Project Draft

Due: October 30, 11:59pm

```
library(tidyverse)
library(broom)
library(caret)
climate <- read_csv("data/climate.data.processed.csv")
knitr::opts_chunk$set(warning = FALSE, message = FALSE, echo = FALSE)
```

INTRODUCTION AND DATA

As the world struggles to convince so many people about the urgency of climate change it needs to be known that the threat we face is not just that of rising sea levels or CO2 levels—it is that of losing our homes. It is that of entire cities having to uproot and move elsewhere because they can no longer sustain themselves. Far from just a small increase in temperature, but a disruption of our lives as we know it.

Each year, tens of millions of people are driven from their homes by floods, storms, and droughts. The Ecological Threat Register, conducted by The Sydney-based Institute for Economics and Peace (IEP), measures ecological threats over 157 independent states and territories. The report projects that as many as 1.2 billion people around the world could be displaced by 2050 (Institute for Economics and Peace, 2020). The adverse effects of global climate change will induce more extreme weather, growing food and water insecurity, and rising sea levels which will cause the number of displaced people to rise (UNHCR, 2019). The IEP report additionally identifies three clusters of ecological hotspots: the Sahel-Horn belt of Africa, from Mauritania to Somalia; the Southern African belt, from Angola to Madagascar, and the Middle East and Central Asian belt, from Syria to Pakistan.

The intersection of climate change and migration requires comprehensive data analysis and solutions to the multidimensional challenges it creates (Podesta, 2019). Therefore, analyzing the dynamics between climate change predictors and displaced people not due to conflict can reveal opportunities for interventions. Our primary goal in this project is to understand the correlation between climate change indicators and the refugee flow from at-risk countries. In order to focus our analysis and delineate a more specific model, we will focus on relationship between climate change indicators and refugee data in the Middle East and Central Asian belt region, an at-risk region identified by The Ecological Threat Register.

Our dataset comes from the World Bank Development Indicators Databank. According to the World Bank, most of the data comes directly from each country in the World Bank Group's national statistical systems. The raw data itself contains many development indicators, and the series name tells us the metric or variable for which we are getting data. Within each series, the data is broken down into the data for each nation for each year between 1960-2019. The dataset additionally contains all the markers that the WDB has tracked in association with climate change in almost every country on Earth. This includes variables such as CO2 emissions levels of every country and agricultural output of each country. Unfortunately, the WDB does not have very complete data for some of the variables. However, we will focus on variables that have sufficient data, unless the variable is unlikely to change much over time.

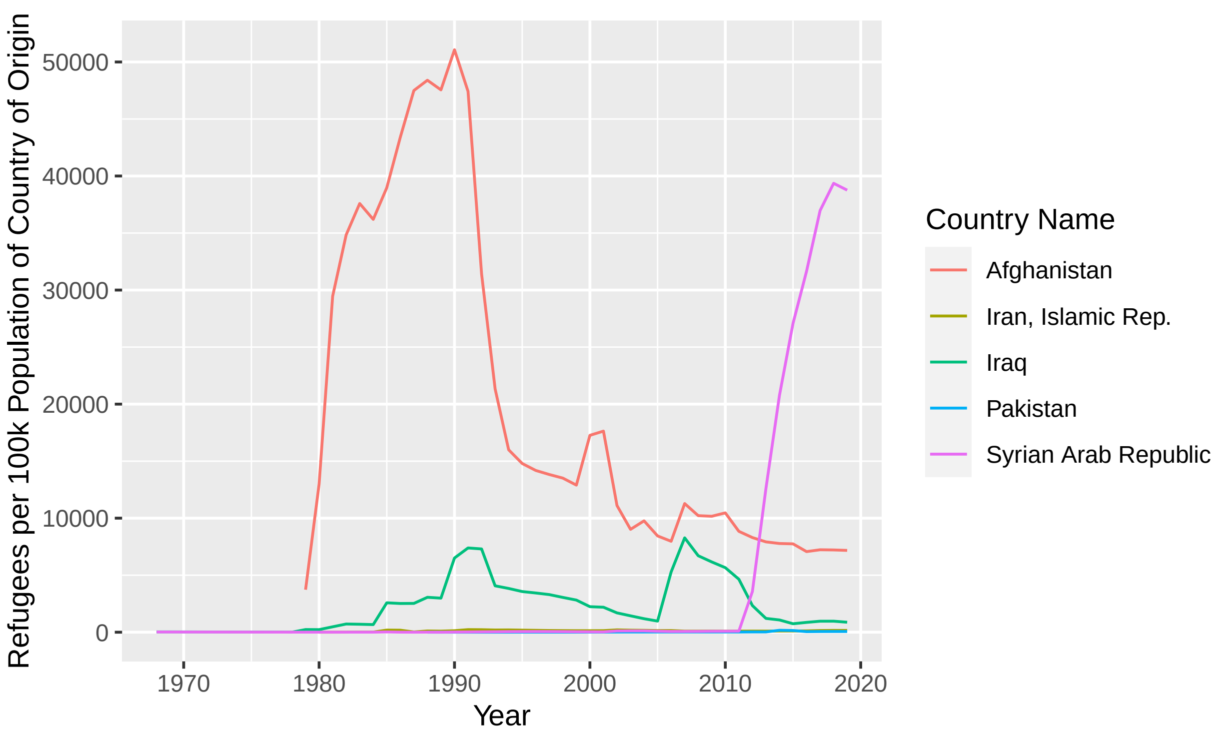
We will start by examining variables in the WDB data that scientific literature identifies as climate change predictors such as CO2 emissions (as measured in kiltons), other greenhouse emissions (in kilotons), land under cereal production, and percentage of arable land. We will also examine refugee flow out of the countries and each country's population. Then, we will analyze the correlation between refugee flow and climate predictors. The predictors will be used to build and evaluate a linear model that attempts to demonstrate if there is a relationship between predictors and refugee flow as well as explain some of the variance in refugee flow. While refugee flow from a country can be influenced by an almost innumerable amount of variables, the hope for the project is not predict all of that variance. A model that is able to significantly predict 20% of the variance in refugee flow would be a useful result. We hypothesize that CO_2 emissions (as measured in kilotons), other greenhouse emissions (in kilotons), land under cereal production, and percentage of arable land will be able to explain, with statistical significance, at least 15% of the variance in refugee flow from countries in the Middle East and Central Asian Belt.

METHODOLOGY

Visualizations and Exploratory Data Analysis

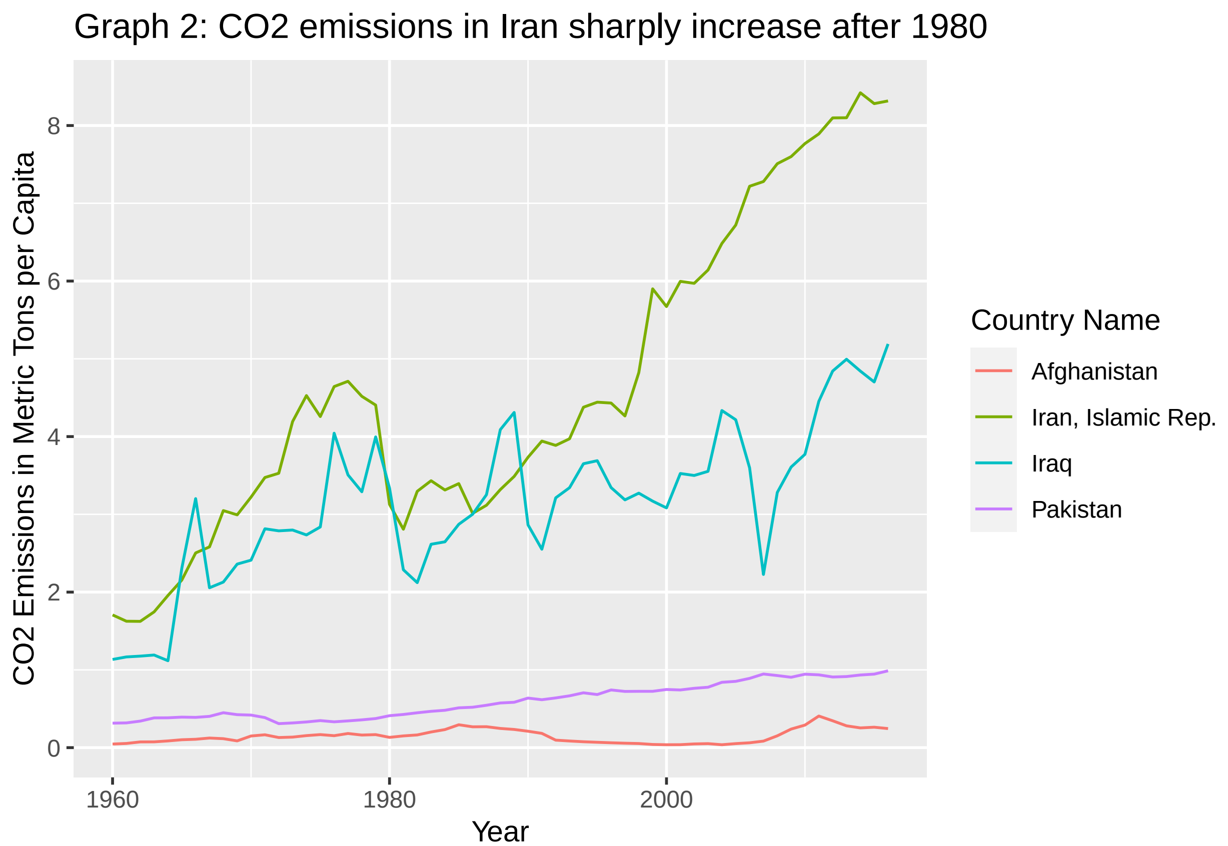
To visualize climate change predictors within each Middle Eastern country over time, we plotted the yearly proportion of population leaving as refugees, CO_2 emissions, N_2O emissions, methane emissions, percent of country's land that was arable, hectares of arable land used for cereal cultivation.

