COVID-19 Death Rate and Case Count Study

John R.

2023-06-03

Introduction

The COVID-19 pandemic spread with unprecedented speed, impacting lives worldwide in a matter of months. Analysis of data that have been taken since the start of the pandemic can help us understand many aspects of this catastrophic event. The analysis that follows seeks to answer the question of whether trends in infection and death rate in the United States and abroad depend on lockdown and masking policies. This report draws interesting observations about these relationships that merit further analysis and consideration.

```
#online location: https://github.com/jore2414/Data-Science-as-a-Field/...
#blob/main/covid19-final.Rmd

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
knitr::opts_chunk$set(echo = TRUE)
```

Import the data

First build an array that contains the persistent location of the data to ensure the analysis is repeatable. The data come from Johns Hopkins University and contain a variety of information related to the country or U.S. state, the number of cases, and the number of deaths as a function of time.

```
# Get current data in the four files

# Breaking the url into two lines so it displays properly when knit to PDF.
# There are other ways of doing this, but they require other libraries to be installed.
url_in1 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19"
url_in2 <- "/master/csse_covid_19_data/csse_covid_19_time_series/"
```

```
url_in <- str_c(url_in1, url_in2)
file_names <-
  c("time_series_covid19_confirmed_global.csv",
    "time_series_covid19_deaths_global.csv",
    "time_series_covid19_confirmed_US.csv",
    "time_series_covid19_deaths_US.csv")
urls <- str_c(url_in,file_names)
```

Now read in the data.

```
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])
```

Tidy the data

The imported global data are organized by location with columns indicating dates that correspond to case or death counts for a given location and date. To make the data more amenable to analysis, it is helpful to organize the data so each date corresponds to a row. Locations will be repeated.

First tidy up the global data

```
#organize global cases by date and eliminate Lat/Long
global_cases <- global_cases %>%
  pivot_longer(cols = -c('Province/State',
                         'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(Lat,Long))

#organize global deaths by date and eliminate Lat/Long
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State',
                         'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat,Long))

#merge global cases with deaths and cast dates from char to object type date
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))

#eliminate rows where cases are 0
global <- global %>% filter(cases > 0)
```

Now tidy up the US data

```
#organize US cases by date and eliminate unneeded data such as Lat/Long
US_cases <- US_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

#organize US deaths by date and eliminate unneeded data such as Lat/Long
US_deaths <- US_deaths %>%
  pivot_longer(cols = -(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

#merge US cases and deaths
US <- US_cases %>%
  full_join(US_deaths)
```

Now introduce some interesting new data (country population) into the global data

```
global <- global %>%
  unite("Combined_Key",
        c(Province_State, Country_Region),
        sep = ", ",
        na.rm = TRUE,
        remove = FALSE)

# Repeat the method from above so the URL displays properly when knit to PDF
uid_lookup_url1 <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/"
uid_lookup_url2 <-
  "master/csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv"
uid_lookup_url <- str_c(uid_lookup_url1, uid_lookup_url2)

uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))

global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, Population,
         Combined_Key)
```

