Final Project Report_Factors that Lead to Global Warming

Rain Hu, Phuc Nguyen, Ben Miller, Jinyao Gao, John Lorge

2022-12-05

Introduction

Purpose of Research

- Global warming has been the center of debates for two decades now. It is especially being picked up in the recent years. Whether these debates have been political, personal, or scientifically based, they are highly polarize with millions of different opinions. All kinds of scientific research have proven that global warming is a collective result of multiple factors. Considering these issues are so compelling and relevant to our daily life, we thought it would be fitting to pick this topic and seek answers through data analysis. We will look at several factors that could have led to global warming and see which one of them has the largest impact on global warming.

The **research questions** we are going to ask are:

1. What is the correlation between CO2 emissions and population growth in the United States?
2. Between population growth and carbon dioxide emission, or both, what is/are the leading factor(s) of global warming?

We thought this was an extremely compelling question because, for years, individuals have argued whether increasing CO2 emissions is directly tied to the increase in human population. As a group, we are very excited to sort and analyze data to answer our questions proposed on global warming, and to play our part in a massive societal argument and issue.

We'd like to explore the **hypothesis** that:

1. the relationship between human population and carbon dioxide is non-linear.
2. increase in human population contributed more to the increase in global temperature than carbon dioxide emission.

Background

Dataset Description

- **In the first dataset**, which is the carbon dioxide(CO2) emission rate, the key variables are the total CO2 emission rate, country and year. The total CO2 emission rate is the total amount of carbon dioxide produced from fossil fuel consumption and cement production. The year indicates the time these data is collected. The country indicates the nation of CO2 emission.
- We extract the data of the United States from 1960 to 2014 because they are the years with data shared by all three datasets. We also changed the unit of the emission from ton to kg.
- **In the second dataset**, which is global temperature change anomalies measured from 1880 to 2016. The variables are the source of data, measurement date, and mean temperature. The source of data is a categorical value that categorizes the data into either GISTEMP or GCAGI depending on the index this data is processed in. The date is in the unit of year. The mean temperature is the global

temperature anomaly based on a specific time frame. The GISTEMP chooses its time frame as 1951-1980; the GCAGI bases its time frame on the 20-century average temperature.

- The mean temperature variable is calculated by anomalies. For every month of the year, the Earth surface temperature is measured via thermometers in thousands of places around the world. At each location, the anomaly is the difference between the temperature measured and the usual temperature. The anomaly is defined as the part of the temperature that does not fit with the temperature scale this location is usually at. Then the planet is divided into 2,592 even squares. The average anomaly is calculated for each square and each day of the year. Then the average of all those values is reported as the mean global temperature anomalies for that year.
- Again, we extract the year 1960 to 2014 of data. We took the mean value of the two methods as our final value. Finally, we took the absolute value of the temperature anomalies because we only care about how much the temperature is off from normal.
- **In the third dataset**, which is the human population in different countries from 1960 to 2018. The key variables are the country name, the year this population is counted, and then the population count in this country. We specifically extract the human population of the United States and the year from 1960 to 2014.
- Other than the three original datasets, we also create two datasets which are the joined dataset from the three. We did this so it is easier to plot graphs.
- **The fourth dataset** is the data of population versus temperature per year from 1960 to 2014. We created it by joining the population data set and temperature data set.
- **The last dataset** is the data of CO2 emission versus temperature per year from 1960 to 2014. We created it by joining the co2 emission data set and temperature data set.

Research Method and Plots

- To explore more about the data set, our team creates some exploratory plots. For our first research question, we asked about the correlation between carbon dioxide emission growth and population in the United States. The x-axis is the year and the y-axis is the emission per capita. We will plot a graph to show the change in emission per capita throughout the year. This graph can show us how carbon dioxide usage for each person changed throughout the years. Since population increase is a sure thing, this will let us understand how much carbon dioxide emission an individual might bring. We will use a scatter plot with a line connecting the dots and a smooth trend line.
- For our second research question, we will investigate how much influence does human population and carbon dioxide emission each has on the global temperature change. We will then compare the two to draw a conclusion on which one is the leading factor in global warming. We are going to generate two graphs. The first graph is the correlation between global temperature and population. In this graph, the x-axis is the population and the y-axis is the temperature. In this way, we can look at the correlation between population increase and temperature anomalies. We will also find the linear correlation coefficient with specific formula and run a hypothesis test to find out which relationship is strong.
- Finally, we will apply a linear regression model to predict the future trend. For question 1, we plan to run a simple linear regression model to predict the future CO2 emission per capita after 2000. For question 2, we plan to combine the graphs in question 2 and apply two simple linear regressions (SLR) to them. The first SLR uses the human population (predictor) to predict the temperature; The second SLR uses the carbon dioxide emission rate (predictor) to predict the temperature.

Analysis

Here, we will first provide some numerical and graphical summary of the data.

Numerical and Graphical Summary

This is the first 5 row of co2 emission data. There is a emission in kilogram column and a column for each corresponding year.

```
## # A tibble: 55 x 3
##    year emission_in_ton emission_in_kg
##   <dbl>           <dbl>          <dbl>
## 1  1960          788300      788300000
## 2  1961          785521      785521000
## 3  1962          814619      814619000
## 4  1963          850622      850622000
## 5  1964          887918      887918000
## # ... with 50 more rows
```

This is the first 5 rows of population of the United States data. There is a year column, each year correspond with a total human population of United States in that year.

