Analyzing COVID Cases and Deaths Through the Lens of Poverty

Data Science Student

January 28, 2023

Introduction

For this project, I would like to evaluate the impact of COVID as it relates to poverty. In particular, I would like to determine whether US states that have higher levels of poverty fared worse throughout the pandemic. In order to undertake this analysis, I will use the John Hopkins COVID time series dataset, which can be found on GitHub, in addition to US poverty statistics available through the United States Department of Agriculture.

John Hopkins Dataset: From John Hopkins University, there are two time series tables available for confirmed cases and deaths in the United States, reported at the county level. They are named time_series_covid19_confirmed_US.csv, time_series_covid19_deaths_US.csv, respectively. The data begin with the start of the pandemic in the US, and they are current as of the present. These will be loaded from the web.

```
base_url <-
"https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_series/"

file_names <- c(
"time_series_covid19_confirmed_US.csv",
"time_series_covid19_deaths_US.csv"
)

urls <- str_c(base_url,file_names)

us_cases <- read_csv(urls[1])
us_deaths <- read_csv(urls[2])
```

USDA Poverty Rate Statistics: The US Department of Agriculture Economic Research Service provides poverty rate statistics. Year 2020 poverty rate information has been chosen to coincide with the beginning of the COVID19 pandemic. No downloadable dataset is available, so this data is added manually to the data model. The original information can be found here: https://data.ers.usda.gov/reports.aspx?ID=17826

```
Province_State <- c("Alabama","Alaska","Arizona","Arkansas","California","Colorado",
"Connecticut","Delaware","District of Columbia","Florida","Georgia","Hawaii","Idaho",
"Illinois","Indiana","Iowa","Kansas","Kentucky","Louisiana","Maine","Maryland",
"Massachusetts","Michigan","Minnesota","Mississippi","Missouri","Montana","National",
"Nebraska","Nevada","New Hampshire","New Jersey","New Mexico","New York",
"North Carolina","North Dakota","Ohio","Oklahoma","Oregon","Pennsylvania",
"Rhode Island","South Carolina","South Dakota","Tennessee","Texas","Utah",
"Vermont","Virginia","Washington","West Virginia","Wisconsin","Wyoming")
```

```
poverty_rate <- c(14.9,9.6,12.8,15.2,11.5,9,9.7,10.9,15,12.4,14,8.9,10.1,11,11.6,
10.2,10.6,14.9,17.8,10.6,9,9.4,12.6,8.3,18.7,12.1,12.4,11.9,9.2,12.5,7,9.4,16.8,
12.7,12.9,10.2,12.6,14.3,11,10.9,10.6,13.8,11.6,13.6,13.4,7.3,9.4,9.2,9.5,15.8,10,9.2)

Poverty_By_State <- data.frame(Province_State, poverty_rate)
```

Data Preparation: To make the John Hopkins data useful, it will be necessary to pivot the columns for the two datasets. Then it will be useful to join the two into a single dataframe. Additionally, the information will be grouped by state, to provide a more aggregated view of the data.

```
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_))
```

```
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_))
```

```
US <- us_cases %>%
  full_join(us_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
US_by_state <- US %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  select(Province_State, Country_Region, date,
         cases, deaths, Population) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

Additional Variables: To better make comparisons between states, a cases per thousand variable and a deaths per thousand variable is added. In addition, the poverty rate information is joined to the single table. This completes the data preparation needed for analysis.

```
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)
```

```
US_state_totals <- merge(x=US_state_totals,y=Poverty_By_State,
          by="Province_State", all.x=TRUE)

US_state_totals <- US_state_totals %>% filter(!is.na(poverty_rate))

head(US_state_totals)
```

```
##   Province_State deaths    cases population cases_per_thou deaths_per_thou
## 1        Alabama  20870  1610535    4903185       328.4671        4.256417
## 2         Alaska   1473   303575     740995       409.6856        1.987868
## 3        Arizona  32775  2398200    7278717       329.4811        4.502854
## 4       Arkansas  12835   995859    3017804       329.9946        4.253093
## 5     California 100027 11983168   39512223       303.2775        2.531546
## 6       Colorado  14024  1747627    5758736       303.4741        2.435257
##   poverty_rate
## 1         14.9
## 2          9.6
## 3         12.8
## 4         15.2
## 5         11.5
## 6          9.0
```

Data Models: I would like to see whether a poverty rate will successfully predict cases per thousand and deaths per thousands with a linear model. The predicted values are added to the data frame below.

