Project Draft

Due: October 30, 11:59pm

```
library(tidyverse)
library(broom)
climate <- read_csv("data/climate_data_final.csv")

#View(climate)

#climate2 <- data.frame(t(climate[-1]))
#colnames(climate2) <- climate[, -1]

#glimpse(climate2)
```

Introduction

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). Moreover, 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 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 indicators 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 socioeconomic factors, and the resulting displacement. In order to make the analysis more manageable, we will focus on the climate change and socioeconomic data of one region. Previous literature has indicated that variables such as rainfall, agricultural yield, and low-lying areas may be associated with internal displacement, and we would like to see if this region's data from the World Development Bank (WDB) are consistent with these claims.

According to the WDB, most of the data from the data set comes directly from each country in the World Bank Group's national statistical systems. The data itself contains many variables, and the series name tells us the metric for which we are getting data. Within each series, the data is broken down into the data for each nation, and each row below the country name corresponds with that country's data for that series name for each year between 2000-2020. This data set 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. The data set also includes points on conflict, fragility and, most importantly, displacement of people within given countries. Unfortunately, the

WDB does not have very complete data for some of the series name variables. However, we will focus mainly on variables that have sufficient data points, unless the variable is unlikely to change much over time, in which case we have decided to turn these into binary variables.

We will start by examining variables in the WDB data such as "Agriculture, forestry, and fishing, value added" (measured in %GDP and %annual growth) to see how important agriculture and these other variables are to the economy of the nation. Other variables include "Urban population living in areas where elevation is below 5 meters (% of total population)", which will tell us the percentage of people live in low-lying areas. We would also like to see if socioeconomic factors and poverty are better predictors of internal displacement not due to conflict. Then, we will analyze the correlation between displacement and statistics such as prevalence of food security in a nation, people using at least basic drinking water services, and GDP per capita. We hypothesize that the percentage of total population below 5 meters of elevation in island nations, the percentage of total population living in slums, poverty, and falling agricultural output will be the strongest correlates and predictors of displacement of people due to climate in following years in regions identified by The Ecological Threat Register.

Methodology

In order to analyze our data, we will run multiple linear regressions on internally displaced people associated with disasters and other variables that have been associated with displacement. We will look at regions of countries in the world in order to make the data analysis easier for us. We would like to see which variables are predictors of internal displacement. Seeing these predictors will allow us to easily see which variables are associated with high amounts of displacement due to disaster. Some of the variables that we will focus on include annual rainfall, agricultural yield, and the percentage of land that is low-lying in the country. For some of the variables with fewer data points but are unlikely to change over the past 20 years, such as the percentage of land that is low-lying, we will turn these variables into binary categorical variables. This will allow us to use these variables in our data analysis.