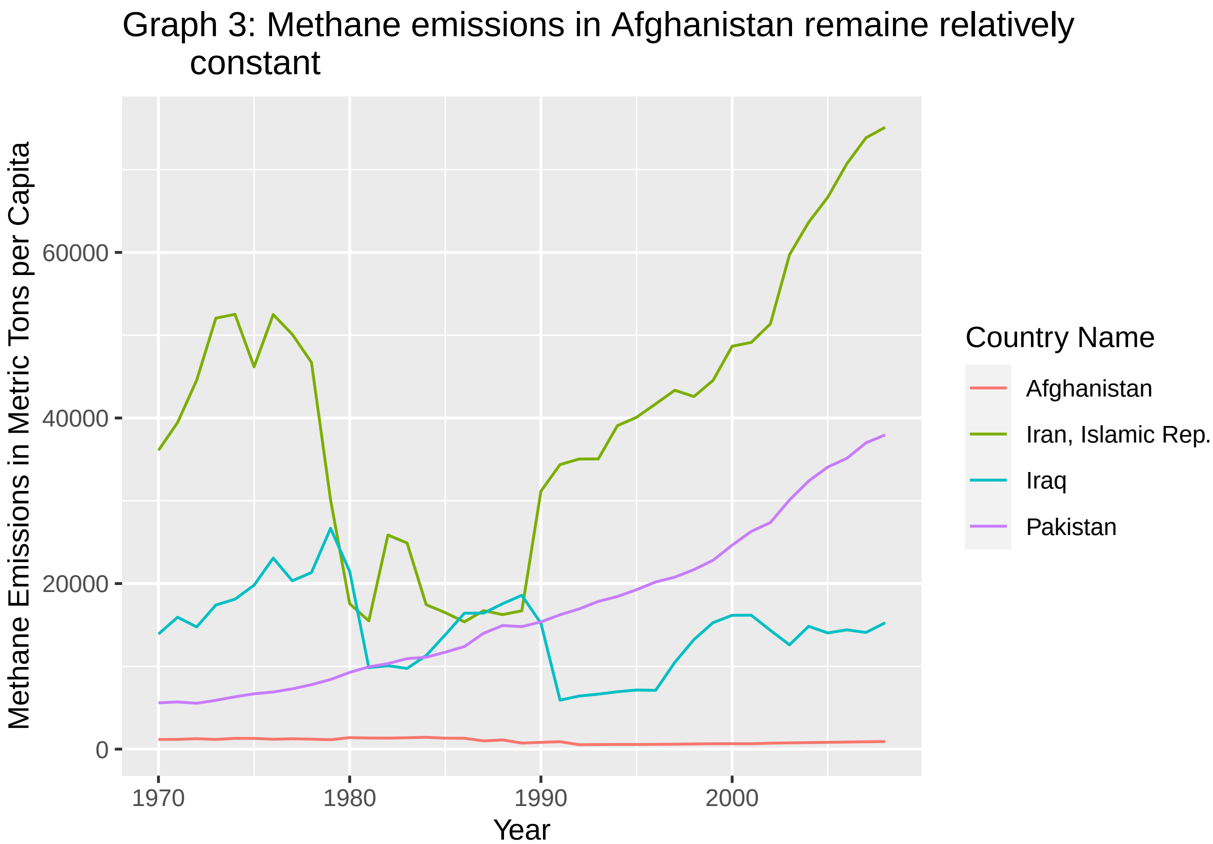
Graph 1: Number of refugees in Iran and Pakistan remained relatively similar while that in Syria increased sharply after 2011 civil war