```
mod_cases <-lm(cases_per_thou ~ poverty_rate, data = US_state_totals)
mod_deaths <-lm(deaths_per_thou ~ poverty_rate, data = US_state_totals)
summary(mod_cases)
```

```
##
## Call:
## lm(formula = cases_per_thou ~ poverty_rate, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.057 -26.705  -2.034  23.579 124.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   262.527     27.565   9.524 9.91e-13 ***
## poverty_rate    3.955      2.303   1.718   0.0921 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 42.57 on 49 degrees of freedom
## Multiple R-squared:  0.0568, Adjusted R-squared:  0.03755
## F-statistic: 2.951 on 1 and 49 DF,  p-value: 0.09215
```

```
summary(mod_deaths)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ poverty_rate, data = US_state_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96422 -0.30730  0.08213  0.43180  1.26959
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.70835    0.42181   1.679   0.0995 .
## poverty_rate  0.21805    0.03523   6.188  1.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6515 on 49 degrees of freedom
## Multiple R-squared:  0.4387, Adjusted R-squared:  0.4272
## F-statistic:  38.3 on 1 and 49 DF,  p-value: 1.198e-07
```

```
US_pov_w_pred <- US_state_totals %>%
    mutate(pred_cases = predict(mod_cases)) %>%
      mutate(pred_deaths = predict(mod_deaths))