Once we identify some of the better predictors of displacement, we would like to run two sample hypothesis tests to see if there are differences in certain variables in countries that are prone to displacement. Specifically, we could run two sample hypothesis tests between high- and low-displacement regions as mentioned in the introduction with relation to certain predictor variables. This way, we could see the extent to how different these variables are for these two classifications. Additionally, we would like to run a Chi-Squared test on the displacement due to disasters and displacement due to conflict. We would like to run this test because it can help us elucidate whether these displacements are related. We would run this test because, if there is a high amount of displacement due to conflict, this could affect the displacement due to disaster, so we would want to see if these variables are independent.

```
climate %>%
  select_at(vars(contains("_10")))
```

```
## # A tibble: 22 x 14
##    `CO2 emissions ~ `Access to clea~ `Agricultural m~ `Average precip~
##    <chr>            <chr>            <chr>           <chr>
##  1 Pakistan         Pakistan         Pakistan        Pakistan
##  2 0.747833829      22.62            65.70273771     ..
##  3 0.74177304       24.14            64.4908243      ..
##  4 0.762850323      25.39            64.39705163     494
##  5 0.776618469      26.63            63.6938039      ..
##  6 0.840018833      28.12            62.9855068      ..
##  7 0.852356025      29.23            62.73491724     ..
##  8 0.890578012      30.5             62.70292513     ..
##  9 0.946883805      31.81            62.35107571     494
## 10 0.926721426      33.05            62.34325407     ..
## # ... with 12 more rows, and 10 more variables: `Cereal production (metric
## #   tons)_10` <chr>, `Crop production index (2004-2006 = 100)_10` <chr>,
```

```
## #   `Droughts, floods, extreme temperatures (% of population, average
## #   1990-2009)_10` <chr>, `Fossil fuel energy consumption (% of
## #   total)_10` <chr>, `PM2.5 air pollution, population exposed to levels
## #   exceeding WHO guideline value (% of total)_10` <chr>, `Population density
## #   (people per sq. km of land area)_10` <chr>, `Population living in areas
## #   where elevation is below 5 meters (% of total population)_10` <chr>, `Net
## #   migration_10` <chr>, `Refugee population by country or territory of
## #   origin_10` <chr>, `Internally displaced persons, new displacement
## #   associated with disasters (number of cases)_10` <chr>
```

```
#Mutate country = "pakistan"
#Save into new file; repeat process; rename the variables to be the same so that
#r-bind can occur
```

```
climate_pakistan <- climate %>%
  filter(`Fossil fuel energy consumption (% of total)_10`!= "..") %>%
  filter(`CO2 emissions (metric tons per capita)_10` != "..")%>%
  filter(`Cereal production (metric tons)_10` != "..")%>%
  filter(`Population density (people per sq. km of land area)_10` != "..")%>%
  filter(`Internally displaced persons, new displacement associated with disasters (number of cases)_10
  filter(`Refugee population by country or territory of origin_10` != "..")%>%
  mutate(fossil_fuel_10 = as.numeric(`Fossil fuel energy consumption (% of total)_10`),
         CO2_emissions_10 = as.numeric(`CO2 emissions (metric tons per capita)_10`),
         cereal_production_10 = as.numeric(`Cereal production (metric tons)_10`),
         population_density_10 = as.numeric(`Population density (people per sq. km of land area)_10`),
         internal_displacement_10 = as.numeric(`Internally displaced persons, new displacement associat
         refugee_10 = as.numeric(`Refugee population by country or territory of origin_10`))%>%

  select(fossil_fuel_10, refugee_10, CO2_emissions_10, cereal_production_10, population_density_10, int
  slice(2:8)
```

```
## Warning: Problem with `mutate()` input `fossil_fuel_10`.
## x NAs introduced by coercion
## i Input `fossil_fuel_10` is `as.numeric(`Fossil fuel energy consumption (% of total)_10`)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## Warning: Problem with `mutate()` input `CO2_emissions_10`.
## x NAs introduced by coercion
## i Input `CO2_emissions_10` is `as.numeric(`CO2 emissions (metric tons per capita)_10`)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## Warning: Problem with `mutate()` input `cereal_production_10`.
## x NAs introduced by coercion
## i Input `cereal_production_10` is `as.numeric(`Cereal production (metric tons)_10`)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## Warning: Problem with `mutate()` input `population_density_10`.
## x NAs introduced by coercion
## i Input `population_density_10` is `as.numeric(`Population density (people per sq. km of land area)_
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## Warning: Problem with `mutate()` input `internal_displacement_10`.
## x NAs introduced by coercion
## i Input `internal_displacement_10` is `as.numeric(`Internally displaced persons, new displacement as
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
## Warning: Problem with `mutate()` input `refugee_10`.
## x NAs introduced by coercion
## i Input `refugee_10` is `as.numeric(`Refugee population by country or territory of origin_10`)`.
```

```
## Warning in mask$eval_all_mutate(dots[[i]]): NAs introduced by coercion
```

```
climate_model <- lm(refugee_10 ~ population_density_10 + fossil_fuel_10 +
                      CO2_emissions_10 + cereal_production_10,
                    data = climate_pakistan)
tidy(climate_model)
```

```
## # A tibble: 5 x 5
##   term                   estimate    std.error statistic p.value
##   <chr>                     <dbl>        <dbl>     <dbl>   <dbl>
## 1 (Intercept)           -8.02e+6 864069.          -9.28   0.0114
## 2 population_density_10  9.45e+3   1534.           6.16   0.0253
## 3 fossil_fuel_10         8.64e+4  11547.           7.48   0.0174
## 4 CO2_emissions_10       7.26e+5 657015.           1.11   0.384
## 5 cereal_production_10  -9.84e-4      0.00617     -0.159  0.888
```

```
climate_model_aug <- augment(climate_model)
climate_model_aug
```