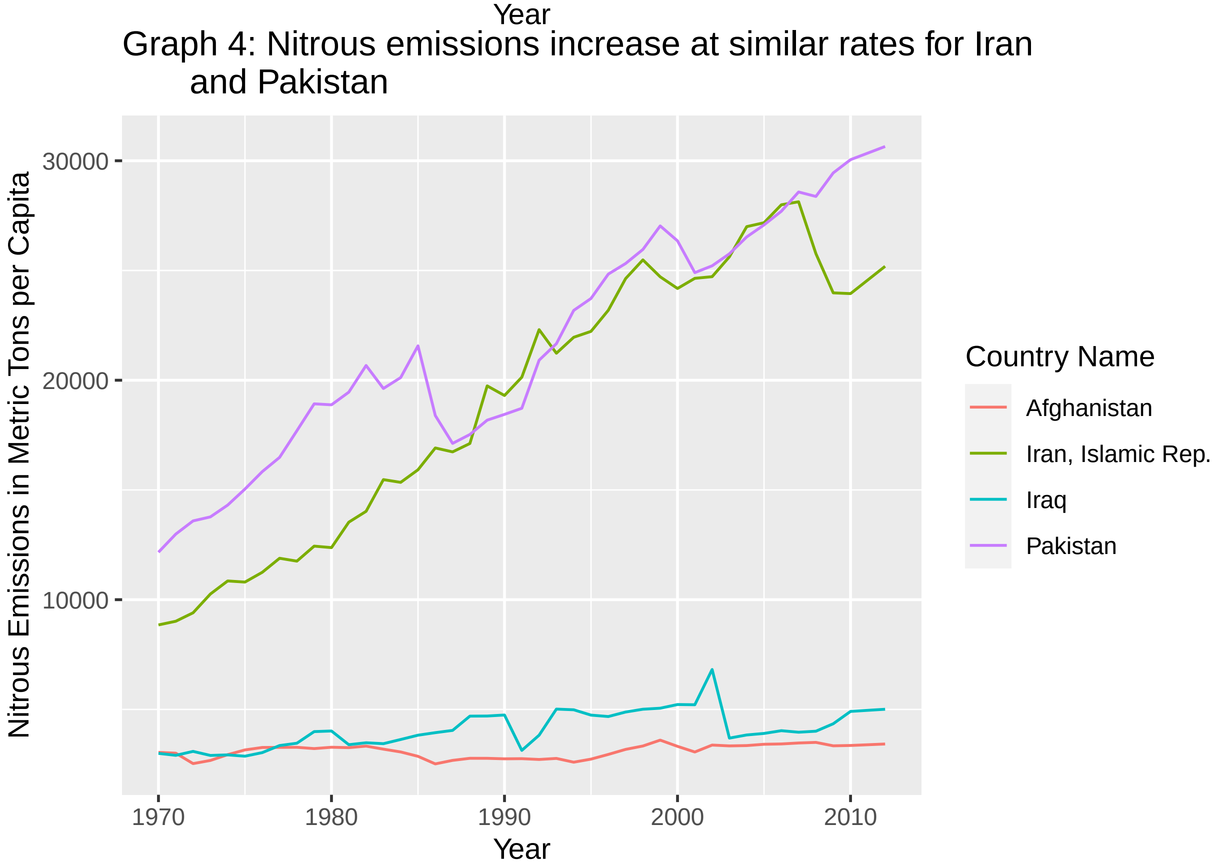


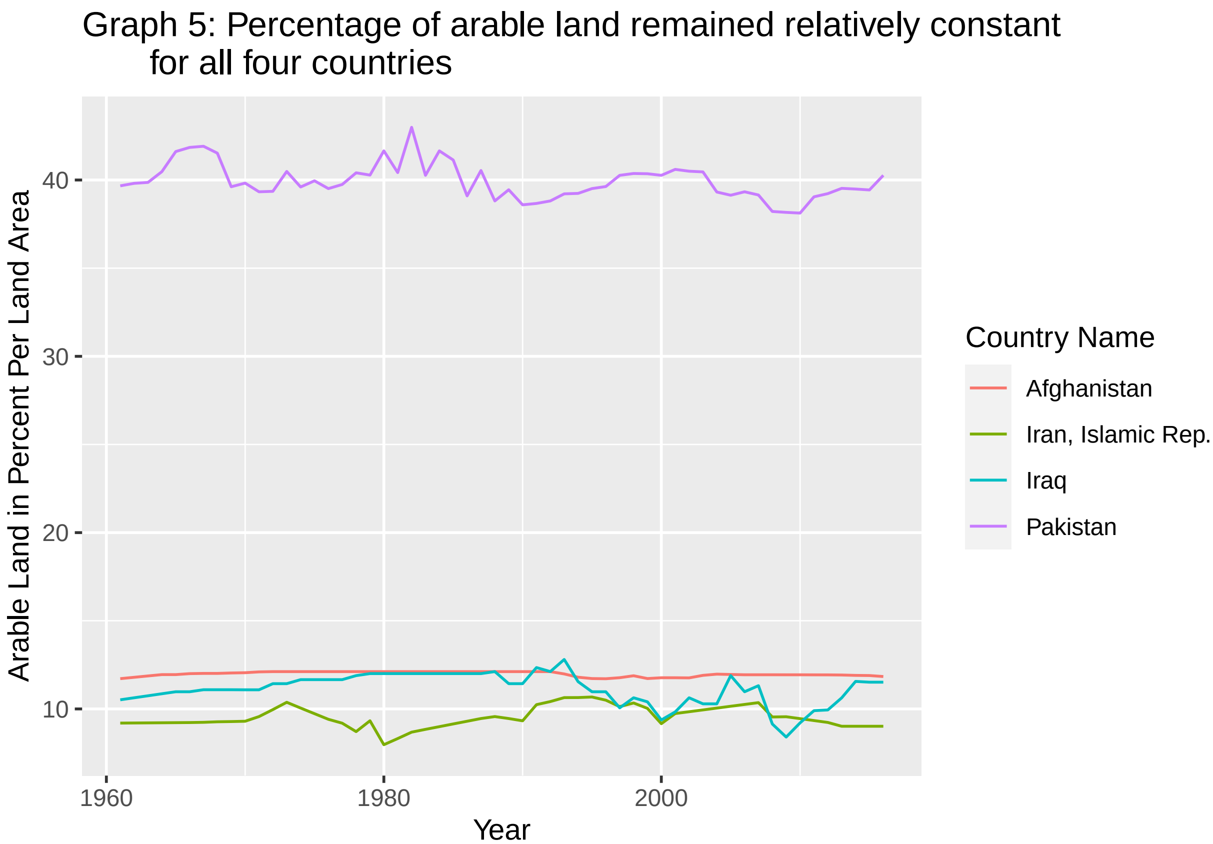
The region as generally defined includes 5 countries, Syria, Iran, Iraq, Afghanistan, and Pakistan. All of these were included in the modeling for our analysis except for Syria. Syria had barely any refugees leaving their country until the start of the Syrian Civil War in 2011. It is clear in the exploratory data analysis, specifically the line graph of refugee population (as proportion of overall population) overtime, that there is a massive spike. That massive increase in refugee population is clearly due to little more than the conflict. Thus data from that country was excluded. However, one can also see significant spikes in refugee population fleeing Afghanistan. Despite this, Afghanistan was included in the model. This is because the refugee population spikes do not correlate as well with periods of war/conflict. The number of refugees fleeing Afghanistan actually decreased dramatically during civil wars in the 1990's. While it is undeniable that Afghanistan did experience periods of civil wars and attacks by foreign powers, that alone is not enough to disqualify them.

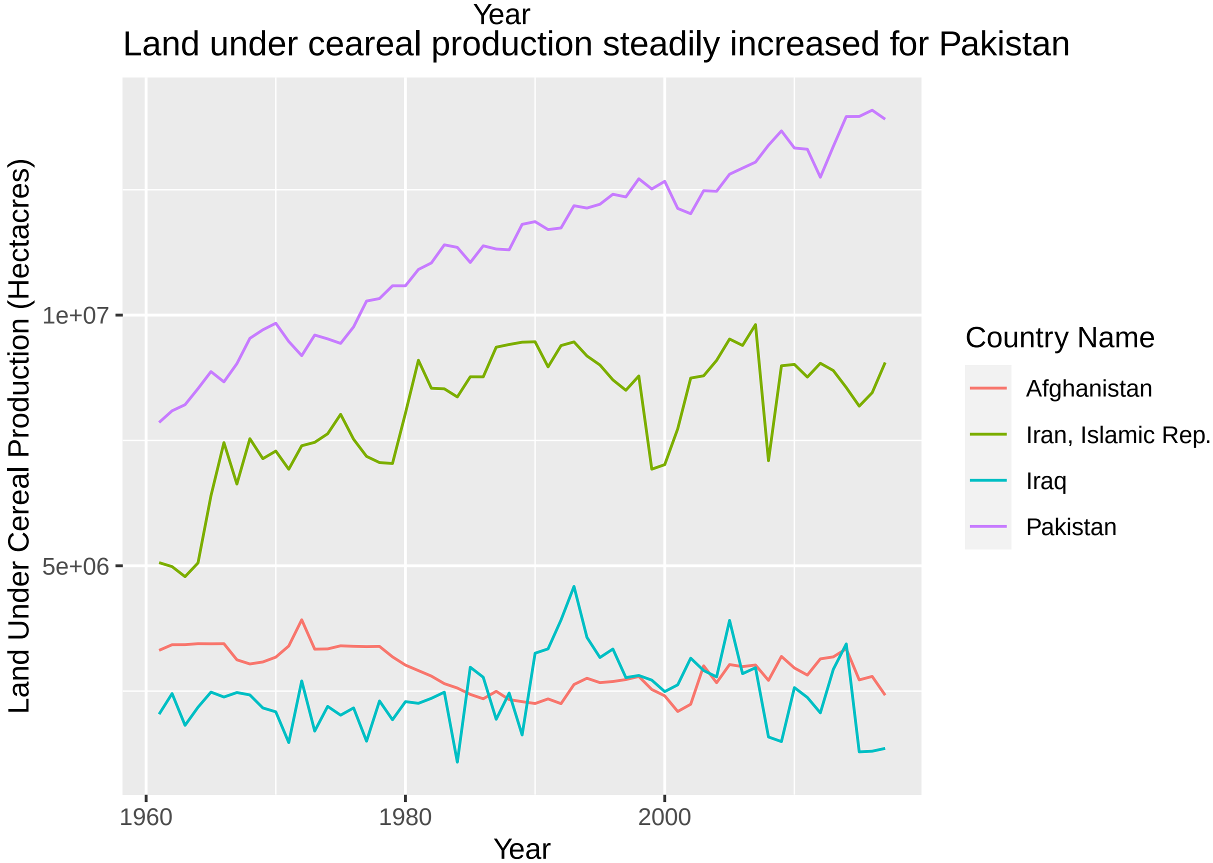
The refugee population fleeing Afghanistan has existed at significant levels regardless of period of conflict or not. This is also the case with Iraq. While the country experienced conflict during the period of study, the refugee trends do not correlate well with that conflict. In fact, the refugee population from Iraq was still slightly decreasing at the start of the American invasion. This was the justification for including Iraq in the data. More broadly, the mere existence of conflict does not justify the exclusion of a country from our analysis. It is unlikely that climate factors alone would push someone to leave their country of residence. Moreover, the goal of this project is not to entirely explain the variance in refugee population with only these factors. That said, we wanted to observe the how influential climate factors are when there are other factors (i.e. conflict) that would increase a person's desire to leave their country of residence. Thus countries that have had significant refugee populations both with and without conflict present are being included.




Graph 3: Methane emissions in Afghanistan remaine relatively constant
Methane Emissions in Metric Tons per Capita
0
20000
40000
60000
1970
1980
1990
2000
Year
Country Name
Afghanistan
Iran, Islamic Rep.
Iraq
Pakistan


Graph 4: Nitrous emissions increase at similar rates for Iran and Pakistan
Nitrous Emissions in Metric Tons per Capita
10000
20000
30000
1970
1980
1990
2000
2010
Year
Country Name
Afghanistan
Iran, Islamic Rep.
Iraq
Pakistan





Graphs 2, 3, and 4 demonstrate that the trend for greenhouse gas emissions differs for each country. Carbon dioxide, nitrous oxide, and methane emissions generally increase in Iran over time. Nitrous oxide emissions for Iraq and Afghanistan increased slightly and percentage of arable land remained relatively constant for all four countries. The percentage of land under cereal production steadily increased for Pakistan while it fluctuated for Iran and Iraq and dipped slightly for Afghanistan.

The percentage of arable land remained relatively constant for all four countries. Thus, we will not use it as a predictor variable in our model as it won't have a significant relationship with the size in refugee population.

Main Effects Linear Regression Model with Log Transformation and Assessing Quality of Fit

In order to analyze our data, we will construct a multiple linear regression model with the amount of land under cereal production, methane emissions, nitrous emissions, and CO2 emissions (all measured in kilotons per capita) as predictor variables and refugees as a proportion of the overall population as the response variable. The model was constructed using data from Afghanistan, Iraq, Iran, and Pakistan. Constructing this model will allow us to investigate the existence and nature of any relationship between the climate change indicators and the refugee flow. It will also allow us to discern which climate change indicators have the most significant impact on refugee flow from a country. Since the distribution of the response variable, proportion of refugees, is skewed, we used a logarithmic transformation in order to more accurately construct a linear model.

```
## # A tibble: 5 x 7
##   term            estimate   std.error statistic  p.value    conf.low  conf.high
##   <chr>              <dbl>       <dbl>     <dbl>    <dbl>       <dbl>      <dbl>
## 1 (Intercept)   0.700      0.389           1.80 7.42e- 2 -0.0698        1.47e+0
## 2 land_cereal  -0.00000158 0.000000163    -9.69 2.75e-16 -0.00000191   -1.26e-6
## 3 co2_tons     -0.372      0.150          -2.49 1.45e- 2 -0.669        -7.55e-2
## 4 nitrous_tons  0.000450   0.0000951       4.73 6.93e- 6  0.000261      6.39e-4
## 5 methane_tons -0.0000545  0.0000329      -1.66 1.00e- 1 -0.000120      1.07e-5
```

```
## # A tibble: 111 x 10
##    `log(prop_refug~ land_cereal co2_tons nitrous_tons methane_tons .fitted
##               <dbl>       <dbl>    <dbl>        <dbl>        <dbl>   <dbl>
##  1           -10.0      2304070    3.29         3459.       21325.   -3.78
##  2            -3.29     3183500    0.167        3215.        1138.   -3.02
##  3            -6.06     1928900    4.00         3988.       26675.   -3.50
##  4            -2.04     3018500    0.132        3275.        1394.   -2.73
##  5           -13.6      8044580    3.13        12372.       17564.   -8.60
##  6            -6.08     2290625    3.34         4017.       21465.   -3.53
##  7            -1.22     2909000    0.151        3257.        1345.   -2.57
##  8           -10.8      9099999    2.81        13531.       15495.   -9.52
##  9            -5.35     2256525    2.29         3392.        9842.   -2.74
## 10            -1.05     2799500    0.163        3327.        1336.   -2.37
## # ... with 101 more rows, and 4 more variables: .std.resid <dbl>, .hat <dbl>,
## #   .sigma <dbl>, .cooksd <dbl>
```

```
## [1] 0.7998283
```

```
##
## Call:
## lm(formula = log(prop_refugee) ~ land_cereal + co2_tons + nitrous_tons +
##     methane_tons, data = climate_filtered)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q     Max
## -6.2600 -0.8001  0.2136  1.1190  2.7183
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.005e-01  3.885e-01   1.803   0.0742 .
## land_cereal  -1.585e-06  1.635e-07  -9.693 2.75e-16 ***
## co2_tons     -3.723e-01  1.497e-01  -2.486   0.0145 *
## nitrous_tons  4.500e-04  9.512e-05   4.731 6.93e-06 ***
## methane_tons -5.452e-05  3.289e-05  -1.658   0.1003
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 106 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7923
## F-statistic: 105.9 on 4 and 106 DF,  p-value: < 2.2e-16
```

