Relationship between country's distance from the equator and COVID-19 cases and deaths

By: Xavier Lim

June 5, 2020

Contents

Introduction

The purpose of this project is to examine the relationship between a country's distance from the equator and their total number of COVID-19 cases and deaths. Earlier on in the pandemic, many hypothesized that warmer temperatures may prevent or slow down COVID-19. Thus, warmer countries may be less susceptible to the disease. However, research has shown that this notion is false. The World Health Organization (WHO) has proven you can catch COVID-19 no matter how sunny or hot the weather is. [1]

To further test this hypothesis, I will test whether countries farther away from the equator (which are assumed to have lower temperatures) tend to have more COVID-19 cases and deaths than countries near the equator (assumed to have warmer temperatures). To determine each country's distance from the equator, their absolute latitude will be calculated prior to the analysis. Then, a correlation analysis will be preformed to determine the correlation between the absolute latitude of a country and their:

- Total number of cases
- Total number of deaths

To account for differences in population across the countries, another correlation analysis will be performed to determine the correlation between the absolute latitude of a country and their:

- Number of cases per million
- Number of deaths per million

If the correlation coefficients are large (close to 1) and positive, it would indicate countries far from the equator (high absolute latitude) tend to have more COVID-19 cases and deaths, while a large negative correlation would indicate countries close to the equator (low absolute latitude) tend to have more COVID-19 cases and deaths.

[1] Myth busters. (n.d.). Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

Data Cleansing

There are two data sets used in this project:

- The **COVID-19 Statistics** data set presents a variety of information about each country and their COVID-19 statistics. However, the variables of interest for this project include each country (`location`) and their: total cases (`total_cases`), total deaths (`total_deaths`), cases per million people (`total_cases_per_million`), and deaths per million people (`total_deaths_per_million`). This dataset was collected from: https://github.com/owid/covid-19-data/tree/master/public/data
- The **Latitude by Country** data set presents each country (`location`) and their latitude (`latitude`) collected from: *https://developers.google.com/public-data/docs/canonical/countries_csv*

First, the current COVID-19 case and death totals of each country will be extracted from the **COVID-19 Statistics** data set. After extracting each country's latitude from the **Latitude by Country** data set, the absolute latitude of each country will be calculated. Next, the two data sets will be joined together to show each country's COVID-19 statistics and their corresponding absolute latitude. Countries without a registered latitude and/or any COVID-19 cases will be removed from the data set due to these variables being necessary to run a correlation analysis later on.

Please note that since there are over 20,000 records in the **COVID-19 Statistics** data set and over 200 records in the **Latitude by Country** data set, only the first 10 entries of each table will be presented throughout the data cleansing portion of this report.

COVID-19 Statistics

```
/* Import Excel File with Each Country's Number of COVID-19 Deaths and Cases per Day*/
proc import
       datafile='/home/u44640293/owid-covid-data.csv'
       out=covid
       dbms=csv
       replace;
       getnames=yes;
       guessingrows= max;


/* Determine Each Country's Current Total Number of COVID-19 Cases & Deaths*/
data current_totals;
       set covid;
       if date="05JUN2020"d;
       keep location total_cases total_deaths total_cases_per_million
            total_deaths_per_million;


/* Sort by Country to Perform Correlation Analysis Later On*/
proc sort data=current_totals out=country_totals;
       by location;

proc print data=country_totals(obs=10);
```

Obs	location	total_cases	total_deaths	total_cases_per_million	total_deaths_per_million
1	Afghanistan	18054	300	463.775	7.706
2	Albania	1197	33	415.943	11.467
3	Algeria	9831	681	224.191	15.53
4	Andorra	852	51	11026.985	660.066
5	Angola	86	4	2.617	0.122
6	Anguilla	3	0	199.973	0
7	Antigua and Barbuda	26	3	265.501	30.635
8	Argentina	19255	588	426.035	13.01
9	Armenia	11221	176	3786.741	59.395
10	Aruba	101	3	945.994	28.09

Latitude by Country

```
/* Import Excel File with Each Country's Latitude*/
proc import
	datafile='/home/u44640293/latitude_data.xlsx'
	out=latitude
	dbms=xlsx
	replace;
	sheet=Sheet1;
	getnames=yes;

proc print data=latitude(obs=10);
```