Analyze the US data by deaths per million

```
#Organize the US data by state, and identify the deaths per million people
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()

#Group the entire US to look at the death rate (deaths_per_mill)
US_totals <- US_by_state %>%
  group_by(Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths*1000000 / Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

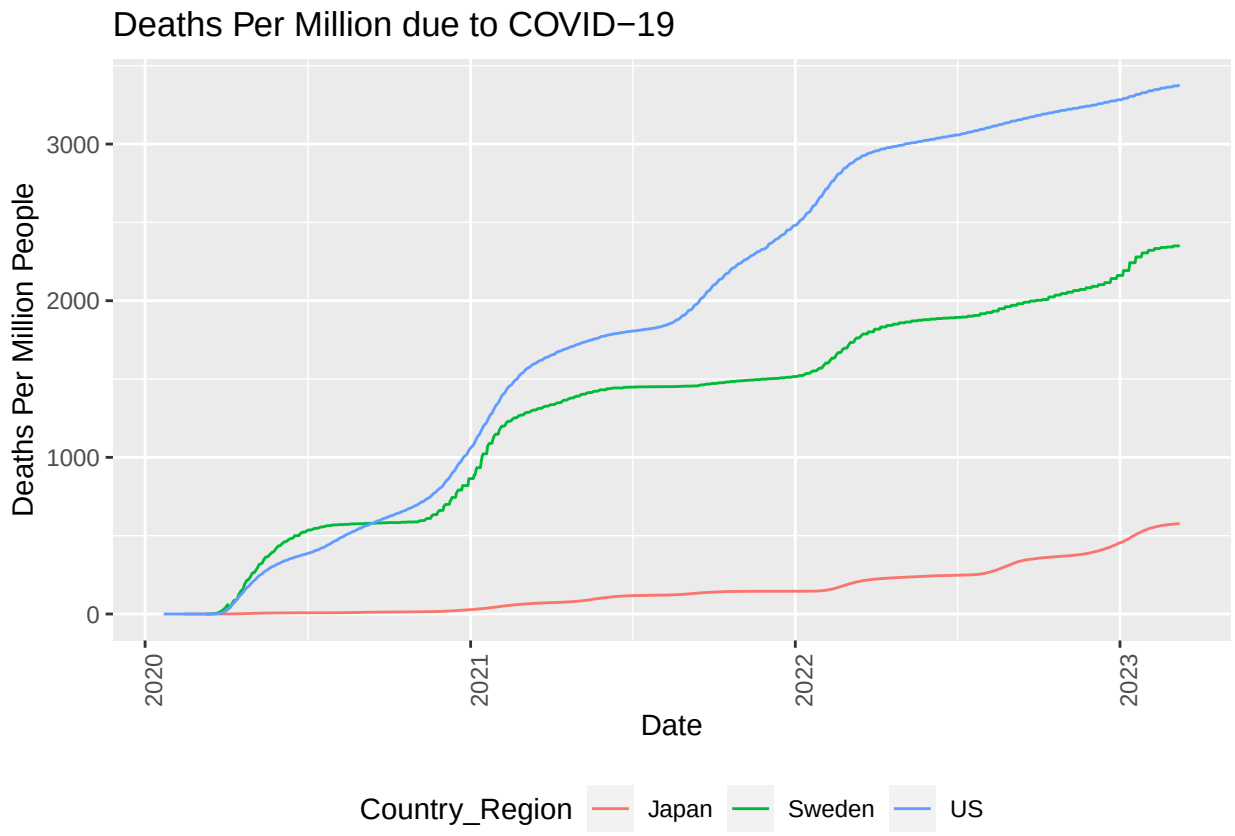
Analyze the data from two countries that did not lockdown, Japan and Sweden, by deaths per million

```
Japan_totals <- global %>%
  filter(Country_Region == "Japan")

Japan_totals <- Japan_totals %>%
  mutate(deaths_per_mill = deaths*1000000 / Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()

Sweden_totals <- global %>%
  filter(Country_Region == "Sweden")

Sweden_totals <- Sweden_totals %>%
  mutate(deaths_per_mill = deaths*1000000 / Population) %>%
  select(Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

Produce and interpret some interesting plots

First plot the death rate versus time in the US, Sweden, and Japan.

```
intl_death_totals <- bind_rows(US_totals, Sweden_totals, Japan_totals)

intl_death_totals %>%
  filter(deaths_per_mill>0) %>%
  ggplot(aes(x = date, y = deaths_per_mill, color=Country_Region)) +
  geom_line()+
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Deaths Per Million due to COVID-19", y = "Deaths Per Million People",
       x = "Date")
```