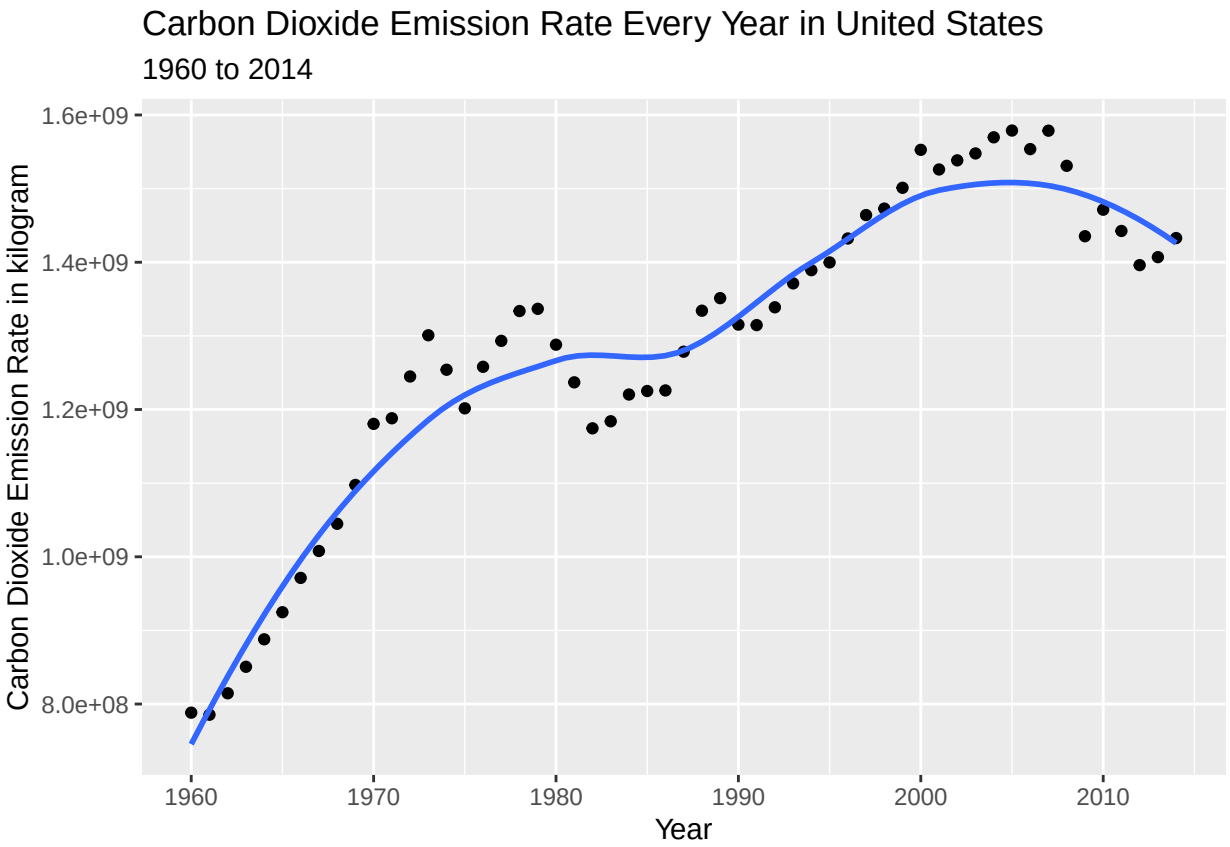
```
## # A tibble: 55 x 2
##    year population
##   <dbl>      <dbl>
## 1  1960  180671000
## 2  1961  183691000
## 3  1962  186538000
## 4  1963  189242000
## 5  1964  191889000
## # ... with 50 more rows
```