Obs	location	latitude
1	Afghanistan	33.93911
2	Albania	41.153332
3	Algeria	28.033886
4	American Samoa	-14.270972
5	Andorra	42.546245
6	Angola	-11.202692
7	Anguilla	18.220554
8	Antarctica	-75.250973
9	Antigua and Barbuda	17.060816
10	Argentina	-38.416097

```
/* Calculate Absolute Value of Each Country's Latitude To Get Their Distance From The
Equator*/
data absLatitude;
	set latitude;
	AbsoluteLatitude=abs(Latitude);

/* Sort by Country to Perform Correlation Analysis Later On*/
proc sort data=absLatitude out=absoluteLatitude;
	by location;

proc print data=absoluteLatitude(obs=10);
```

Obs	location	latitude	AbsoluteLatitude
1	Afghanistan	33.93911	33.9391
2	Albania	41.153332	41.1533
3	Algeria	28.033886	28.0339
4	American Samoa	-14.270972	14.2710
5	Andorra	42.546245	42.5462
6	Angola	-11.202692	11.2027
7	Anguilla	18.220554	18.2206
8	Antarctica	-75.250973	75.2510
9	Antigua and Barbuda	17.060816	17.0608
10	Argentina	-38.416097	38.4161

COVID-19 Statistics and Absolute Latitude by Country

```
/*Merge Each Country's Absolute Latitude and Total COVID-19 Cases and Deaths*/
data covidAll;
   MERGE absoluteLatitude country_totals;
   BY location;

proc print data=covidAll(obs=10);
```

Obs	location	latitude	AbsoluteLatitude	total_cases	total_deaths	total_cases_per_million	total_deaths_per_million
1	Afghanistan	33.93911	33.9391	18054	300	463.775	7.706
2	Albania	41.153332	41.1533	1197	33	415.943	11.467
3	Algeria	28.033886	28.0339	9831	681	224.191	15.53
4	American Samoa	-14.270972	14.2710	.	.	.	.
5	Andorra	42.546245	42.5462	852	51	11026.985	660.066
6	Angola	-11.202692	11.2027	86	4	2.617	0.122
7	Anguilla	18.220554	18.2206	3	0	199.973	0
8	Antarctica	-75.250973	75.2510	.	.	.	.
9	Antigua	17.060816	17.0608	26	3	265.501	30.635
10	Argentina	-38.416097	38.4161	19255	588	426.035	13.01

```
/*Remove Countries Without a Registered Latitude and/or Any COVID Cases*/
data covidFinal;
      set covidAll;
      if latitude = ' ' or total_cases = ' ' then delete;

proc print data=covidFinal(obs=10);
```

Obs	location	latitude	AbsoluteLatitude	total_cases	total_deaths	total_cases_per_million	total_deaths_per_million
1	Afghanista	33.93911	33.9391	18054	300	463.775	7.706
2	Albania	41.153332	41.1533	1197	33	415.943	11.467
3	Algeria	28.033886	28.0339	9831	681	224.191	15.53
4	Andorra	42.546245	42.5462	852	51	11026.985	660.066
5	Angola	-11.202692	11.2027	86	4	2.617	0.122
6	Anguilla	18.220554	18.2206	3	0	199.973	0
7	Antigua	17.060816	17.0608	26	3	265.501	30.635
8	Argentina	-38.416097	38.4161	19255	588	426.035	13.01
9	Armenia	40.069099	40.0691	11221	176	3786.741	59.395
10	Aruba	12.52111	12.5211	101	3	945.994	28.099

Macro Functions

The function below takes in a variable of interest, a cutoff value for the variable of interest, and the data set the variable belongs to. It returns an updated data set with only countries above the cutoff value for the specified variable. This function will be used later on to set which countries will be labelled on a graph.

```
/*Macro Function to Set Cutoff Value to be Labelled on Graph*/
%macro label_cutoff(cutoff_variable=, cutoff_value=, dataSet=);
data labels;
	set &dataSet;
	Label = location;
	if &cutoff_variable < &cutoff_value then
		Label = " ";
put Label;
%mend;
```

The function below takes in the name of a country to be excluded from a data set and the name of the data set. It will return an updated data set without the selected country. This function will be used later on to exclude countries that appear to be outliers.

```
/*Macro Function to Remove Outlier Countries from Data Set*/
%macro remove_country(country_to_remove=, dataSet=);
data covidWithout;
	set &dataSet;
	if location ~= &country_to_remove;
put covidWithout;
%mend;
```

Correlation Analysis

```
/*Correlation Between Absolute Latitude and Variables of Interest*/
proc corr data=covidFinal;
      var absoluteLatitude;
      with total_cases total_deaths total_cases_per_million
           total_deaths_per_million;
```