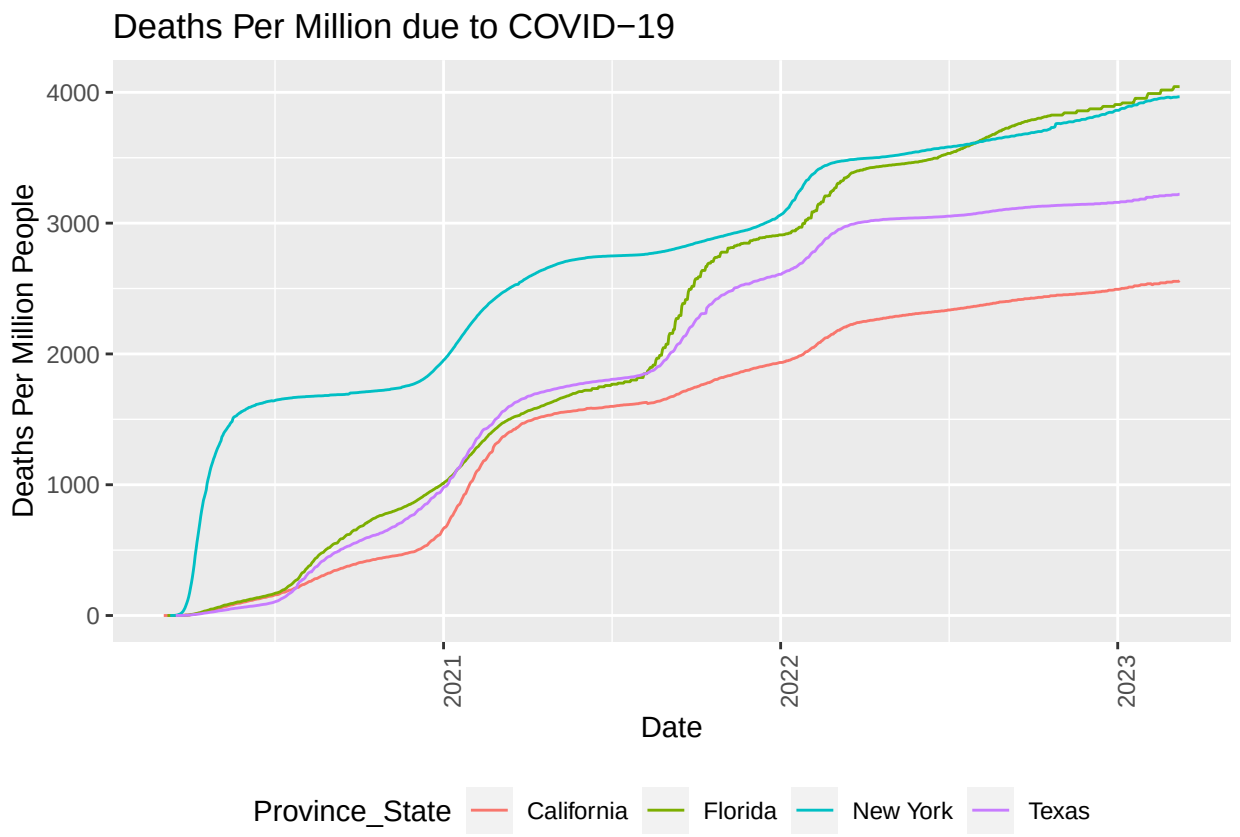


Interpretation

Considering Japan and Sweden did not lockdown and all three countries have modern health care systems, it is a little bit counter-intuitive that the COVID-19 death rates in Japan and Sweden were, in fact, lower than the death rates in the United States. Because different countries have different methods for reporting that could bias the data, it would be interesting to compare U.S. states with very different lockdown policies. Florida and Texas are two large states known for minimal to no lockdown policies, while California and New York had stricter lockdown policies. Next, we will compare the death rates between these states.