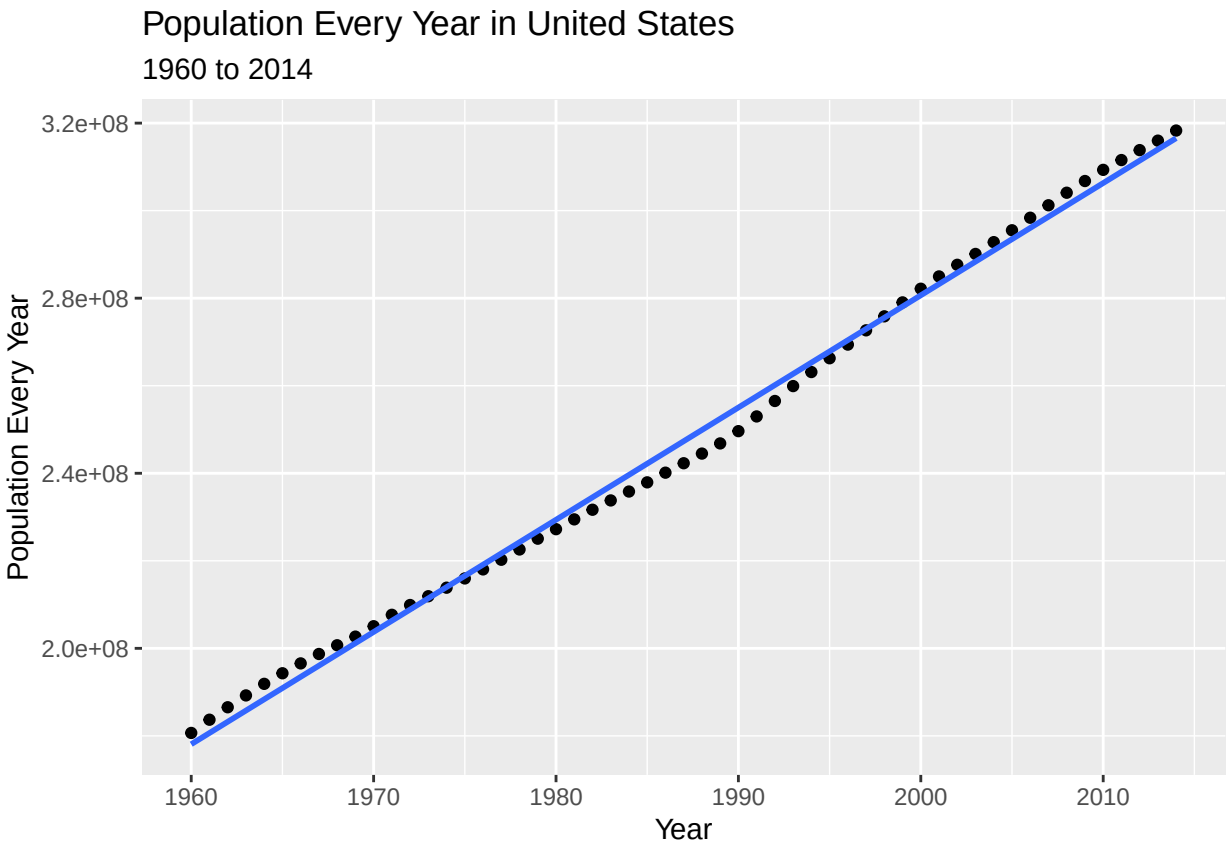
This is the first 5 rows of the temperature of the world data. There is a temperature anomalies absolute value for each year.

```
## # A tibble: 55 x 2
##    year temperature
##   <dbl>       <dbl>
## 1  1960    0.000200
## 2  1961    0.0638
## 3  1962    0.0594
## 4  1963    0.0834
## 5  1964    0.175
## # ... with 50 more rows
```

First Research Question

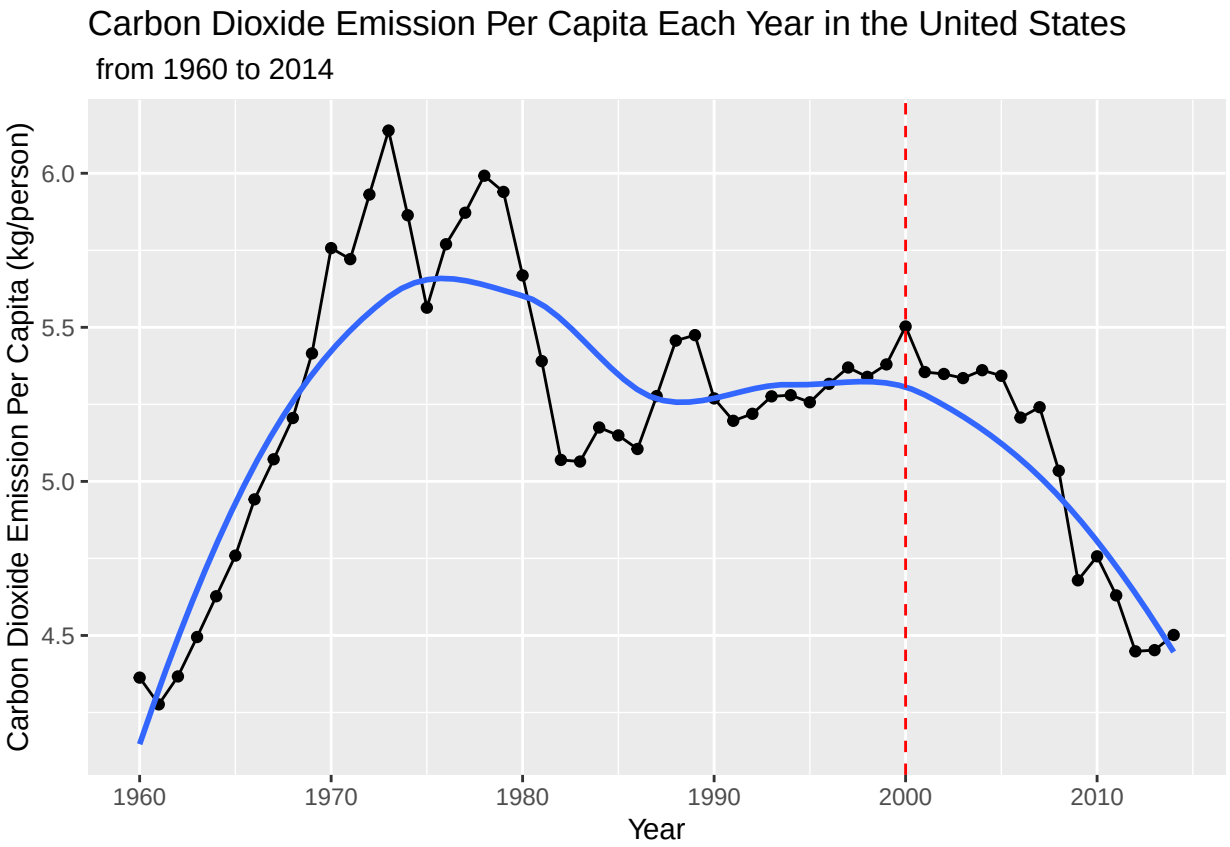
To explore more about the data set, our team creates some exploratory plots. For our **first research question**, we asked about the correlation between carbon dioxide emission growth and population in the United States. Before we examine this question, we will first take a look at some preliminary graph. We will graph carbon dioxide emission versus year and human population versus year separately.