Pearson Correlation Coefficients, N = 200	
	AbsoluteLatitude
total_cases	0.11176
total_deaths	0.13621
total_cases_per_million	0.27827
total_deaths_per_million	0.35219

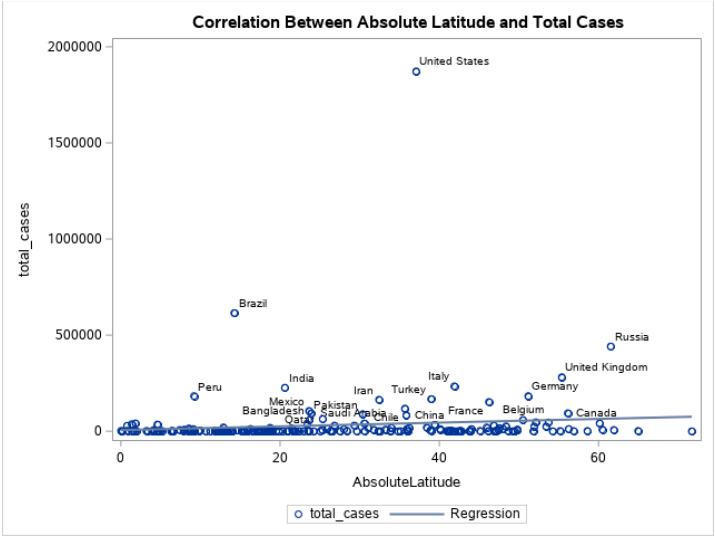
As you can see from the results of the correlations, the direction of all correlations are positive since all correlation coefficients are positive. This means being farther from the equator (higher absolute latitude) is associated with more COVID-19 cases and deaths. However, it is important to note that these relationships vary in terms of correlation strength:

- Absolute Latitude is moderately correlated with total COVID-19 deaths per million (0.35)
- Absolute Latitude is weakly correlated with total COVID-19 cases (0.11), deaths (0.14), and cases per million (0.28)

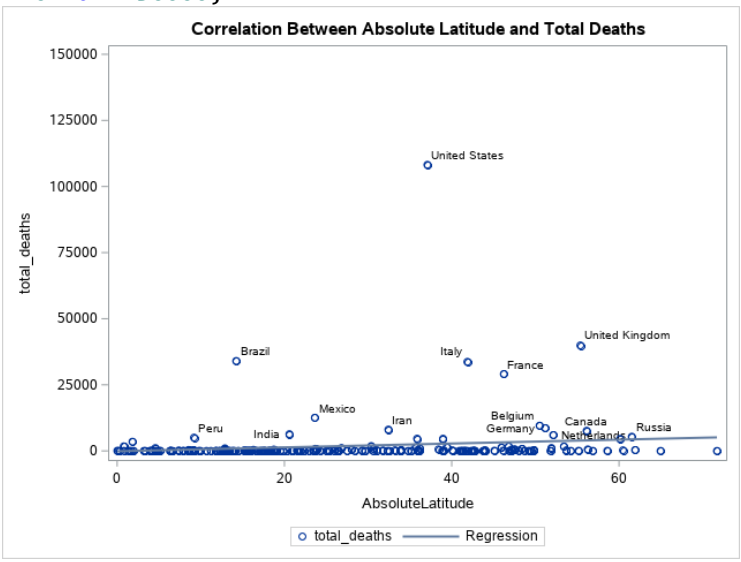
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
total_cases	200	31780	147040	6355960	3.00000	1872660
total_deaths	200	1819	9095	363703	0	108211
total_cases_per_million	200	1441	2832	288255	0.89400	22124
total_deaths_per_million	200	55.04664	143.06653	11009	0	1238
AbsoluteLatitude	200	26.65217	16.94389	5330	0.02356	71.70694

Taking a look at the summary statistics, there is quite a lot of variability in the number of cases, deaths, cases per million, and deaths per million across the 200 countries in the data set. To get a better visual representation of the variability in the data, let's take a look at some scatterplots illustrating the relationship between absolute latitude and the various variables.

```
/*Scatterplot Showing Relationship Between Absolute Latitude and Total Cases*/
%label_cutoff(cutoff_variable=total_cases, cutoff_value=50000,dataSet=covidFinal);
proc sgplot;
title "Correlation Between Absolute Latitude and Total Cases";
	scatter x=AbsoluteLatitude y=total_cases / datalabel=Label;
	reg x=AbsoluteLatitude y=total_cases;
	yaxis min=0 max=2000000;
```