Plot the death rates of Texas, Florida, California, and Texas

```
Texas <- filter(US_by_state, Province_State == "Texas")
Florida <- filter(US_by_state, Province_State == "Florida")
NewYork <- filter(US_by_state, Province_State == "New York")
California <- filter(US_by_state, Province_State == "California")

state_death_totals <- bind_rows(Texas,Florida,NewYork,California)

state_death_totals %>%
  filter(deaths_per_mill>0) %>%
  ggplot(aes(x = date, y = deaths_per_mill, color=Province_State)) +
  geom_line()+
  theme(legend.position="bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Deaths Per Million due to COVID-19", y = "Deaths Per Million People", x = "Date")
```



Interpretation

This plot shows another interesting result. The two states that opposed lockdown measures, Texas and Florida, witnessed death rates that were generally in between the death rates of two states with strict lockdown policies, California and New York. The U.S. state data combined with international data suggest that lockdown measures are not necessarily a strong predictor of the death rate within a large population.

Other factors such as population density, age, lifestyle, obesity, and data collection methods could influence both the actual death rates and the calculated death rates.

Model the data

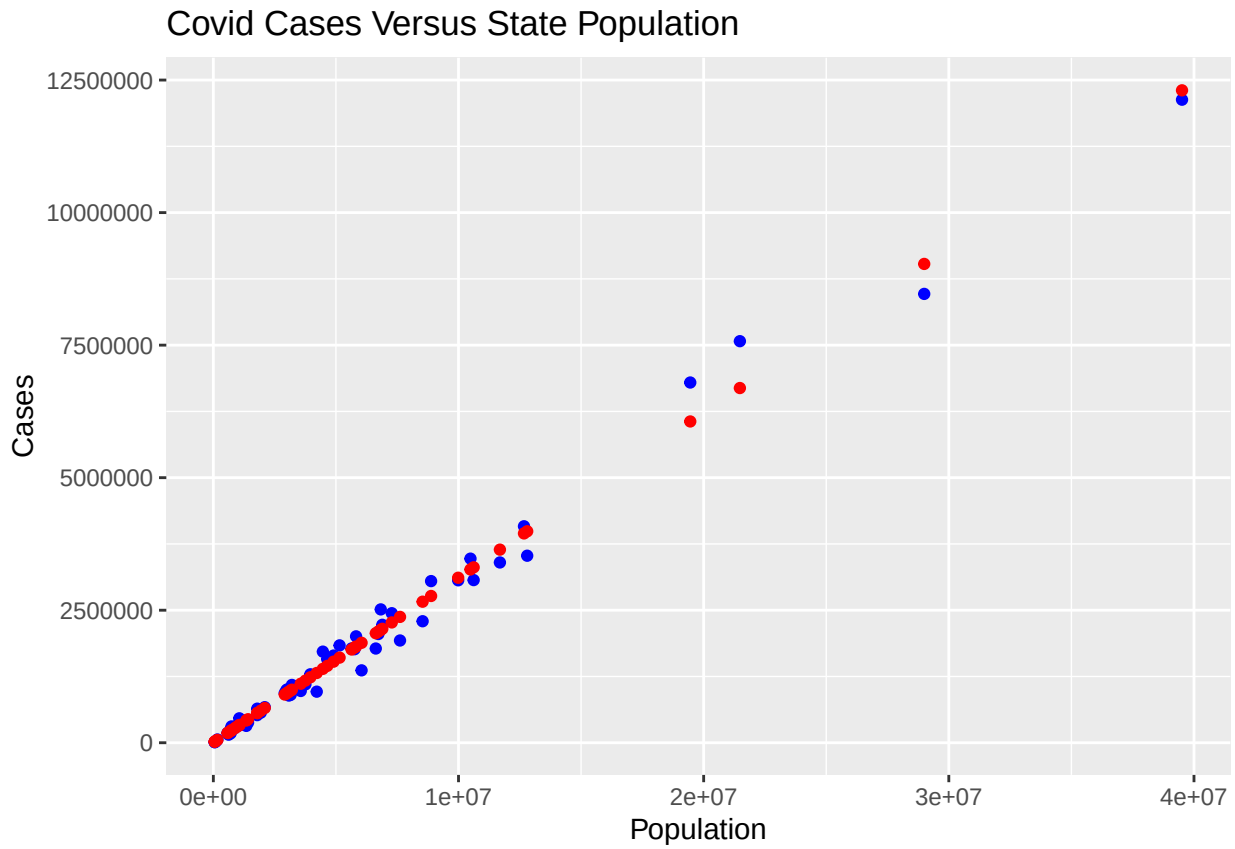
Having observed in isolated cases that the effectiveness of lockdown policies is not obvious, it would be interesting to look at a broader scale whether the policies of any states or countires led to a drastic reduction in the number of cases relative to what may be predicted. These will show up as outliers when shown against a simple linear model.

```
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000* cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases > 0, population >0)

country_totals <- global %>%
  group_by(Country_Region) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000* cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases > 0, population >0)

US_model <- lm(cases ~ population, data = US_state_totals)
global_model <- lm(cases ~ population, data = country_totals)

US_tot_w_pred <- US_state_totals %>% mutate(pred = predict(US_model))
country_tot_w_pred <- country_totals %>% mutate(pred = predict(global_model))
```