Carbon Dioxide Emission Rate Every Year in United States
1960 to 2014
Carbon Dioxide Emission Rate in kilogram
1.6e+09
1.4e+09
1.2e+09
1.0e+09
8.0e+08
1960
1970
1980
1990
2000
2010
Year

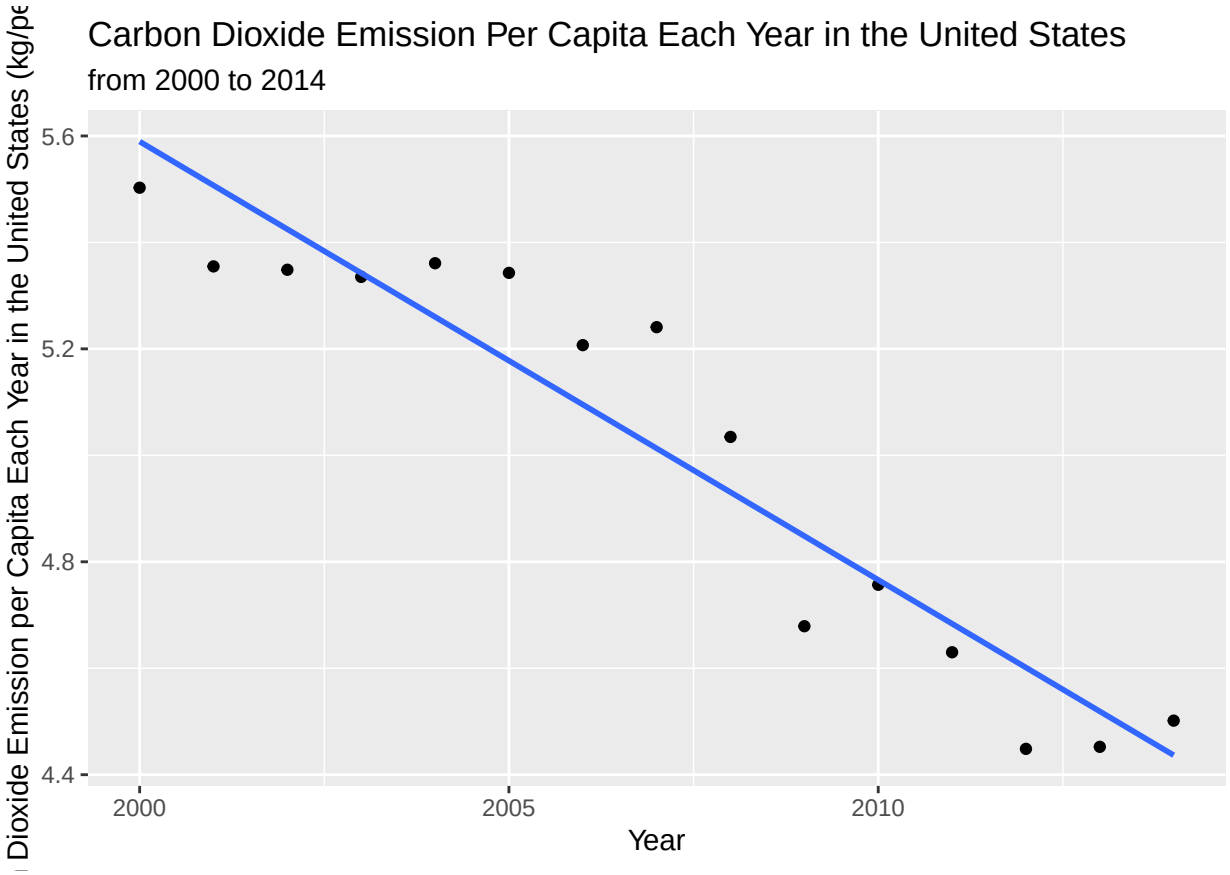


As we can see, both graph demonstrate an increasing trend within years.

To evaluate the relationship between population and CO2 emission, we introduce another variable with the formula $\frac{CO2emission}{population}$. This is the CO2 emission per capita, we refers to it in the dataset as 'proportion'. We will first plot the CO2 emission per capita versus year graph by making year the independent variable and CO2 emission per capita the dependent variable. We also placed a vertical line at year 2000 to point out an important change in the trend.