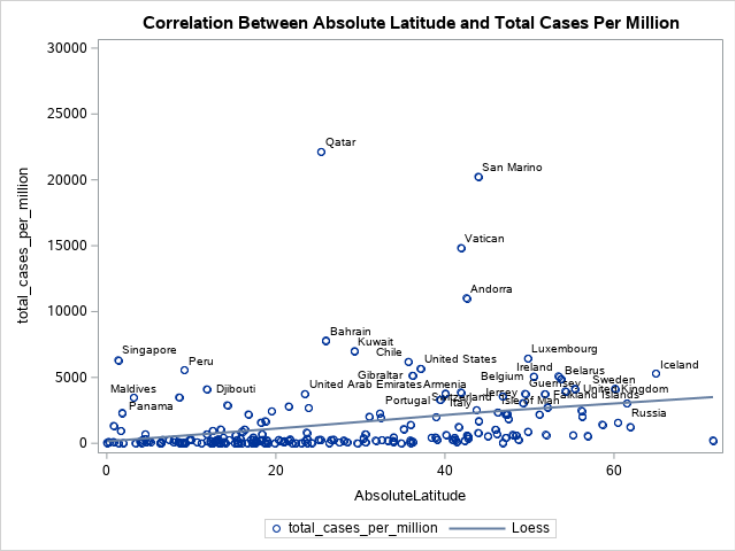



```
/*Scatterplot Showing Relationship Between Absolute Latitude and Total Deaths*/
%label_cutoff(cutoff_variable=total_deaths, cutoff_value=5000,dataSet=covidFinal);
proc sgplot;
title "Correlation Between Absolute Latitude and Total Deaths";
	scatter x=AbsoluteLatitude y=total_deaths / datalabel=Label;
	reg x=AbsoluteLatitude y=total_deaths;
	yaxis min=0 max=150000;
```



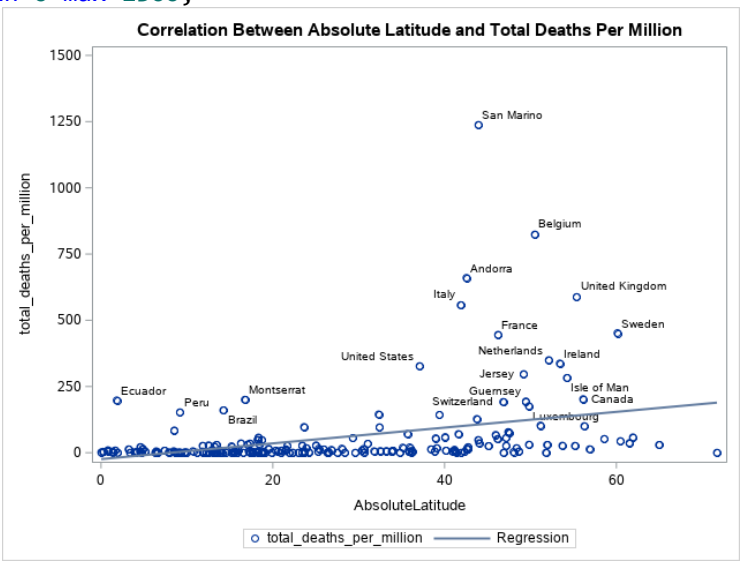
Based on the two graphs above, it is evident that there is quite a lot of variability in the number of total cases and deaths across the different countries. Overall, the general trend is positive which is outlined by the upward-sloping regression line in each graph which suggests countries farther from the equator have more cases and deaths. However, this correlation is very weak as there are countries both close to the equator (ex. Brazil, Peru, India, and Mexico) and far from the equator (ex. UK, Russia, and Canada) that appear to have many cases and deaths. Now to account for differences in population, let's take a look at the relationship between the absolute latitude of a country and their number of cases and deaths per million people.

```
/*Scatterplot Showing Relationship Between Absolute Latitude and Cases Per Million*/
%label_cutoff(cutoff_variable=total_cases_per_million,cutoff_value=3000,
dataSet=covidFinal);
proc sgplot;
title "Correlation Between Absolute Latitude and Total Cases Per Million";
	scatter x=AbsoluteLatitude y=total_cases_per_million / datalabel=Label;
	reg x=AbsoluteLatitude y=total_cases_per_million;
	yaxis min=0 max=30000;
```