Plot the models

Plot the linear regression and actual data of cases versus state population

```
US_tot_w_pred %>% ggplot() +
  geom_point(aes(x = population, y = cases), color = "blue") +
  geom_point(aes(x = population, y = pred), color = "red") +
  labs(title = "Covid Cases Versus State Population", y = "Cases", x = "Population")+
  theme(legend.position="bottom")
```




```
summary(US_model)
```

```
##
## Call:
## lm(formula = cases ~ population, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -565364  -70751   -2556   83869  884049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.678e+03  4.409e+04   0.061    0.952
## population  3.114e-01  4.742e-03  65.668   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253700 on 54 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9874
## F-statistic:  4312 on 1 and 54 DF,  p-value: < 2.2e-16
```

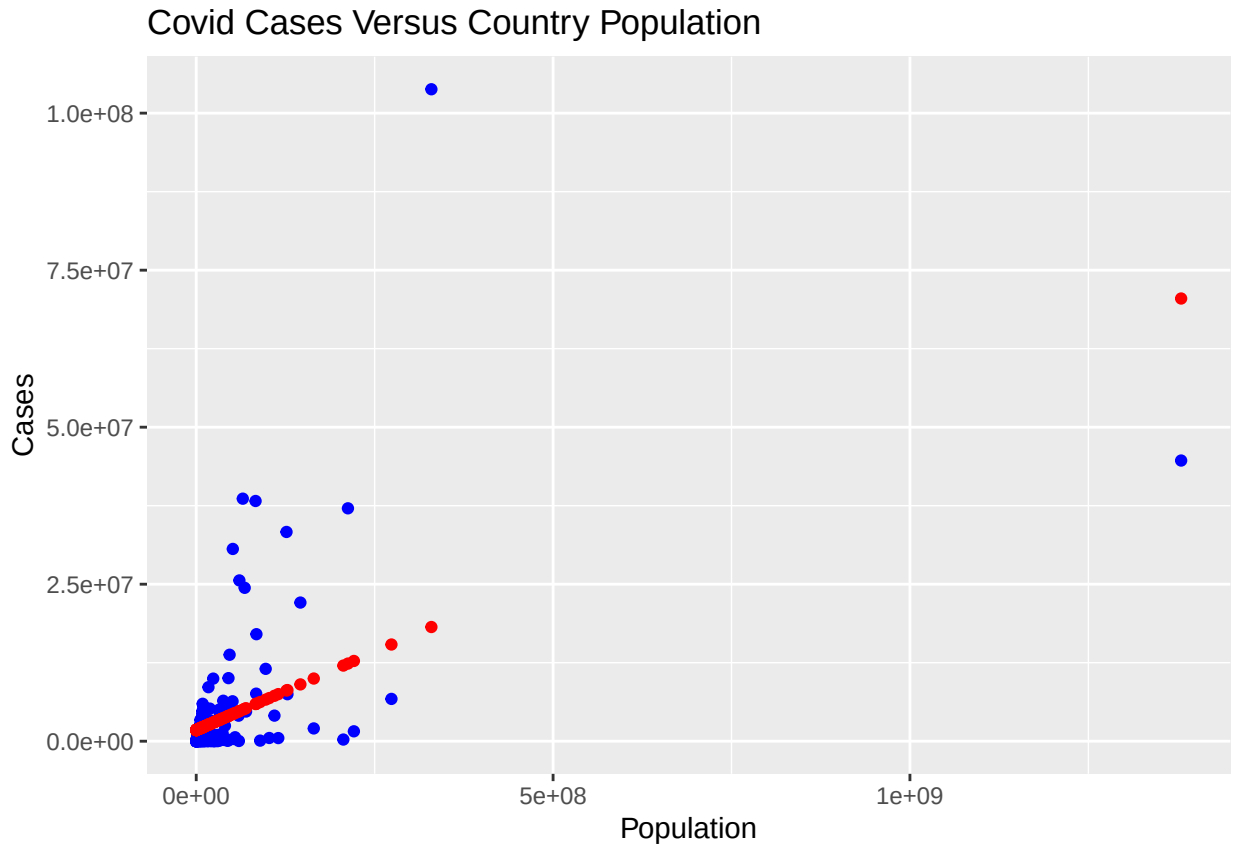
Interpretation

The US data show extremely high correlation between cases and population. It would be interesting to dig deeper to see whether the widely varying policies from state-to-state really did not have a strong impact on

the spread of COVID infections. Bias in data collection may have skewed some of the outliers back toward the mean. Let's see if the same can be observed in the global data set.

Plot the linear regression and actual data of cases versus state population

```
country_tot_w_pred %>% ggplot() +
  geom_point(aes(x = population, y = cases), color = "blue") +
  geom_point(aes(x = population, y = pred), color = "red") +
  labs(title = "Covid Cases Versus Country Population", y = "Cases", x = "Population")+
  theme(legend.position="bottom")
```




```
summary(global_model)
```

```
##
## Call:
## lm(formula = cases ~ population, data = country_totals)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -25795046  -2173404  -1740719   -742791  85619284
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 1.781e+06  6.555e+05   2.716   0.0072 **
## population  4.979e-02  5.825e-03   8.547  3.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8744000 on 192 degrees of freedom
## Multiple R-squared:  0.2756, Adjusted R-squared:  0.2718
## F-statistic: 73.05 on 1 and 192 DF,  p-value: 3.897e-15
```

Interpretation