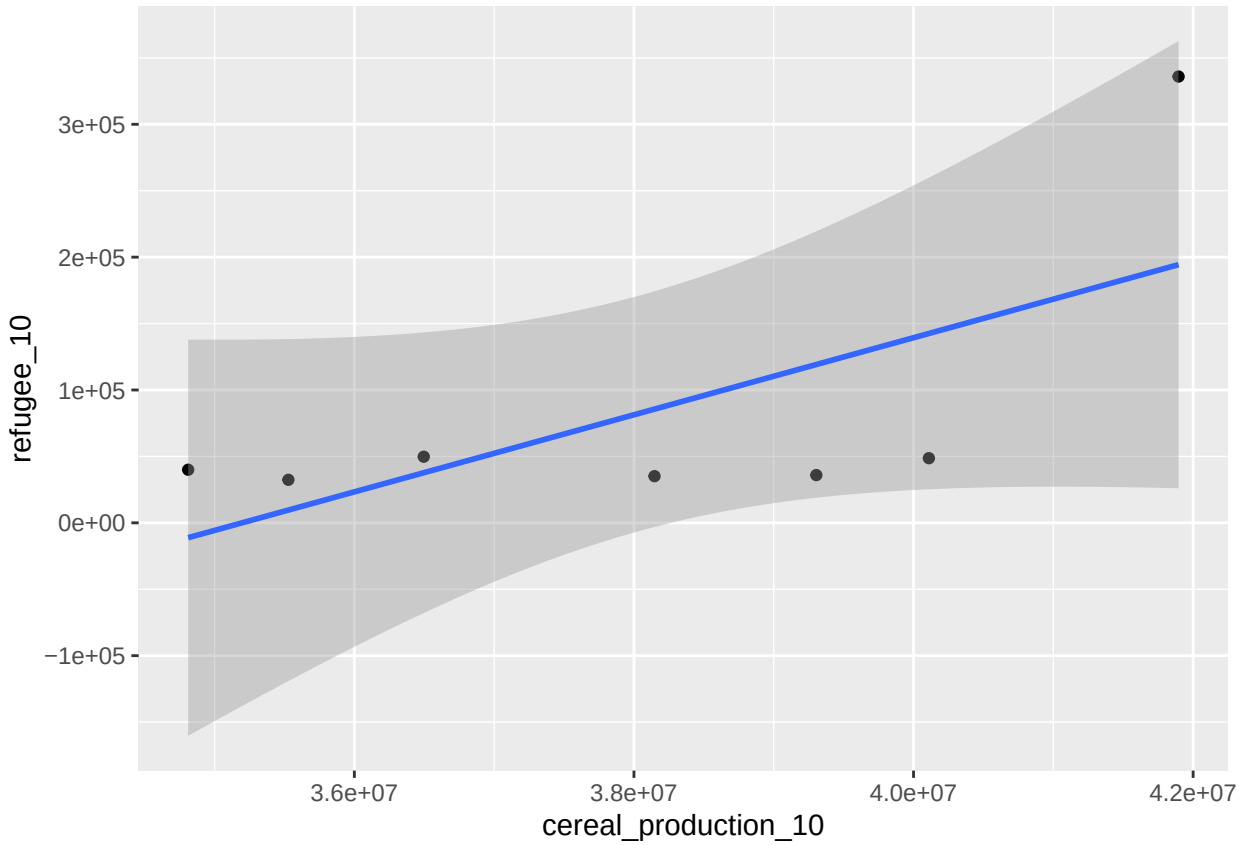
```
## # A tibble: 7 x 11
##   refugee_10 population_dens~ fossil_fuel_10 CO2_emissions_10 cereal_producti~
##        <dbl>            <dbl>          <dbl>            <dbl>            <dbl>
## 1      32391             223.           61.6            0.927         35528100
## 2      35129             228.           61.0            0.905         38147050
## 3      39979             233.           60.3            0.945         34811258
## 4      35948             238.           59.7            0.936         39304580
## 5      49778             243.           