Firstly, we can clearly see that the graph is not linear; It is fluctuating from 1960 to 2014. From 1960 to 1970, the carbon dioxide emission per capita drastically increased. Then, there is a short period of dip from 1975 to 1990. After that, the increasing trend resumes from 1990 to 2000. From 2000 to 2014, the trend is pretty consistently decreasing. We placed the vertical line at year 2000 for several reasons. Firstly, we are in the 21st century. This is a new era for human to use natural resources. Our awareness of using renewable energy have changed in time. Secondly, we can see a linearly decreasing trend from 2000 and all of the points before that seem irrelevant. Because of that, we decide to only use the data after 2000 as the predictor for the carbon dioxide emission per capita of future years. After filtering all the years before 2000, we graphed a linear regression line:



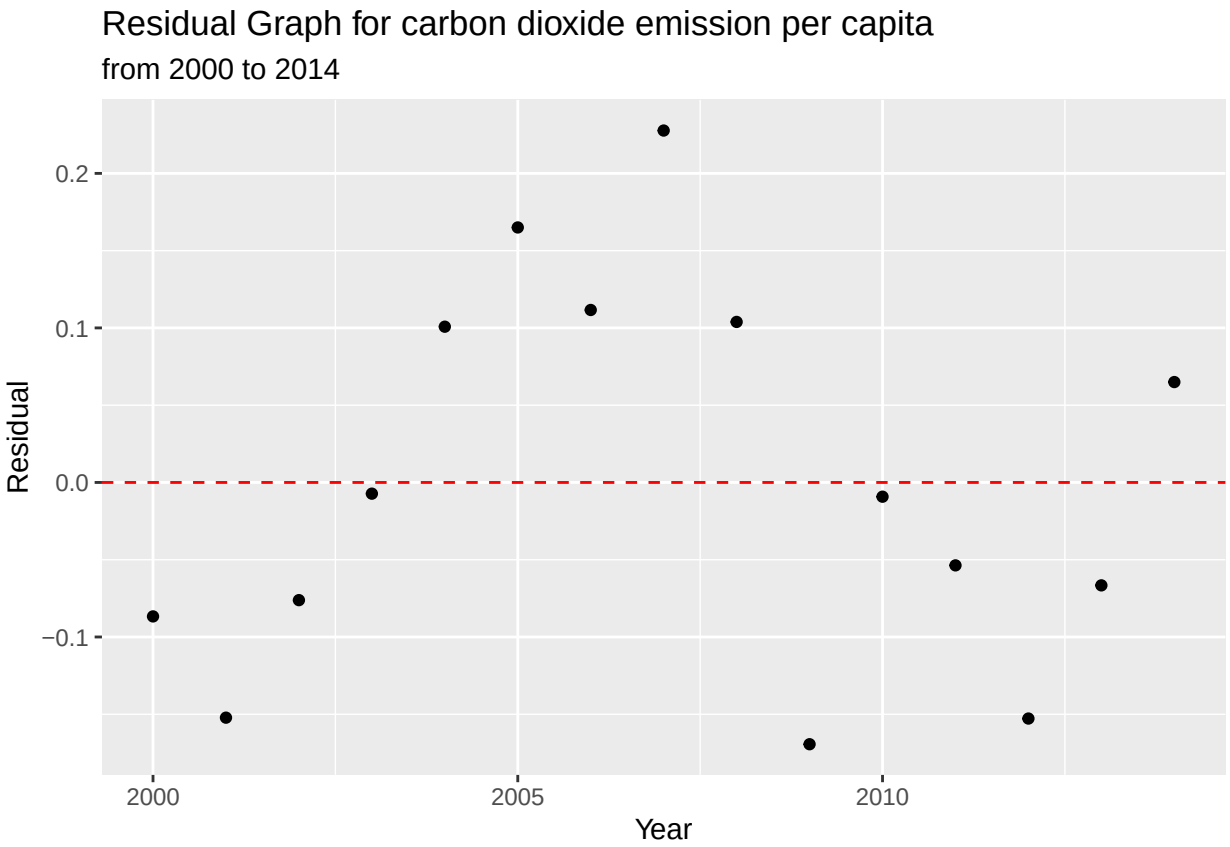
To assess the linear correlation, we will firstly introduce a formula that can help us measure the correlation strength between two variables. This is the *correlation coefficient* `r`. The correlation coefficient measures the strength of a linear relationship between two quantitative variables. The formula for correlation coefficient is:

$$r = \mathsf{Corr}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

It can also be calculated through a base R function: cor(). We will use the base R function to calculate the correlation coefficient between the co2 emission per capita and year.

```
## [1] -0.9482985
```

