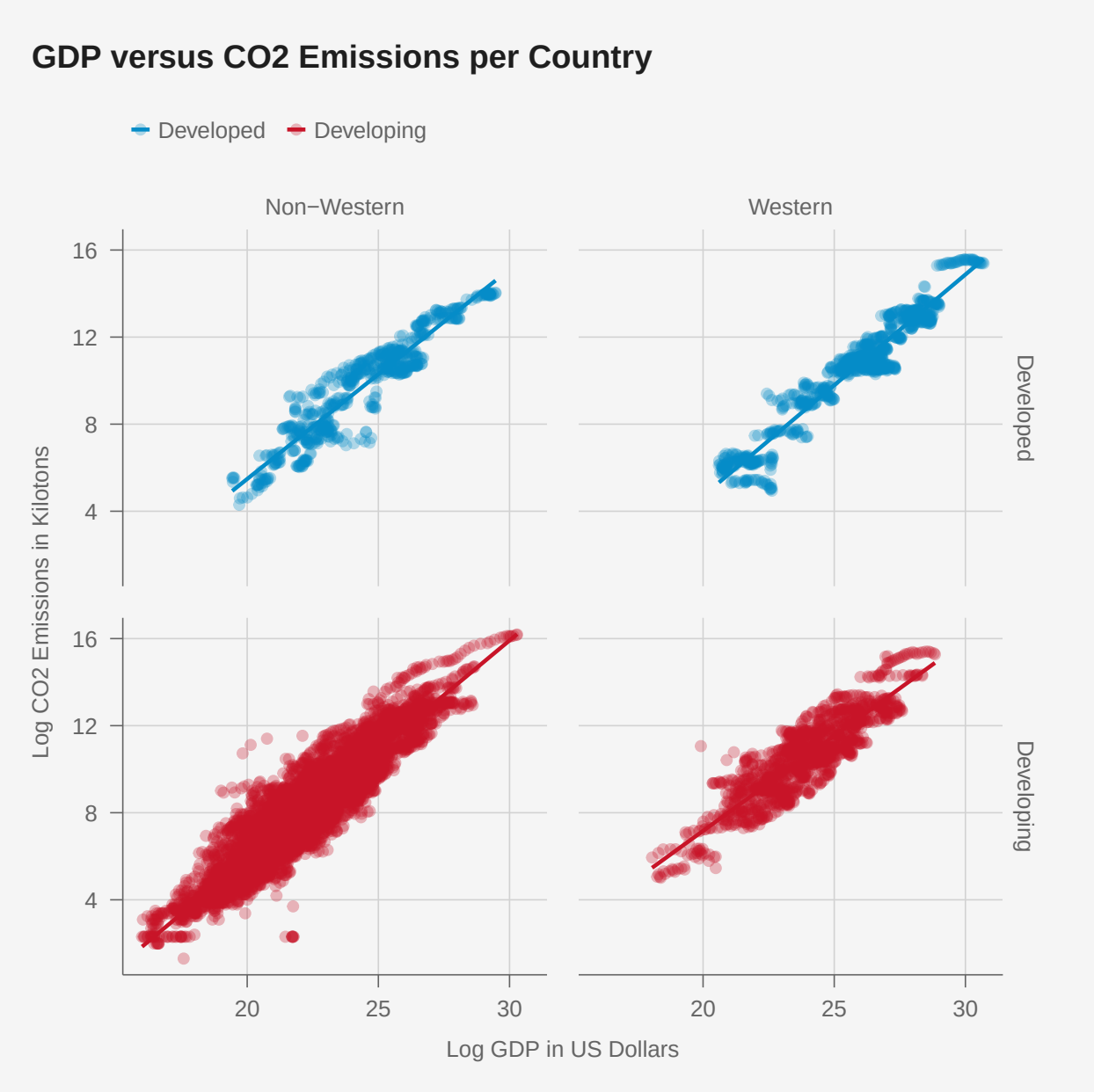
Since the p-value is lower than our significance level (alpha = 0.05), then we reject the null hypothesis that the mean log CO_2 emissions in kt for developed and developing countries is the same. Therefore, there is convincing evidence to conclude that there is a difference in the mean CO_2 emissions in Kilotons for developed and developing countries. More specifically, developed countries have higher mean CO_2 emissions. This result makes sense since wealthier countries produce and consume much more than poor countries.