head(US_pov_w_pred)
```

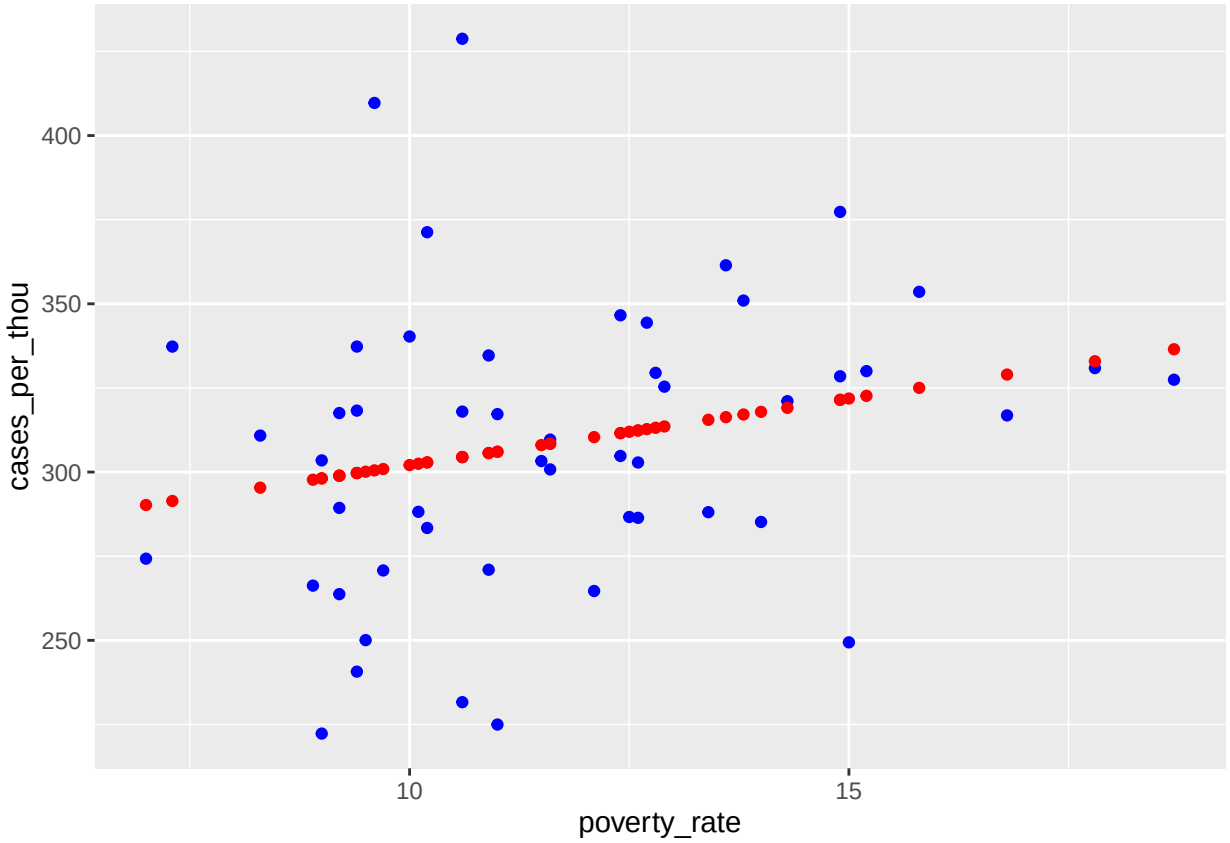
```
##   Province_State deaths    cases population cases_per_thou deaths_per_thou
## 1        Alabama  20870  1610535    4903185       328.4671        4.256417
## 2         Alaska   1473   303575     740995       409.6856        1.987868
## 3        Arizona  32775  2398200    7278717       329.4811        4.502854
## 4       Arkansas  12835   995859    3017804       329.9946        4.253093
## 5     California 100027 11983168   39512223       303.2775        2.531546
## 6       Colorado  14024  1747627    5758736       303.4741        2.435257
##   poverty_rate pred_cases pred_deaths
## 1         14.9   321.4611    3.957296
## 2          9.6   300.4978    2.801631
## 3         12.8   313.1549    3.499391
## 4         15.2   322.6477    4.022711
## 5         11.5   308.0130    3.215926
## 6          9.0   298.1246    2.670801
```

Predicting COVID Cases based on Poverty Rate: The R-Squared value for this model is 0.03755, which indicates that there is essentially no efficacy in using the poverty rate to predict the number of cases of COVID. The correlation coefficient is .24, again suggesting that Poverty rate is not a strong predictor here. In the graph below, the red predicted dots do little to capture any trend.

```
x <- US_pov_w_pred["cases_per_thou"]
y <- US_pov_w_pred["poverty_rate"]
cor(x, y)
```