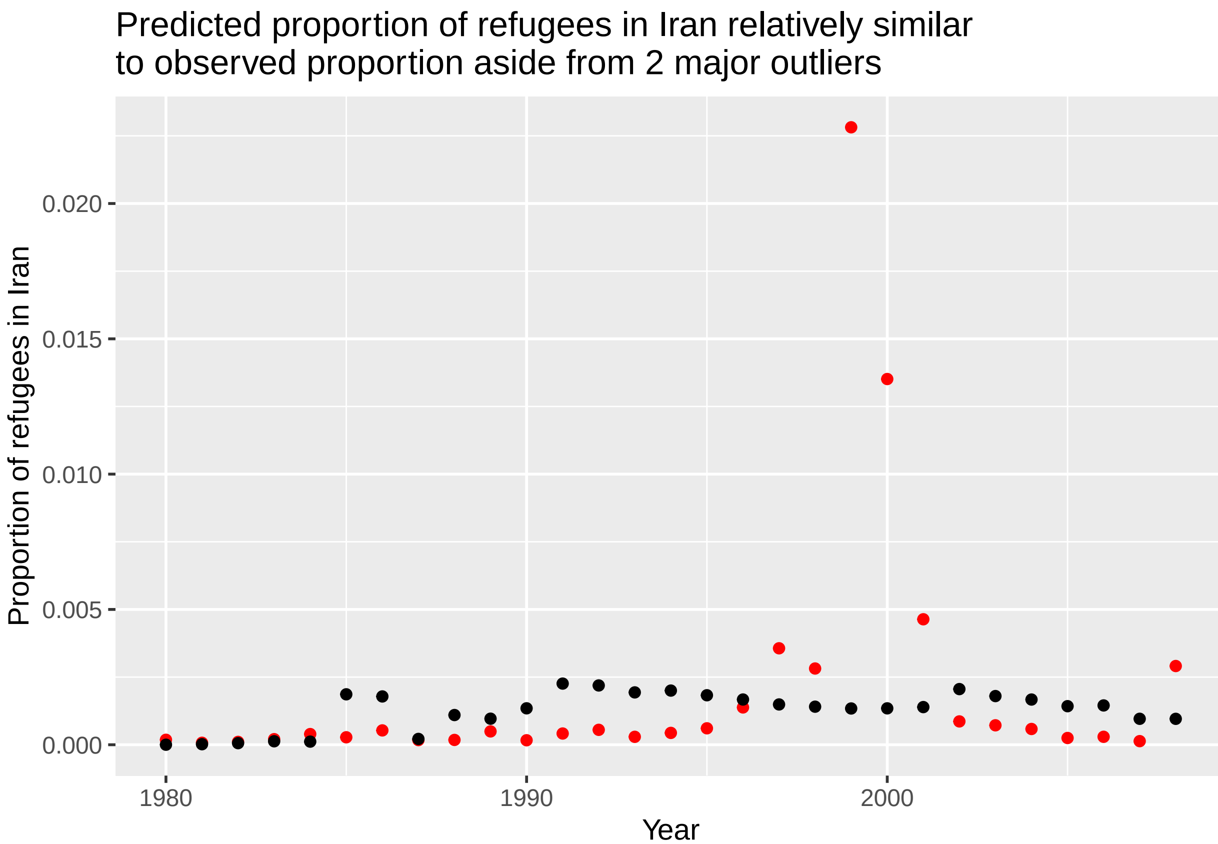
```
## # A tibble: 5 x 2
##   term         estimate
##   <chr>           <dbl>
## 1 (Intercept)     2.01
## 2 land_cereal     1.00
## 3 co2_tons        0.689
## 4 nitrous_tons    1.00
## 5 methane_tons    1.00
```

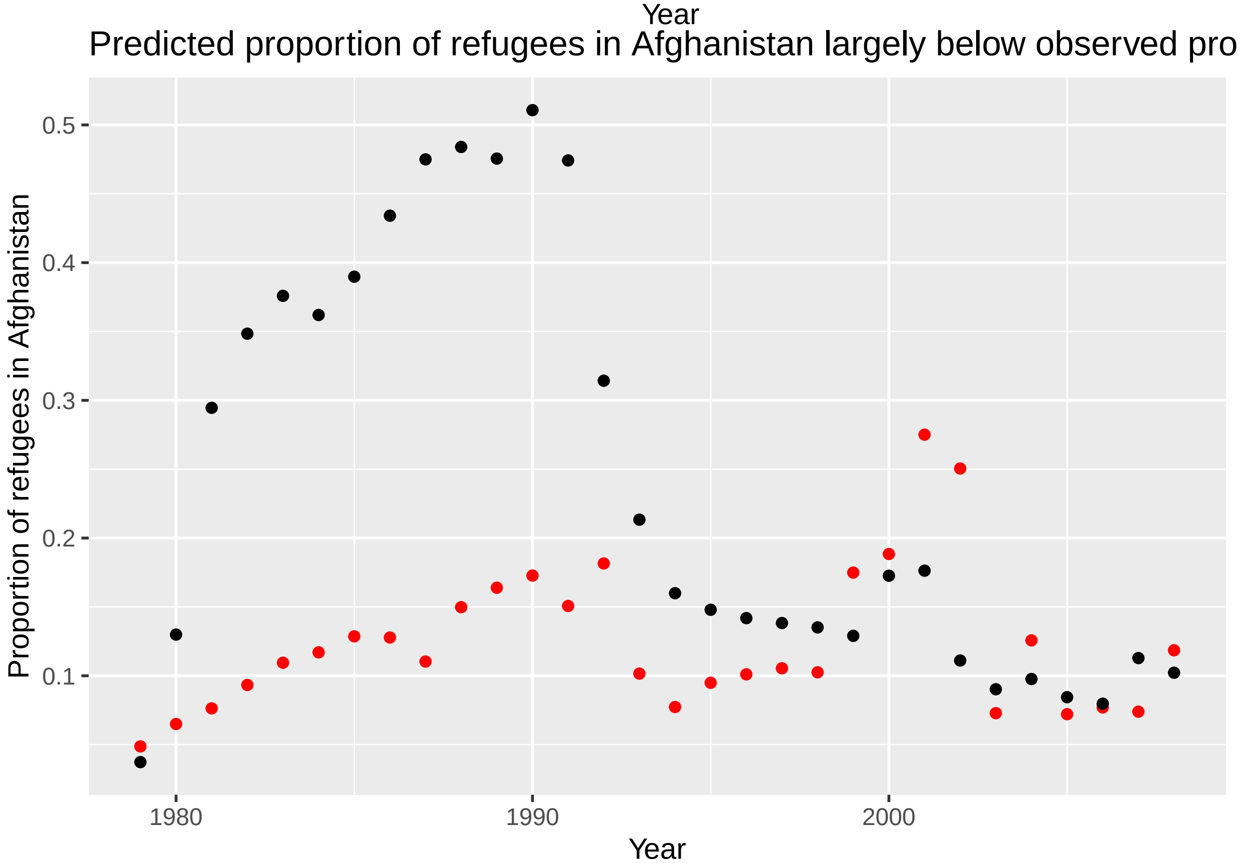
The model that we constructed can be seen here:

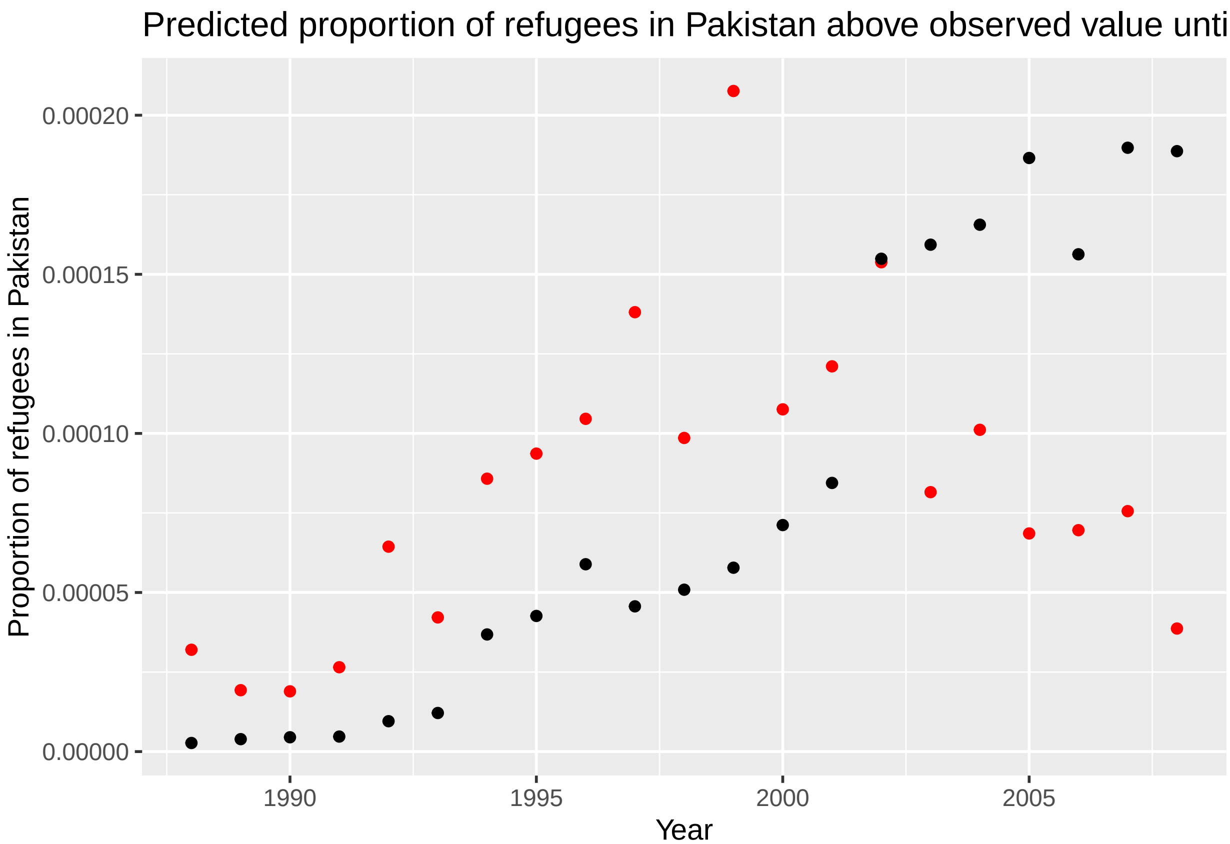
Predicted proportion of refugees = e^(7.004603e-01 + -1.584535e-06 * land_cereal + -3.723390e-01 * co2_tons + 4.499857e-04 * nitrous_tons + -5.452494e-05 * methane_tons)

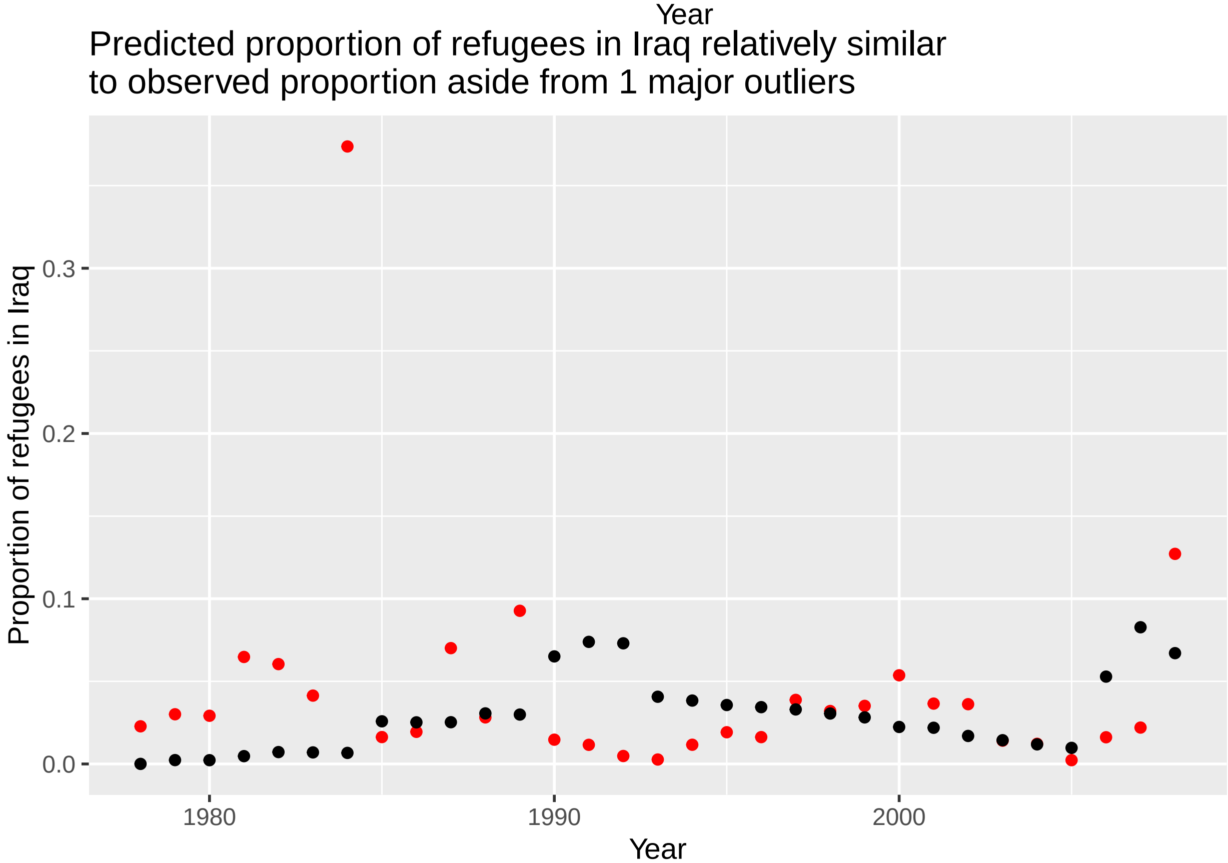
We obtained the following 95% of confidence intervals for the slope corresponding to each predictor variable conditional on all other predictors in our model:

Land under cereal production: (0.9999981, 0.9999987), CO2 emissions: (0.5121042, 0.9273253), N20 emissions: (1.0002614, 1.0006388), Methane emissions: (0.9998803, 1.0000107)


Predicted proportion of refugees in Iran relatively similar
to observed proportion aside from 2 major outliers
Proportion of refugees in Iran
0.020
0.015
0.010
0.005
0.000
1980
1990
2000
Year

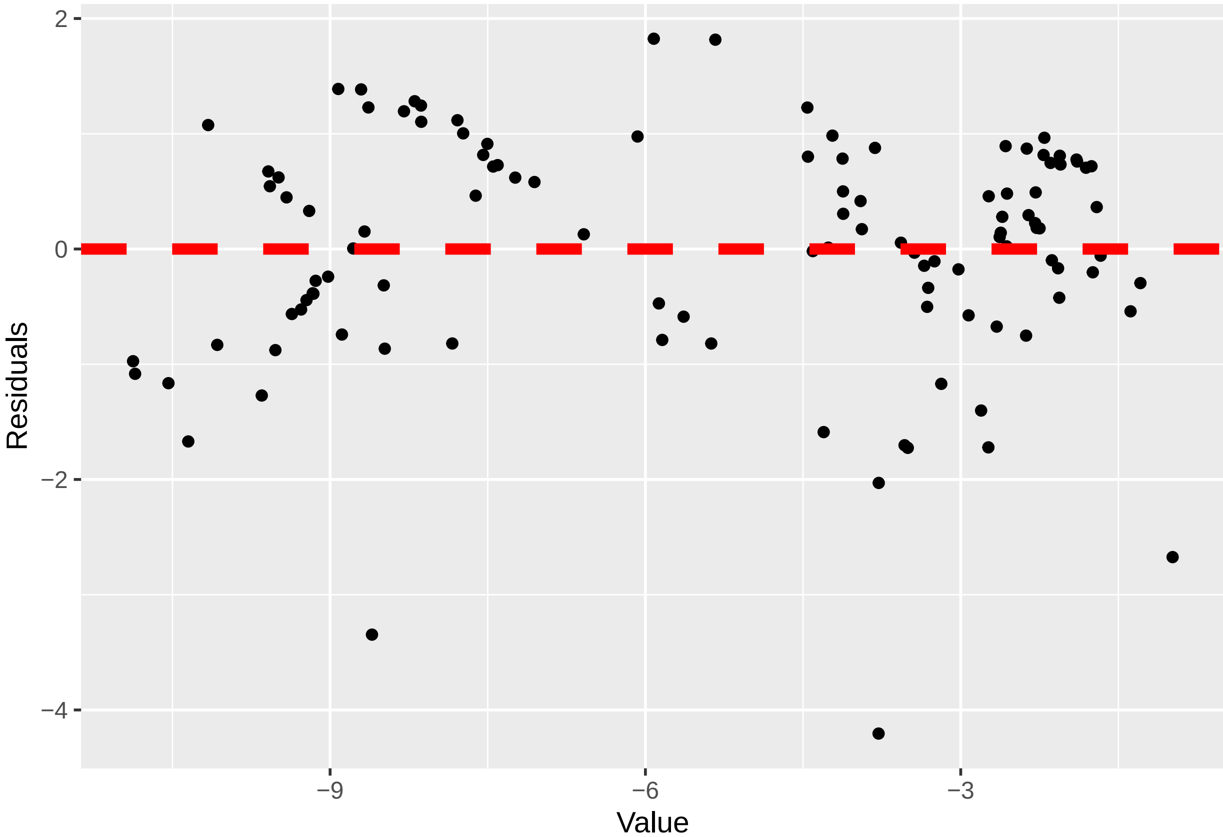

Predicted proportion of refugees in Afghanistan largely below observed pro
Proportion of refugees in Afghanistan
0.5
0.4
0.3
0.2
0.1
1980
1990
2000
Year