59.5            0.908         36496350
## 6      48678             248.           59.3            0.914         40109711
## 7     335947             253.           61.6            0.934         41895811
## # ... with 6 more variables: .fitted <dbl>, .resid <dbl>, .std.resid <dbl>,
## #   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>
```

```
glance(climate_model)%>%
  pull(r.squared)
```

```
## [1] 0.9858425
```

```
ggplot(climate_pakistan, mapping = aes(x = cereal_production_10, y=refugee_10)) +
  geom_point()+
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

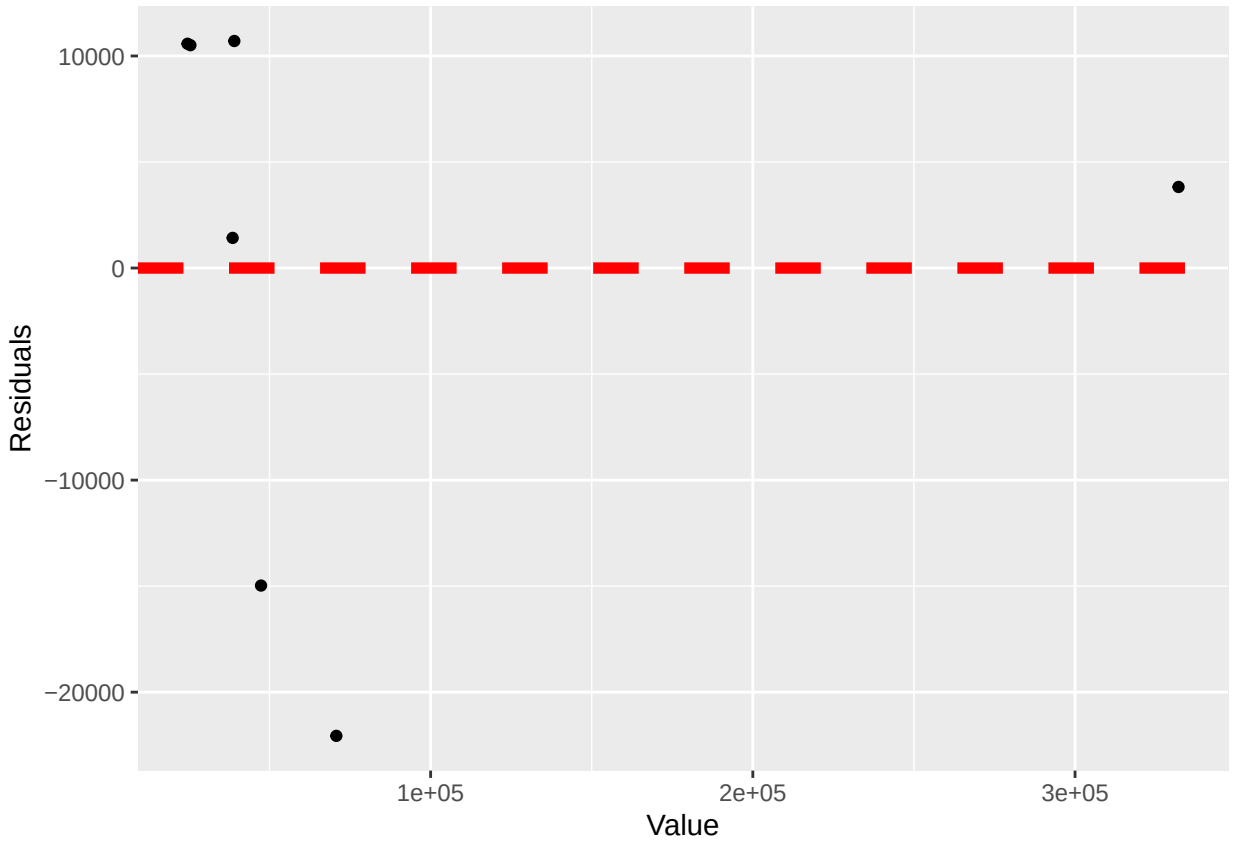