One key assumption for this report is independence between samples. This would imply that every economy only utilizes the resources they have available in their territory. In the past, this may have held true. However, given the modern world we live in today, this assumption is incorrect, since many economies rely on each other for stability. For instance, the United States relies on other countries for a continuous supply of food and goods, while other countries rely on developed economies, like the United States, to buy the goods they produce. This means that a developed country's wealth allows it to import goods from other developed or developing economies, thus raising that country's GDP and CO_2 emissions as demand increases, incentivizing an increase in supply. This is a cycle, meaning the more a country's goods are demanded, the more that country is incentivized to produce, which will increase GDP and emissions. Therefore, we cannot actually conclude independence across samples due to the shown global economic dependence. Thus, hypothesis testing may not be the best method to show whether developed countries pollute more than developing countries. Nevertheless, we know there is a real connection between a country's wealth and its environmental impact, but using a hypothesis test on this specific data set may not be appropriate, even if the result is what we expect.

The data used is also not adequate for a Chi-Square test, since there are multiple observations for every country. This could be fixed by narrowing the study to a single year so that every country only has one entry, and taking a random sample of the countries. Nevertheless, the goal of this study is to find a historical relationship between economic development and impact on the environment, so we would have to perform individual Chi-Square tests for every year since 1960, which would be very time consuming.

4 Interpretation of the Results

```
complete %>%
  ggplot(
    aes(x = lggdp,
        y = lgemissions,
        col = status)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(
    x = 'Log GDP in US Dollars',
    y = 'Log CO2 Emissions in Kilotons',
    title = 'GDP versus CO2 Emissions per Country',
    col = NULL
  ) +
  facet_grid(status ~ region) +
  theme(plot.title = element_text(size = 20),
        axis.title = element_text(size = 18),
        legend.text = element_text(size = 16)) +
  theme_pub()
```



It seems that developing countries have higher CO_2 emissions at the same level of total GDP compared to developed countries. The plot also shows that there seems to be a strong, positive correlation between a country's GDP and its CO_2 emissions regardless of the economic development status, which agrees with our linear regression models. Lastly, the plot shows that, in general, countries in the Western world have higher CO_2 emissions than countries in the Non-Western world.

5 Conclusion and Reflection

The results of the regression model show that GDP is indeed a good indicator of CO_2 emissions, which makes sense because GDP is measurement of consumption and production, so the more a country produces and consumes, the more the country is going to pollute due to a higher demand of natural resources.

Although the hypothesis test showed that developed economies pollute more, I expected the difference to be much larger. Based on other statistics, such as kg of meat consumed per year and cars per family, it makes sense that developed countries will have much higher CO_2 emissions, since they consume colossal amounts of natural resources. The fact that the report shows that developed countries emit similar amounts of CO_2 can only be taken with a grain of salt, since there are many confounding factors that influence a country's emissions. More than anything, I believe we are ignoring the exploitation of natural resources in developing countries, since many of the most powerful companies, such as Tesla and Rio Tinto, own mines in developing countries, so the GDP and emissions are counted for the developing country, but the developed country that owns the company is actually responsible for the emissions and the benefits of production almost never show in the local economy. Therefore, I believe the results of the t-test are "toned down", meaning the difference in means should be much larger.

In terms of the model and t-test, there are some factors that can be improved on in the future. In this report I used GDP per capita to separate developed and developing countries, so my LSRL model only determined whether GDP is a good predictor of CO_2 emissions for developed and developing countries. In the future, I believe that collecting data of GDP per capita and CO_2 emissions per capita will help produce a LSRL model and a hypothesis test that can determine whether GDP per capita is a good predictor of individual emissions and whether individuals impact the environment differently depending on their wealth. If possible, it could be better to test for separate social groups within the country, since many of the wealthiest individuals in developing countries are not a good representation of the rest of the population. I believe that only looking at total GDP and total CO_2 emissions can overlook many confounding variables, such as national wealth disparity, so reducing the model to compare per capita data will help create more accurate results. Another point of improvement is to minimize the number of NA values for GDP and CO_2 emissions. In my data set, many countries have years with no record of their total GDP and/or CO_2 emissions, so the data set is not truly from 1960 to 2019. This is much harder to solve, but it is a factor to improve on for future data collection.