In the global data, total case count is far less correlated with population than it is in the U.S. There could be a variety of reasons for this, and many could be considered sources of bias: wider variability in policy measures, different data collection methods across countries, and government manipulation of data. Additionally, there were many countries reporting an unrealistically low number of cases, potentially because they did not have access to appropriate testing equipment.

Conclusion

Even a data set that seems simple on its surface, such as Covid case and death counts can reveal rich and counterintuitive insights into the impact of policies, and the importance of collecting data in a consistent manner. This report presents an initial view of whether Covid policies within the United States and between the United States, Japan, and Sweden impacted Covid death rates. Perhaps surprisingly, both within the United State and between the three countries, the data did not reveal an obvious trend that lockdown measures strongly influenced death rates over time. There may be some bias here as states and countries without lockdown measures may have weaker reporting incentive.

Further analysis of case counts within the United States showed that state population was a strong predictor of the total number of cases, and there were no obvious outliers among the states. Knowing that various states had very different approaches to lockdown and masking measures, with some imposing strict measures and some with virtually no measures, I would have expected higher variability in the data relative to the model if policies were impactful in preventing Covid cases. That said, state by state data may be biased based on local and regional collection methods, and this is something that would need to be further analyzed prior to making a strong conclusion.

The global data for Covid cases versus population show a much weaker correlation than the US data. This does not prove that other countries had more effective policies to prevent the spread of infection. Instead, it begs questions about how data are collected and managed (i.e. there are likely systemic biases) within a global event, and highlights that the influence of these factors needs to be well understood prior to making strong conclusions from even reputable data sources. This would be an appropriate avenue for continued investigation.