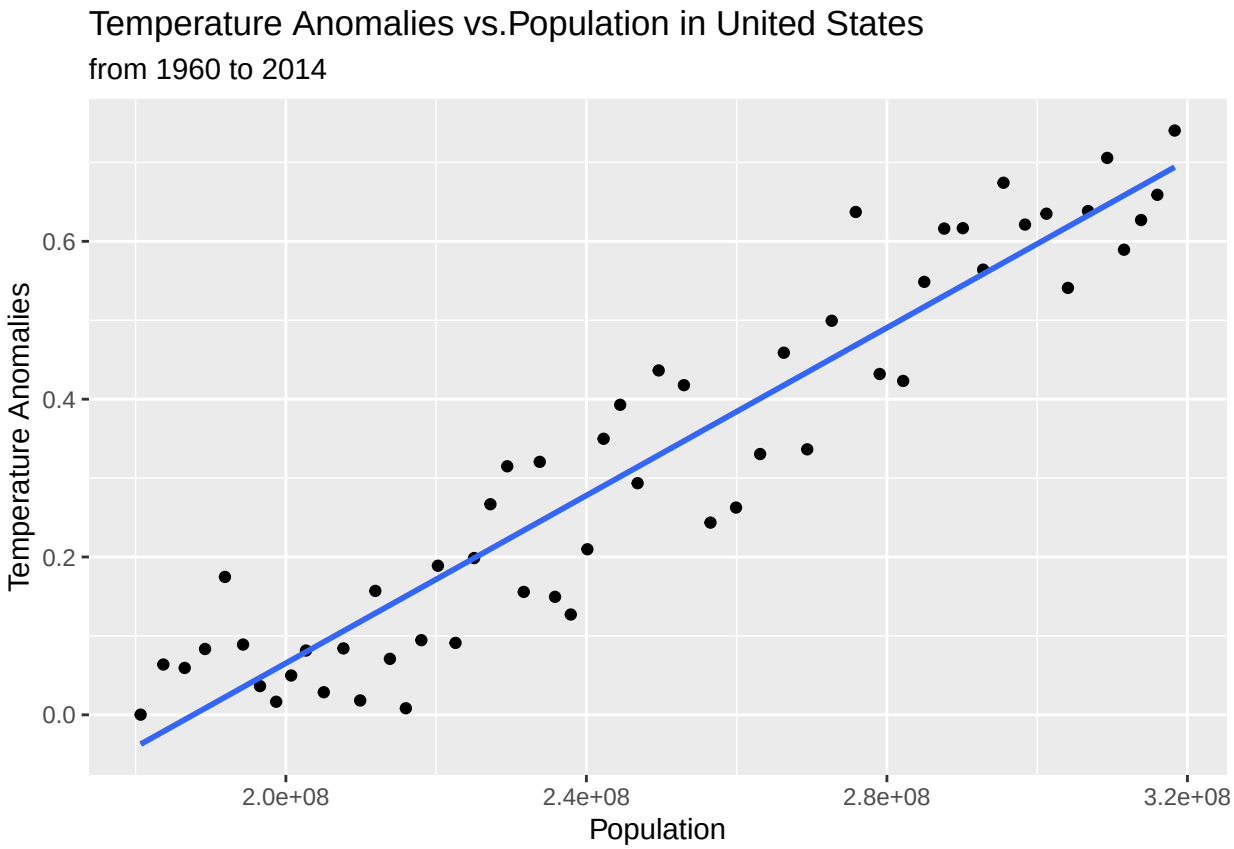
As we can see, the correlation coefficient value is approximately -0.95 which yields a strong negative correlation. To confirm if simple linear regression is the right model for this prediction, we need to look at the residual plot. The residual value is the difference between the actual value and the predicted value on the fitted line. The formula is $r = y_i - \hat{y}_i$. In our case, it is the different between the predicted co2 emission per capita and the actual co2 per capita measured and calculated for a particular year. Below, we have plotted our residual graph.