```
  labs(x = "Residuals")
```

```
## $x
## [1] "Residuals"
##
## attr(,"class")
## [1] "labels"
```

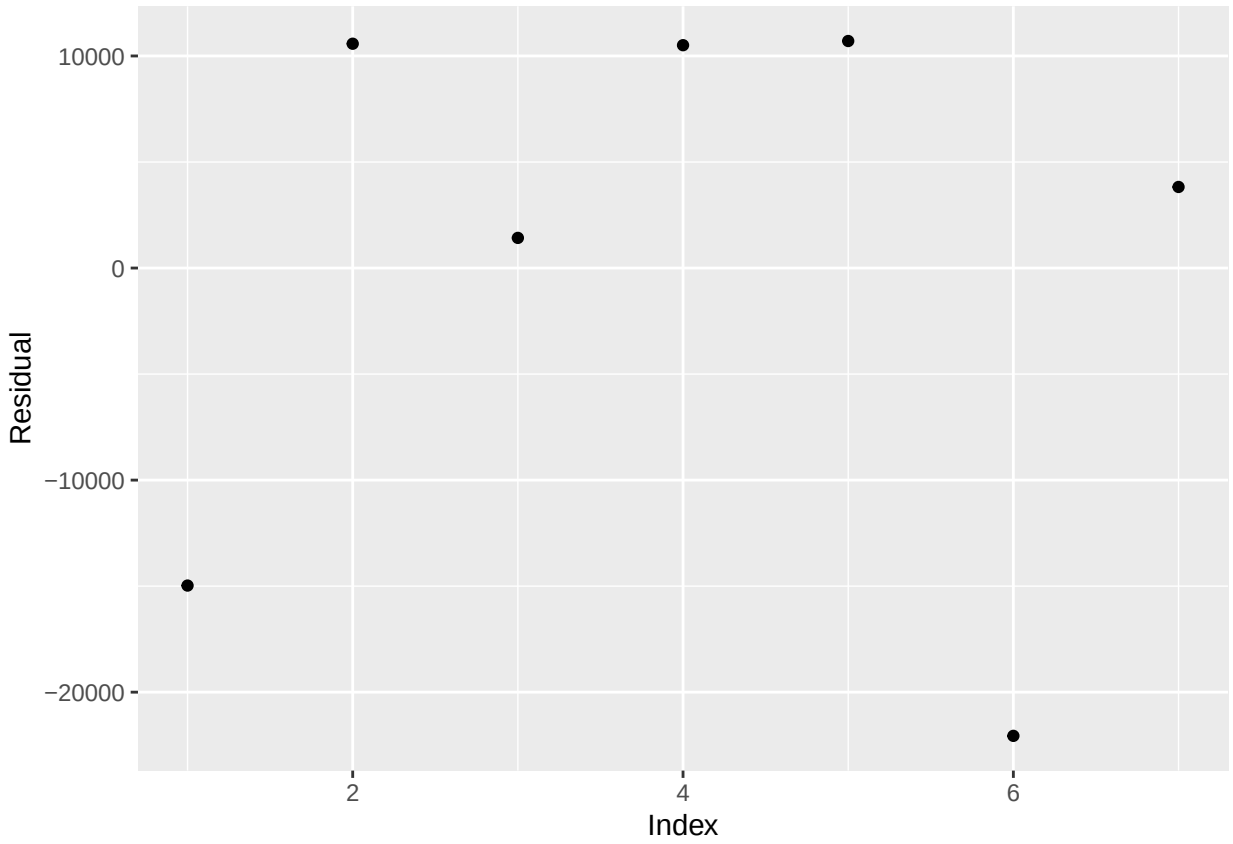
Since the dataset includes many different predictor variables, it was first necessary to understand which would be the most powerful holding the others constant. In order to make inferences in regression, certain conditions must be met. In order to simplify the process, a simple model with one predictor and one response was created. The first two are linearity which requires that the relationship between the variables and the predictor 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.