It is evident that there is quite a lot of variability in the number of cases per million across the different countries. Overall, the general trend is positive which is outlined by the upward-sloping regression line which suggests countries farther from the equator have more cases per million. Similar to the graph looking at total cases and deaths, this correlation is also very weak as there are countries both close to the equator (ex. Singapore and Peru) and far from the equator (ex. Iceland and Sweden) that appear to have many cases per million.

```
/*Scatterplot Showing Relationship Between Absolute Latitude and Deaths Per Milion*/
%label_cutoff(cutoff_variable=total_deaths_per_million, cutoff_value=150,
dataSet=covidFinal);
proc sgplot;
title "Correlation Between Absolute Latitude and Total Deaths Per Million";
	scatter x=AbsoluteLatitude y=total_deaths_per_million / datalabel=Label;
	reg x=AbsoluteLatitude y=total_deaths_per_million;
	yaxis min=0 max=1500;
```

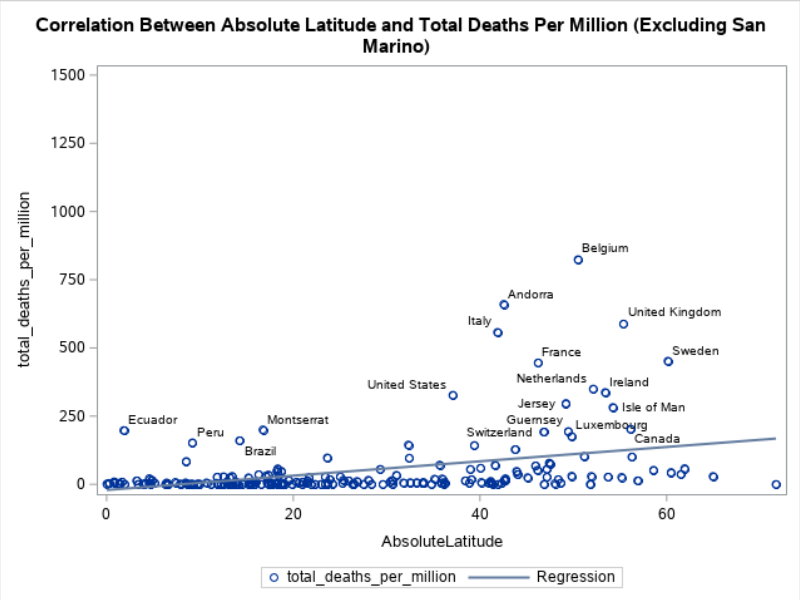


Now, looking at deaths per million, there appears to be less variability across the different countries. In addition, the correlation is much stronger since there are countries far from the equator (ex. Belgium, UK, Sweden) that have far more deaths per million compared to countries near the equator. San Marino appears to be a clear outlier as it has a lot more deaths per million than any other countries. Let's remove it to see if we get a clearer trend.

```
/*Remove San Marino*/
%remove_country(country_to_remove="San Marino", dataSet=covidFinal);
proc corr;
	VAR AbsoluteLatitude total_deaths_per_million;
```

Pearson Correlation Coefficients, N = 199		
	AbsoluteLatitude	total_deaths_per_million
AbsoluteLatitude	1.00000	0.38355
total_deaths_per_million	0.38355	1.00000

```
%label_cutoff(cutoff_variable=total_deaths_per_million, cutoff_value=150,
dataSet=covidWithout);
proc sgplot;
title "Correlation Between Absolute Latitude and Total Deaths Per Million (Excluding
San Marino)";
	scatter x=AbsoluteLatitude y=total_deaths_per_million / datalabel=Label;
	reg x=AbsoluteLatitude y=total_deaths_per_million;
	yaxis min=0 max=1500;
```



After removing San Marino from the data set, the correlation did in fact increase from 0.35 to 0.38.

Conclusion

The absolute latitude of a country has a:

- moderate positive correlation with COVID-19 deaths per million
- weak positive correlation with COVID-19 cases, deaths, and cases per million

In conclusion, there appears to be close to no correlation between a country's distance from the equator and their number of COVID-19 cases and deaths. However, there is a moderate positive correlation between a country's distance from the equator and COVID-19 deaths per million. Thus, although countries far from the equator are not necessarily more likely to contract COVID-19, it may be important to explore whether countries far from the equator are more susceptible to die from COVID-19 once they get it. However, there may be other confounding variables impacting the correlation analysis results. For example, countries far from the equator may tend to have an older population distribution leading to a higher COVID-19 death rate.