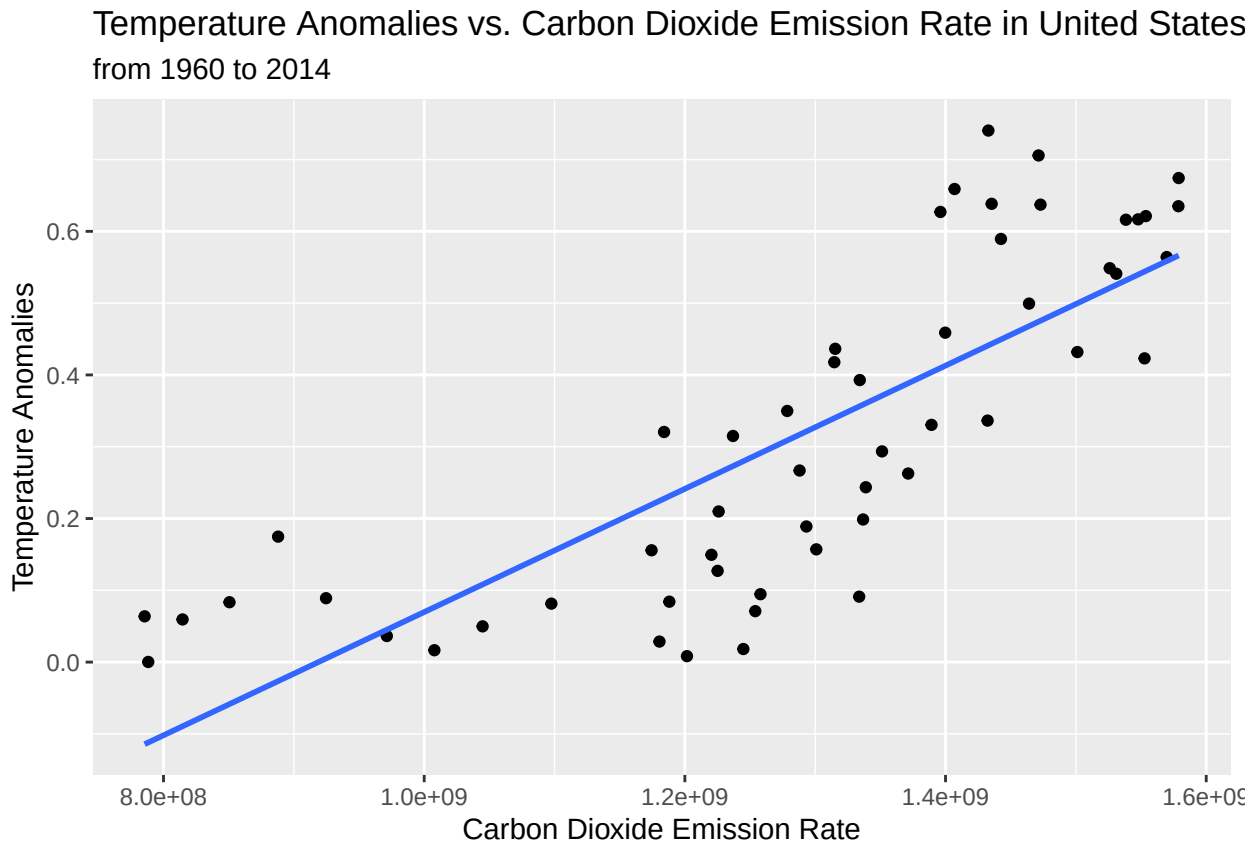


From the residual graph, we can tell that even the point above and below y = 0 is pretty evenly distributed, the residual points seems to demonstrate a different trend other than linear. We can see that there is a clear pattern for the residual point instead of randomly distributed. Thus, we can conclude that linear regression is not the best choice.

Second Research Question

To answer the second research question, we will find the rate of change (slope) of the linear regression of temperature versus population and temperature versus CO2 emissions seperatly. In both graphs, temperature is dependent variable (on the y-axis). We will then compare two values to see which one demonstrate greater effect on temperature change. Again, we will first look at some preliminary graphs. The first graph draws the relationship between temperature and human population. The second graph draws the relationship between temperature and co2 emission rate.


Temperature Anomalies vs.Population in United States
from 1960 to 2014
Temperature Anomalies
0.0
0.2
0.4
0.6
2.0e+08
2.4e+08
2.8e+08
3.2e+08
Population



From the graph, we can vaguely see that the temperature versus population slopes is larger. However, we want to look the slope numerically to see how much is one larger than the other. We will calculate the exact slope and make comparison between them. To calculate the slope, we will use the following formula:

$$\hat{b_1} = r * \frac{s_y}{s_x}$$

.

```
## [1] 5.313601e-09
```

```
## [1] 8.581691e-10
```

We will then construct a 95% confidence interval for the slopes. We will use the standard error to calculate the confidence interval. To calculate the standard error of the slope, we use the this formula:

$$s_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

```
## [1] 2.564334e-10
```

```
## [1] 8.994103e-11
```

2.564334e-10 is the population slope's standard error and 8.994103e-11 is the CO2 emission slope's standard error. Then to calculate the 95% confidence interval, we use this formula:

$$\theta : \hat{\theta} \pm t_{n-2} s_{\hat{\theta}}$$

The value of t_{n-2} can be calculated using qt(0.975, n-2). Because of the two parameters, we use n-2 as our degrees of freedom, n is the total number of data.

```
## [1] 4.799261e-09 5.827942e-09
```

```
## [1] 6.777702e-10 1.038568e-09
```

The first line is the population slope's 95% confidence interval and the second line is the CO2 emission slope's 95% confidence interval. We can see that the smallest value of the population slope is still bigger than the CO2 emission slope.

To confirm that the slope for the human population growth is larger than the slope for the carbon dioxide emission, we will use the hypothesis test to test it. We will use the standard steps to run through this hypothesis test:

1. Test model:

Let $\hat{B}_{1_{pop}}$ be the slope for the simple linear regression of the temperature anomalies vs the human popoulation growth

Let $\hat{B}_{1_{co2}}$ be the slope for the simple linear regression of the temperature anomalies vs the carbon dioxide emission

$$X_1 \mid \hat{B}_{1_{pop}} \sim Norm(\bar{B}_{1_{pop}}, s_{pop}) X_2 \mid \hat{B}_{1_{co2}} \sim Norm(\bar{B}_{1_{co2}}, s_{co2})$$

2. State hypothesis

$$H_0 : \hat{B}_{1_{pop}} - \hat{B}_{1_{co2}} = 0 H_a : \hat{B}_{1_{pop}} - \hat{B}_{1_{co2}} > 0$$

3. Test statistic

The test statistics are $\hat{B}_{1_{pop}}$ and $\hat{B}_{1_{co2}}$

4. Determine the sampling distribution for the test statistic when the null hypothesis is true

If the hypothesis is true then $\hat{B}_{1_{pop}} = \hat{B}_{1_{co2}}$. Then, we can refer to the following distribution:

$$\hat{B}_{1_{pop}} - \hat{B}_{1_{co2}} \sim Norm(0, SE_{combine})$$

where the standard error of those two is:

$$SE_{combine} = \sqrt{s_1^2 + s_2^2}$$

In this formula, s_1 and s_2 are the standard errors of the two slopes calculated above.