```
ggplot(climate_model_aug, mapping = aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, lwd = 2, col = "red", lty = 2) +
  labs(x = "Value", y = "Residuals")
```



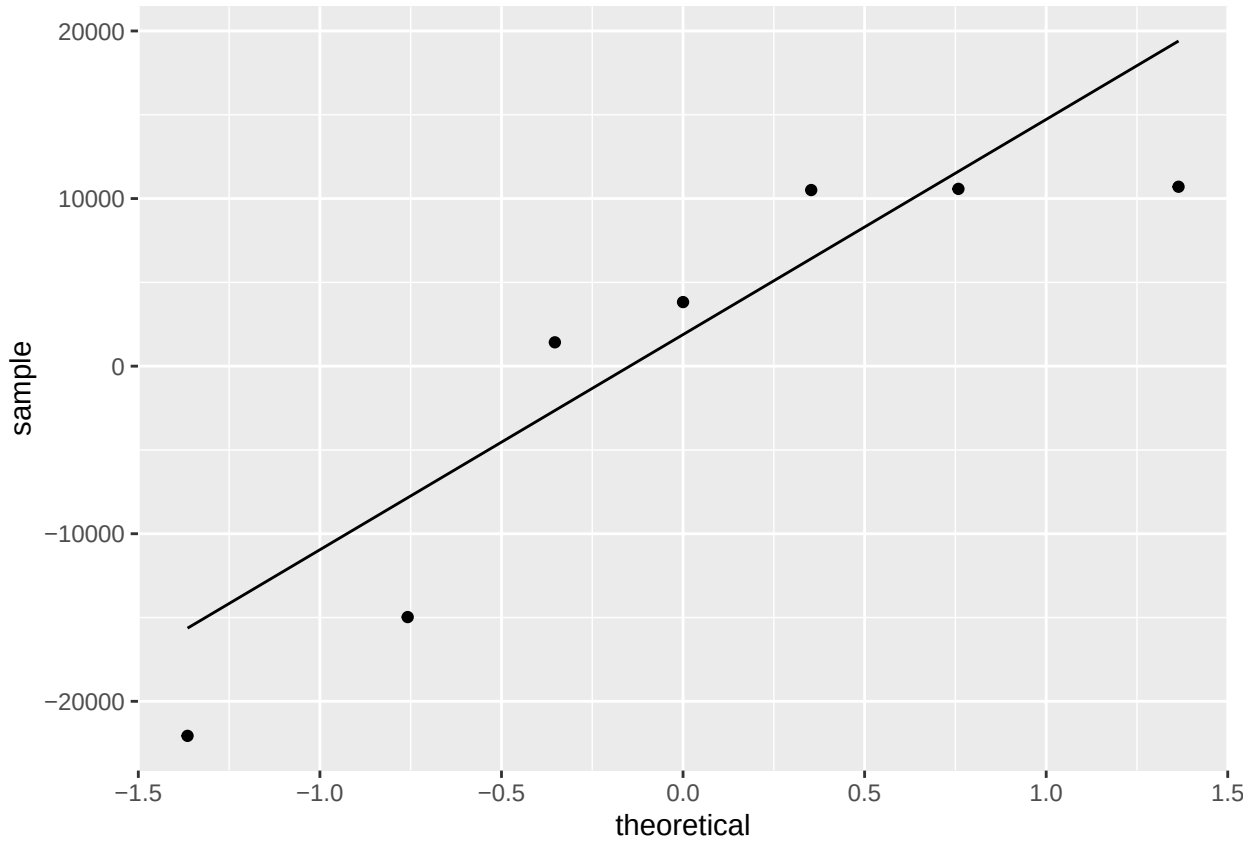
The next is independence which requires that the residuals be independent. This is checked very easily by simply plotting all the residuals on a graph.