Predicted proportion of refugees in Pakistan above observed value until
0.00020
0.00015
0.00010
0.00005
0.00000
Proportion of refugees in Pakistan
1990
1995
2000
2005
Year

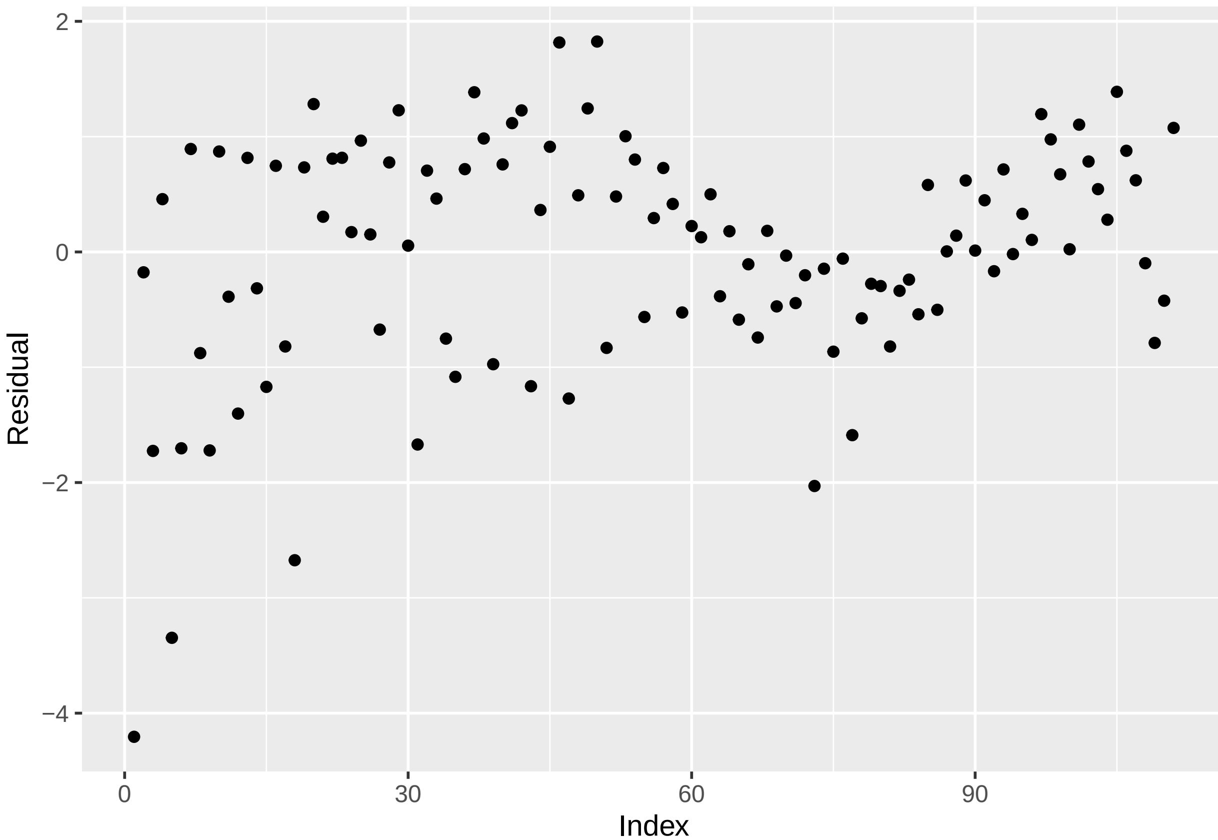

Predicted proportion of refugees in Iraq relatively similar
to observed proportion aside from 1 major outliers
0.3
0.2
0.1
0.0
Proportion of refugees in Iraq
1980
1990
2000
Year

Diagnostic Plots for Main Effects Linear Regression Model with Log Transformation

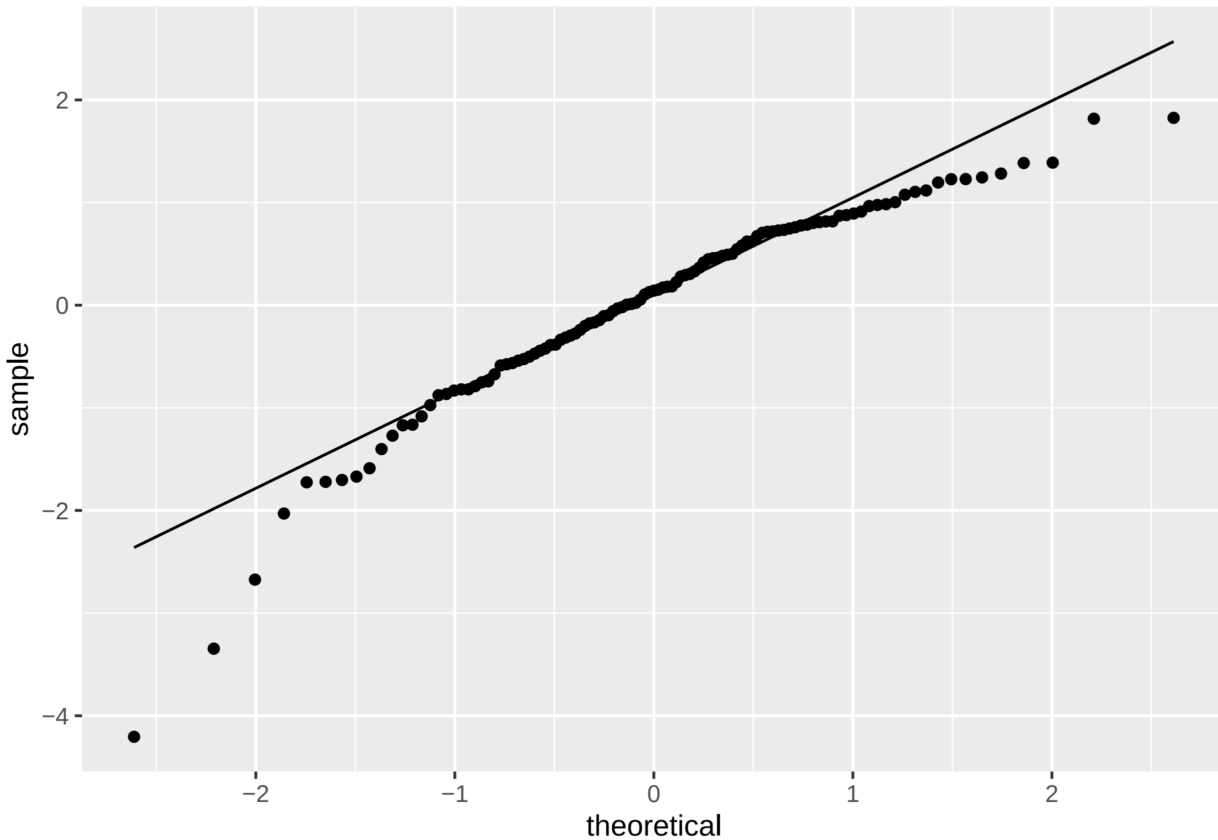
Since the dataset includes many different predictor variables, it was first necessary to understand which would be the most powerful holding the others constant. In order to make inferences in regression, certain conditions must be met. The first two conditions that must be met are linearity which means the relationship between the variables and the predictors be linear and equal variance which means that residuals have relatively constant variance. These conditions were checked by creating a basic linear model and plotting the residuals such that we could examine their variance and linearity.