5. Any outcome the is greater than $\hat{B}_{1_{pop}} - \hat{B}_{1_{co2}}$ would be considered as extreme value that is against our null hypothesis. We will calculate the area of those outcomes under the curve of $Norm(0, SE_{combine})$ and call it p-value. If p-value is smaller than 0.05, then there is an evidence against our hypothesis.

6. Calculate the p-value

To calculate the p-value, we will use t-stat to calculate it. We have the following formula for calculating the t-stat:

$$\hat{t} = \frac{(\hat{B}_{1_{pop}} - \hat{B}_{1_{co2}}) - 0}{SE_{combine}}$$

Firstly, we will calulate the value of $SE_{combine}$

```
## [1] 2.717489e-10
```

Then, we will calculate the value of t-stat:

```
## [1] 16.3954
```

Next, we will calculate the p-value using that above value of t-stat with n - 2 degree of freedom. Before calculating, we can observe that t-stat > 0. Therefore, we will only consider the area on the t-curve that has $x \geq \hat{t}$ because we only test if $\hat{B}_{1_{pop}} - \hat{B}_{1_{co2}} > 0$ which is a one-sided test.

```
## [1] 1.038032e-22
```

Using the pt() function, we can see that the p-value is really close to 0, which is many times smaller than 0.01.

Discussion

For our **research question 1**, we conclude the following interpretation:

- There seems to be a more complex relationship between the co2 emission and population. We tried to fit the trend using the simple linear regression model, but the residual graph proves that simple linear regression model is not the best model to represent the relationship. The linear regression line and the correlation coefficient can be deceptive. Our guess is that a 4th degree polynomial (quartic function) would fits the trend better.
- There is a potential shortcoming of this analysis. We used data starting from 2000 to predict the carbon dioxide emission per capita in the future. It have demonstrates a decreasing trend since 2000. However, at some point, this prediction line is going to reach zero. This is not a realistic prediction because it is impossible for carbon dioxide emission per capita to reach zero. The correct trend will be approaching zero but never reaches zero. We do not have enough data/information to come up with a better prediction.

For our **research question 2,** we conclude the following interpretation:

- From our graphs, we can see that population and temperature are both increasing in times. Between population and carbon dioxide emission, population have a larger influence on the global temperature anomalies. This is because population have a larger slope, meaning the same unit increase of population lead to a larger anomalies in global temperature. We further confirmed this conclusion by constructing a 95% confidence confidence interval. The following is the interpretation from the 95% confidence interval. **We are 95% confident that slope of population versus year from 1960 to 2014 is from 4.799261e-09 to 5.827942e-09. We are 95% confident that the slope of emission in kg versus year from 1960 to 2014 is from 6.777702e-10 to 1.038568e-09.** The lower bound of population slope is still larger than the upper bound of co2 emission slope.
- We then run a hypothesis test, and conclude that **there is a strong evidence (p-value = 1.04e-22, one sided test) that the slope of the human population growth will be greater than the slope of the carbon dioxide emission.**
- For temperature anomalies versus co2 emission, linear regression seems to fit, but a more complicated correlation might better explain the relationship between the two. Our guess is that an exponential relationship would work better to make this prediction.
- The potential short-coming of this analysis is that even though the co2 emission rate and human population are both increasing in time, there is still a possibility that individual human are emitting lesser co2 but our model did not take specifically looked at this possibility.

Some potential future directions for additional work include:

- Asking if humans are the major cause for carbon dioxide? What are the other factors that would lead to a high carbon dioxide emissions rate?
- If we were to use a different method, we would choose another regression model. Possibly a model that is non-linear. To better answer the question, we need data that are more recent data, beyond 2014.

Summary

- Overall, our research conclude that there is a complex correlation between Co2 Emissions and Population Growth in the United States from 2000 calculated by CO2 emission per capita vs year. This is because our residual graph have shown an non-linear relationship. Considering this complex correlation, between the two, population growth has a more significant impact than CO2 emission because the population has a larger rate of growth in temperature.

Reference

Data Source Citation

- **CO2 emission:** Boden, T.A., G. Marland, and R.J. Andres. 2013. Global, Regional, and National Fossil-Fuel CO2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. doi 10.3334/CDIAC/00001_V2013
- **Global Temperature:** GISTEMP Team, 2022: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. Dataset accessed 20YY-MM-DD at https://data.giss.nasa.gov/gistemp/. Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. J. Geophys. Res. Atmos., 124, no. 12, 6307-6326, doi:10.1029/2018JD029522.
- **Human Population:** According to the World Bank site, the data was sourced by these researchers from the World Bank Database, which includes all data from: (1) United Nations Population Division. World Population Prospects, (2) United Nations Statistical Division. Population and Vital Statistics Reprot (various years), (3) Census reports and other statistical publications from national statistical offices, (4) Eurostat: Demographic Statistics, (5) Secretariat of the Pacific Community: Statistics and Demography Programe, and (6) U.S. Census Bureau: International Database.

Explanation for temperature anomal

- Gardiner L. (2022). How to measure global average temperature in five easy steps. *Center for Science Education*. Retrieved from: https://scied.ucar.edu/image/measure-global-average-temperature-five-easy-steps