```
ggplot(data = climate_model_aug,
       aes(x = 1:nrow(climate_pakistan),
           y = .resid)) +
  geom_point() +
  labs(x = "Index", y = "Residual")
```



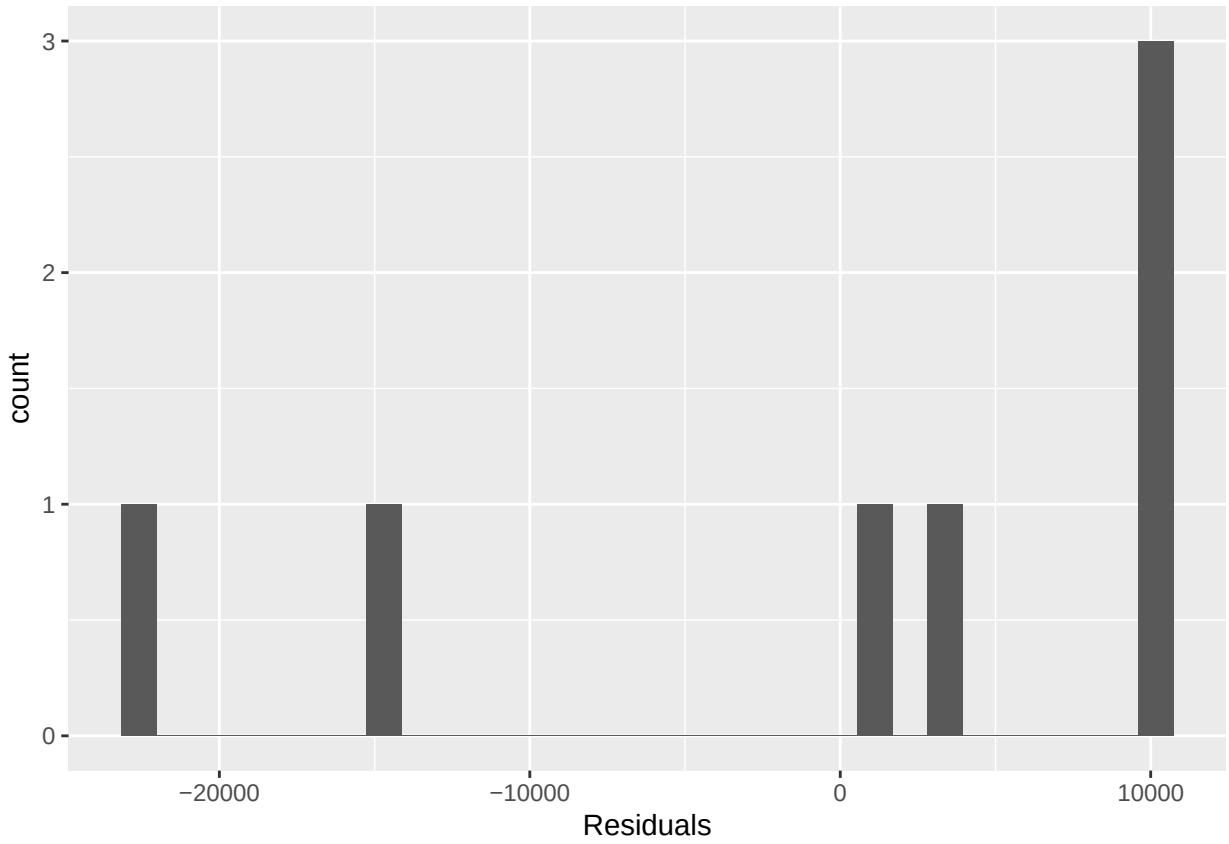
The next requirement is that of normality. The checks for this involved both a plotting of the residuals on a histogram as well as the creation of a Q-Q plot. The histogram was examined for a potential relationship between the residuals and the fit of the model to the qq line was checked too.

```
ggplot(climate_model_aug, mapping = aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```




```
ggplot(climate_model_aug, mapping = aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Given that the data passed these conditions checks we decided to perform them with a much more complicated model in order to observe the power of our predictors in the presence of each other.

```
#climateData <- climate %>%
  #mutate(MiddleEastAndCentralAsianBelt = ifelse(`Country Name` ==
  #"Syrian Arab Republic" |
        #`Country Name` == "Iraq" | `Country Name` == "Afghanistan" |
  #`Country Name` == "Pakistan" #|
       # `Country Name` == "Iran, Islamic Rep.", 1, 0)) %>%
  #mutate(SouthernAfricanBelt = ifelse(`Country Name` == "Angola" | `
  #Country Name` == "Zambia" |
        #`Country Name` == "Zimbabwe" | `Country Name` == "Mozambique" |
  # `Country Name` == "Malawi" |
       # `Country Name` == "Madagascar", 1, 0)) %>%
  #mutate(SahelHornBelt = ifelse(`Country Name` == "Senegal" |
  #`Country Name` == "Mauritania" |
      #  `Country Name` == "Mali" |`Country Name` == "Burkina Faso" |
  #`Country Name` == "Niger" |
      #  `Country Name` == "Chad" | `Country Name` == "Sudan" |
  #`Country Name` == "Eritrea", 1, 0))
```

```
climateData2 <- climateData %>%
  filter(`Series Name` == "CO2 emissions (metric tons per capita)" |
           `Series Name` == "Net migration" )
  #filter("")

climate2 <- data.frame(t(climateData2[-1]))
#climate2 %>%
```

```
# filter()
```

Results

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.

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 and Concerns:? 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. MORE HERE