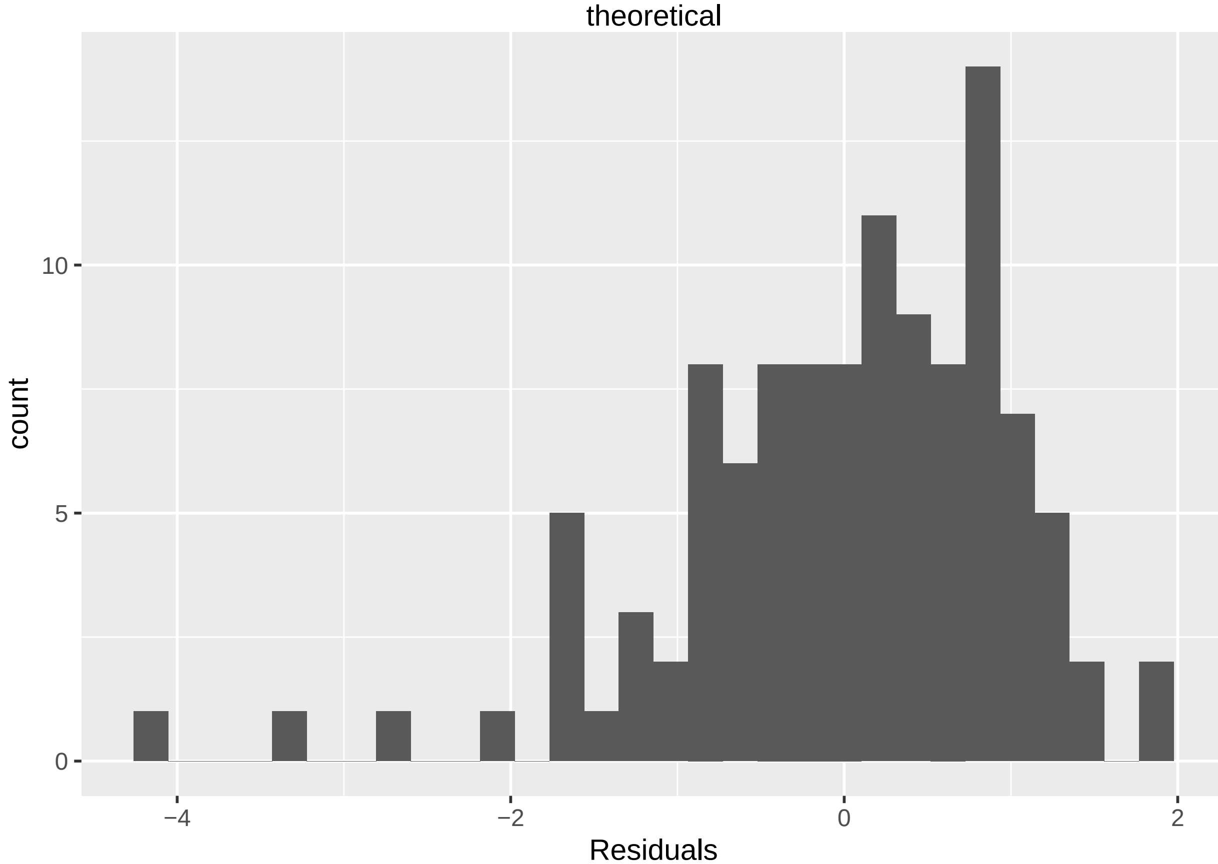


The next condition is independence which requires that the residuals be independent. We will evaluate independence by simply plotting all the residuals on a graph.



The next requirement is that of normality. The checks for this involved both a plotting of the residuals on a histogram as well as the creation of a Q-Q plot. The histogram was examined for a potential relationship between the residuals and the fit of the model to the qq line was checked too.


sample
theoretical
2
0
−2
−4
−2
−1
0
1
2


count
Residuals
10
5
0
−4
−2
0
2

Multiple Linear Regression Model with Log Transformation and Interactions

Next, we constructed a linear regression model with a logarithmic transformation and with interactions between the three greenhouse gas emissions to better understand the relationship between the predictor variables.

```
## # A tibble: 9 x 7
##   term              estimate  std.error statistic  p.value conf.low  conf.high
##   <chr>                <dbl>      <dbl>     <dbl>    <dbl>    <dbl>      <dbl>
## 1 (Intercept)        2.05e+0    3.76e-1     5.44  3.73e- 7  1.30e+0    2.79e+0
## 2 land_cereal       -1.85e-6    2.62e-7    -7.05  2.17e-10 -2.37e-6   -1.33e-6
## 3 co2_tons          -1.06e-1    2.42e-1    -0.439 6.62e- 1 -5.87e-1    3.74e-1
## 4 nitrous_tons       4.09e-4    1.61e-4     2.53  1.29e- 2  8.83e-5    7.29e-4
## 5 methane_tons      -2.80e-4    5.27e-5    -5.31  6.44e- 7 -3.84e-4   -1.75e-4
## 6 co2_tons:nitro~   -9.92e-5    3.69e-5    -2.69  8.35e- 3 -1.72e-4   -2.60e-5
## 7 co2_tons:metha~   -5.25e-6    9.70e-6    -0.541 5.90e- 1 -2.45e-5    1.40e-5
## 8 nitrous_tons:m~    1.19e-8    1.87e-9     6.38  5.37e- 9  8.20e-9    1.56e-8
## 9 land_cereal:co~    2.57e-7    6.79e-8     3.78  2.67e- 4  1.22e-7    3.91e-7
```

```
## # A tibble: 111 x 11
##    .rownames `log(prop_refug~ land_cereal co2_tons nitrous_tons methane_tons
##    <chr>                <dbl>       <dbl>    <dbl>        <dbl>        <dbl>
##  1 75                  -10.0      2304070    3.29          3459.       21325.
##  2 77                   -3.29     3183500    0.167         3215.        1138.
##  3 79                   -6.06     1928900    4.00          3988.       26675.
##  4 81                   -2.04     3018500    0.132         3275.        1394.
##  5 82                  -13.6      8044580    3.13         12372.       17564.
##  6 83                   -6.08     2290625    3.34          4017.       21465.
##  7 85                   -1.22     2909000    0.151         3257.        1345.
##  8 86                  -10.8      9099999    2.81         13531.       15495.
##  9 87                   -5.35     2256525    2.29          3392.        9842.
## 10 89                   -1.05     2799500    0.163         3327.        1336.
## # ... with 101 more rows, and 5 more variables: .fitted <dbl>,
## #   .std.resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>
```

```
## [1] 0.8942522
```

```
##
## Call:
## lm(formula = log(prop_refugee) ~ land_cereal + co2_tons + nitrous_tons +
##     methane_tons + co2_tons * nitrous_tons + co2_tons * methane_tons +
##     nitrous_tons * methane_tons + co2_tons * land_cereal, data = climate_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5013 -0.4023  0.0801  0.6416  2.1880
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.045e+00  3.762e-01   5.436 3.73e-07 ***
## land_cereal              -1.848e-06  2.620e-07  -7.051 2.17e-10 ***
## co2_tons                 -1.064e-01  2.424e-01  -0.439 0.661722
## nitrous_tons              4.086e-04  1.615e-04   2.531 0.012917 *
```

```
## methane_tons              -2.799e-04  5.270e-05  -5.310 6.44e-07 ***
## co2_tons:nitrous_tons     -9.918e-05  3.687e-05  -2.690 0.008352 **
## co2_tons:methane_tons     -5.250e-06  9.701e-06  -0.541 0.589602
## nitrous_tons:methane_tons  1.190e-08  1.866e-09   6.376 5.37e-09 ***
## land_cereal:co2_tons       2.565e-07  6.792e-08   3.777 0.000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.135 on 102 degrees of freedom
##   (133 observations deleted due to missingness)
## Multiple R-squared:  0.8943, Adjusted R-squared:  0.886
## F-statistic: 107.8 on 8 and 102 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 9 x 2
##   term                      estimate
##   <chr>                        <dbl>
## 1 (Intercept)                  7.73
## 2 land_cereal                  1
## 3 co2_tons                     0.899
## 4 nitrous_tons                 1
## 5 methane_tons                 1
## 6 co2_tons:nitrous_tons        1
## 7 co2_tons:methane_tons        1
## 8 nitrous_tons:methane_tons    1
## 9 land_cereal:co2_tons         1
```

The model that we constructed can be seen here:

Predicted proportion of refugees = e^(2.045090 - 1.847728e-06 * land_cereal - 1.063508e-01 * co2_tons + 4.085926e-04 * nitrous_tons - 2.798573e-04 * methane_tons - 9.918373e-05 (co2_tons * nitrous_tons) - 5.249514e-06 (co2_tons * methane_tons) + 1.189883e-08 (nitrous_tons * methane_tons) * 2.565460e-07 (land_cereal * co2_tons))

We obtained the following 95% of confidence intervals for the slope corresponding to each predictor conditional on all other predictors in our model:

Land under cereal production: (0.9999981, 0.9999987), CO2 emissions: (0.5121042, 0.9273253), N20 emissions: (1.0002614, 1.0006388), Methane emissions: (0.9998803, 1.0000107)