```
##                 poverty_rate
## cases_per_thou    0.2383278
```

```
US_pov_w_pred %>% ggplot() +
    geom_point(aes(x = poverty_rate, y = cases_per_thou), color = "blue") +
    geom_point(aes(x = poverty_rate, y = pred_cases), color = "red")
```

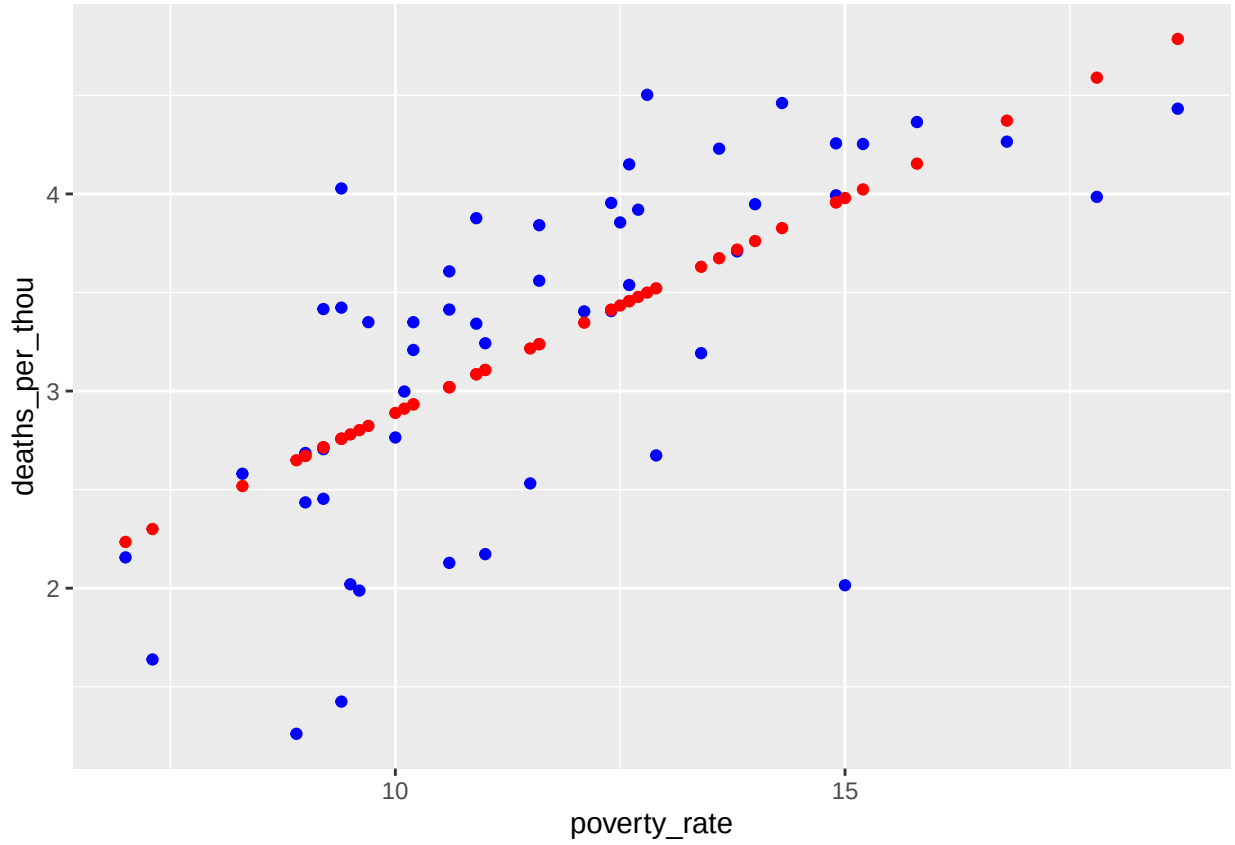


Predicting COVID Deaths based on Poverty Rate: Interestingly, the R-Squared value for this second model is 0.4272, which makes it much more effective than the previous one. In addition, the correlation coefficient between the two variables in .662, suggesting that there is a moderate correlation between poverty rate and the number of COVID deaths. The higher the poverty rate, the more likely it is that someone will die of COVID.

```
x <- US_pov_w_pred["deaths_per_thou"]
y <- US_pov_w_pred["poverty_rate"]
cor(x, y)
```

```
##                  poverty_rate
## deaths_per_thou    0.6623442
```

```
US_pov_w_pred %>% ggplot() +
    geom_point(aes(x = poverty_rate, y = deaths_per_thou), color = "blue") +
    geom_point(aes(x = poverty_rate, y = pred_deaths), color = "red")
```



Conclusions: The two models suggest that COVID spread with little regard to the relative wealth of individuals in the United States. Individuals in states with lower poverty rates did not escape being infected more than those with higher poverty rates. However, overall, more individuals in states with higher poverty rates died than those in states with lower poverty rates. Correlation does not necessarily mean causation. However, one interpretation of the results is that individuals in high poverty states had fewer resources to combat the illness or less access to necessary health care.

Possible sources of bias in the analysis include data that was not compiled accurately. In addition, other variables which have more explanatory power might not have been selected due to preconceived ideas about how diseases affected a country. It is also possible that I am suffering from confirmation bias, in which I would like to find a strong correlation between poverty and disease impact.

Loaded Libraries

```
(.packages())
```

```
##  [1] "lubridate" "magrittr"  "forcats"   "stringr"   "dplyr"     "purrr"
##  [7] "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
```