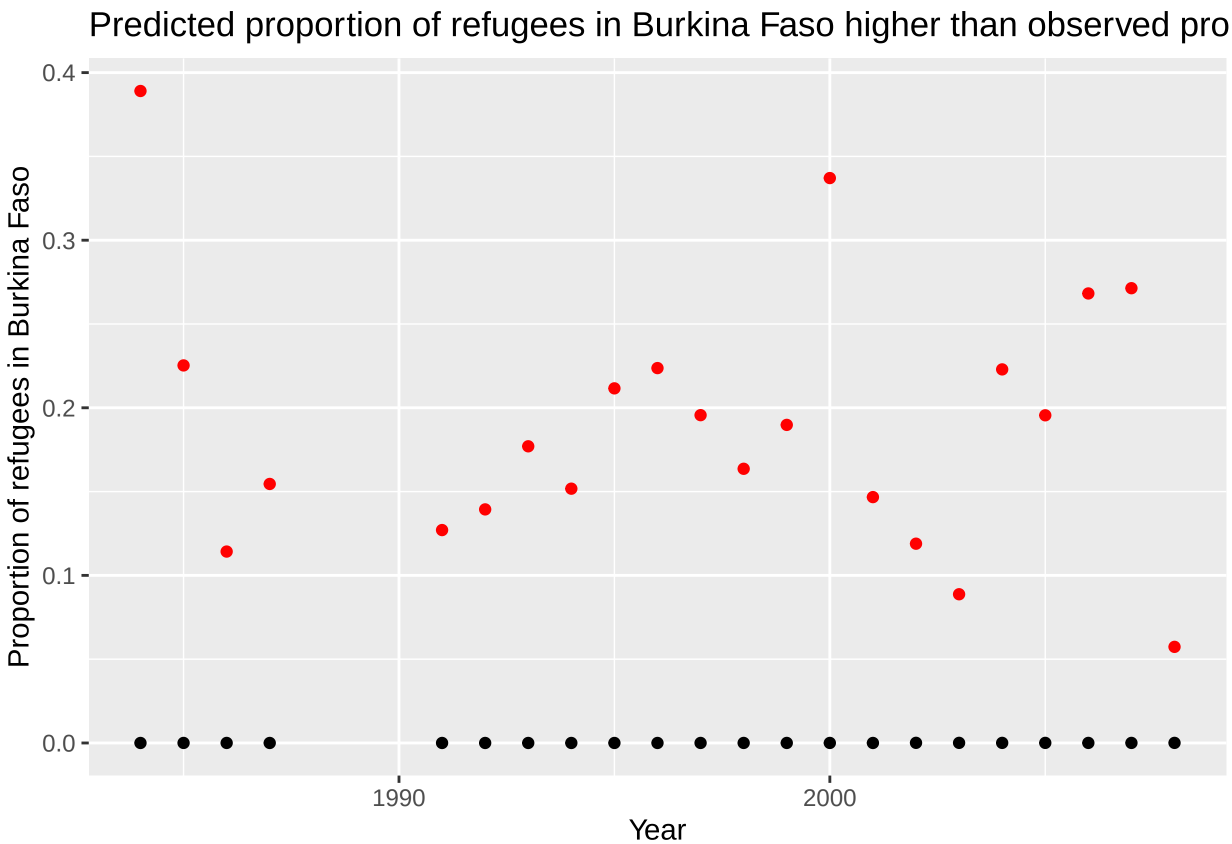
Testing for Multicollinearity

To further analyze the relationship between the predictor variables, we tested for mutlicollinearity. This was to determine if two or more explanatory variables in our multiple regression model were highly linearly related.

```
##  land_cereal     co2_tons nitrous_tons methane_tons
##    20.132819     4.343368    41.210613    16.117866
```

Applying the Model to Another Country

To further evaluate our model, we applied it to another at risk region as identified by the Ecological Threat Register, the Sahel Belt of Africa. The Sahel Belt includes Senegal, Mauritania, Mali, Burkina Faso, Niger, Chad, Sudan, Eritrea. These countries have similar conflict levels and similar economic situations thus lending it to another application of our model.



RESULTS

Main Effects Linear Regression Model with Log Transformation and Assessing Quality of Fit

Our fitted main-effects multiple linear regression model with a logarithmic transformation is as follows:

Predicted proportion of refugees = e^(7.004603e-01 + -1.584535e-06 * land_cereal + -3.723390e-01 * co2_tons + 4.499857e-04 * nitrous_tons + -5.452494e-05 * methane_tons))

At the alpha = 0.05 significance level, the fitted log coefficients for each of our predictors in the log transformation relative to the baseline category of standard refugee rate were statistically significant except for methane emissions. Therefore, there is sufficient evidence to suggest that the true log coefficients corresponding to the predictor variables hectares of land for cereal production, metric tons of CO2 produced per capita, and kilotons of nitrous oxide produced compared with the log of the proportion of refugees in the country is not equal to 0. There is some relationship between these predictor variables and change in the proportion of refugees in the countries studied.

We are particularly interested in testing to see whether there is a log relationship between carbon dioxide emissions and the proportion of refugees in a population.

H0: There is no relationship between carbon emissions per capita and the log proportion of refugees. The coefficient in the log transformation is 0.

H1: There is a relationship between carbon emission per capita and the log proportion of refugees. The coefficient in the log transformation is not 0.

Under the null hypothesis, our test statistic follows a standard normal distribution. The value of our test statistic is equal to approximately -2.486452, which corresponds to a p-value of 1.446306e-02. Thus, at the $\alpha = 0.05$, we reject our null hypothesis; we have sufficient evidence to suggest that there is a relationship between the metric tons of CO2 produced per capita and the log of the proportion of refugees. The estimated log transformation coefficient was -3.723390e-01 for CO2 emissions. Therefore, we are 95% confidence that the proportion of refugees to decreases by a factor of 0.5121042 - 0.9273253, holding all other variables constant. The 95% confidence intervals for nitrous oxide emissions and methane emissions respectively are (1.0002614 - 1.0006388) and (0.9998803 - 1.0000107), which suggests that compared carbon dioxide emissions, nitrous and methane emissions have a lesser impact on the proportion of refugees in a population.

The adjusted coefficient of determination (r squared) for the linear regression model was 0.792. This means 79.2% of the variability in the proportion of refugees can be explained by the model. The adjusted coefficient of determination was used since our model uses multiple predictors.

Diagnostic Plots for Main Effects Linear Regression Model with Log Transformation

Independence: The residuals do not exhibit any pattern or abnormal clusters.

Linearity: There is an approximately symmetric distribution above and below the y=0 line.

Equal Variance: The vertical variation or spread of the residuals is not approximately equal. There is less variance at the extremes of the plot indicating that there is not equal variance.

Normality: The residuals appear to have a normal distribution.

Multiple Linear Regression Model with Log Transformation and Interactions

We also constructed a linear model with interactions between each of the predictor variables and a logarithmic transformation. Each interaction tested (CO2 emissions and N2O emissions, CO2 emissions and methane emissions, N2O emissions and methane emissions, and land under cereal production and CO2 emissions) in addition to nitrous oxide gas emissions and methane emissions, only increased the proportion of refugees by a factor of 1 holding all other variables constant. Thus, we found that this model had limited usefulness compared to the model that analyzed main effects.

Testing for Multicollinearity

The VIF scores of the variables output here indicate high degree of multicollinearity with several of the variables. Everything expect for CO2 emissions had a VIF greater than 5 which was our established level of concern. This was expected given that we were including variables that measured emissions. Something that is a source of emissions is likely not a source for only a single type of emissions meaning that as one type of emissions increases, the other kinds of emissions will also increase. Thus there was a great deal of correlation between our variables. That interaction between the variables made the point predictions in our test data from the Sahel very inaccurrate. Methods for resolving these problems are addressed in the limitations and concerns sections.

Applying Model to Another Country

DISCUSSION

Given the rapidly accelerating nature of climate change, it is imperative that more data analysis be done on the massive impacts it has. Our goal was to see what climate change factors like CO2 emissions and agricultural outputs would influence refugee flow from a country. Refugee flow was picked as a terminal measure of climate change impacts on a country. With so many people denying that climate change impacts their lives because it just makes some days hotter, we decided to show a very concrete and serious impact on the lives of millions of people. Using a linear model allowed us to see not only the direct effects of several different variables all at play at once, but also their interactions. In all of the countries we analyzed, climate change was not the only factor that influenced refugee flow. However, the ability to explain even 20% of the variance in something as intensely complicated as refugee flow with a linear model would be quite the feat. In this experiment we were actually able to obtain and adjusted R squared of ADJ R SQUARED HERE. GOOD result: This was surprising and could be due to LIMITATIONS, but the conditions for inference were mostly satisfied leading us to believe we had constructed a robust model. BAD result: While we were not able to construct a model that explained a significant amount of the variance in refugee flow, this was still a good exercise in trying to uncover relationships that could help the general understanding of migration.

Our variables were X X X X X. Some like X and X were chosen because they are very obvious markers of human impact on the climate. Additionally X and X were chose because they are downstream markers of anthropogenic climate change that are closer to affecting human lives. There was also an interaction effect between X and X measured for THIS REASON.

The significance of the main effects for X X X were Y Y Y respectively. These are great results indicating there likely is a relationship between these variables and refugee flow.

While we did not formally predict a direction for the relationship between our predictors and refugee flow, our general thought was that as markers of climate change worsened (i.e. emissions increased) the refugee flow from a country would increase. However, the opposite was true for some emissions markers, most clearly for CO2 emissions in kilotons per capita. This was likely the case because CO2 emissions are not just a marker of a worsening climate. They are also a marker of the economic development of a country. SOURCE HERE As almost every country in the world has industrialized, their CO2 emissions have increased. Additionally, while the impacts of climate change can take many years to accumulate to a disastrous level, the impacts of rapid economic development can be felt much faster. Those impacts are usually more job opportunities and decreases in poverty, both of which would make someone more likely to stay in the current country of residence. Overall, this data proves a general relationship all over the world which is that rapid economic development may make immediate living situations easier, but it will also have impacts on the environment such as increased emissions.

Overall, this not only served as an excellent exercise for us in applying linear regression to the real world, but also can act as the basis for future work trying to develop a quantitative prediction of refugee flow in the future.

Limitations, Concerns, and Future Investigation

The main concern with the validity of our models here was the lack of data. The World Bank data which was used was missing for several years and we had to develop our regional groupings due to lack of collation on the part of the World Bank. Moreover, robust, comprehensive data on refugee populations from each country doesn't begin until the late 1980s and early 1990s whereas other metrics such as carbon dioxide emissions begin in the 1960s. In addition to the lack of data, the data on refugee populations was a raw sum and didn't disaggregate the data based on reason for refugee status. Further studies on the relationship between climate predictors and displaced populations should look at refugees and displaced people due to climate change. Though the World Bank data includes data for internally displaced people due to disasters, this metric lacks enough data for analysis.

Were we to repeat our analysis, we would include and adjust for more potential confounders to improve the prediction accuracy of our model. The test for multicollinearity demonstrated high correlations between some of the predictor variables, especially related to greenhouse gas emissions. We can address multicollinearity by linearly combining the highly correlated independent variables. In context, having a variable for greenhouse gas emissions makes sense within the literature since greenhouse gas emissions are often measured together.