Social Factors of 2020 COVID-19 Incidence Rates

Introduction

In 2020, the COVID-19 pandemic exposed significant disparities in how different communities were affected, raising important questions about the role of social vulnerabilities in shaping public health outcomes.

The goal of this project is to explore two questions:

1. What social factors are most correlated to state-level incidence rates of COVID-19 in 2020?
2. Are those social factors still important on more granular (i.e. county) levels?

Why should you care?

The findings from this project could provide actionable insights for policymakers, public health officials, and researchers. Understanding which social factors most strongly correlate with COVID-19 incidence can help prioritize resource allocation in future health crises, inform interventions to protect our most vulnerable populations, and form deeper understanding of the relationship between social structures and public health.

This project makes use of the Social Vulnerability Index (SVI) data from the CDC. Data was downloaded for the year of 2020. Social factors include total populations, unemployment populations, minority populations, and many more.

More detailed descriptions of variables can be found here: SVI Data

I will be using the `us-states` and `us-counties-all` data to supplement the number of COVID cases in my analysis.

The project will involve three main steps:

- Perform a correlation analysis on state-level data to find which social factors appear to be most important.
- Perform a multi-regression analysis to control for the effects of other variables.
- Examine if these factors still hold a strong effect on the county-level by doing another correlation analysis.

Analysis

Data Cleaning and Prepartion

Here are all packages used in this project:

```
library(tidyverse) # For basic data manipulation.
library(lubridate) # To easily sort by date.
library(readr)     # To read in .csv files.
library(maps)      # To create a choropleth map visualization.
library(patchwork) # To patch our maps together.
library(knitr)     # To neatly display data.
```

To prepare our data, we first want to filter our state-level COVID data to only represent total cases for the year of 2020.

Since the SVI data is by county, we also need to aggregate it so that it is by state.

Then, we can merge them into one tibble, which will then be used for our analysis.

```
# Load in our datasets.

covid <- read_csv("us-states.csv")
svi <- read_csv("svi_interactive_map_2020.csv")

# Filter the COVID data for cases in 2020 only, then pick the totals (end of year).

covid_2020 <- covid %>%
  filter(year(date) == 2020)

covid_2020 <- covid_2020 %>%
  group_by(state) %>%
  filter(date == max(date)) %>%
  ungroup()

# Aggregating the SVI data to be grouped by state instead of by county.

svi_state_data <- svi %>%
  group_by(STATE) %>%
  summarise(
    across(where(is.double), ~ sum(.x, na.rm = TRUE)),
    .groups = "drop"
  )

# To merge easily by state, lower-case the SVI columns to match the COVID file's columns.

colnames(svi_state_data) <- tolower(colnames(svi_state_data))

# Creating our combined dataset.

covid_states <- covid_2020 %>%
  left_join(svi_state_data, by = "state")

# Creating a column for incidence rates by state.

covid_states <- covid_states %>%
  dplyr::mutate(IncidenceRate = (cases / e_totpop) * 100)
```

Table 1: Final State-Level COVID dataset.

date	state	fips	cases	deaths	area_sqmi	e_totpop	m_totpop	e_hu	m_hu
2020-12-31	Alabama	01	361226	4827	50647.11	4893186	530	2270398	7241
2020-12-31	Alaska	02	46740	198	571022.12	736990	1176	318370	3404
2020-12-31	Arizona	04	523829	8879	113653.37	7174064	0	3040595	4301
2020-12-31	Arkansas	05	225138	3676	51992.82	3011873	0	1379778	6656
2020-12-31	California	06	2307860	25965	155858.26	39346023	344	14210945	15958

Note: Only the first few out of 185 columns are shown here.

State-Level Correlation and Multi-Regression Analysis

```
# Creating percentages of population for relevant columns.

# It should be noted, the SVI data has percentage columns for their data already.
# However, the percentages are for each county, not state. The logic I use below is a simple
# way of getting percentages by state.

covid_states <- covid_states %>%
  mutate(across(starts_with("e_"), ~ (.x / e_totpop) * 100, .names = "pct_{.col}"))

# Selecting the percentages and our IncidenceRate column to analyze.

correlation_data <- covid_states %>%
  dplyr::select(IncidenceRate, starts_with("pct_"))

# Calculating the correlation matrix.

correlation_matrix <- cor(correlation_data, use = "pairwise.complete.obs")

# To see our strongest correlations, we'll now take the matrix data and turn it
# into a table which we can easily sort and display.

correlation_long <- as.data.frame(as.table(correlation_matrix))
top_correlations <- correlation_long %>%
  filter(Var1 == "IncidenceRate" & Var2 != "IncidenceRate") %>%
  arrange(desc(abs(Freq))) %>%
  slice_head(n = 10)
```

Table 2: Top 10 State-Level Correlations to Incidence Rates

Var1	Var2	Freq
IncidenceRate	pct_e_age17	0.60
IncidenceRate	pct_e_hburd	-0.45
IncidenceRate	pct_e_asian	-0.40
IncidenceRate	pct_e_twomore	-0.32
IncidenceRate	pct_e_aian	0.31
IncidenceRate	pct_e_uninsur	0.30
IncidenceRate	pct_e_nhpi	-0.30
IncidenceRate	pct_e_unemp	-0.29
IncidenceRate	pct_e_age65	-0.29
IncidenceRate	pct_e_noveh	-0.26

The three strongest positive correlations with COVID-19 incidence rates are

`pct_e_age17`: Estimated percentage of the population aged 17 and below.

`pct_e_aian`: Estimated percentage of the population identifying as American Indian/Alaska Native.

`pct_e_uninsur`: Estimated percentage of the population without health insurance.

The three strongest negative correlations are

`pct_e_hburd`: Estimated percentage of households with high housing cost burdens.

`pct_e_asian`: Estimated percentage of the population identifying as Asian.

`pct_e_twomore`: Estimated percentage of the population identifying as two or more races.

To further explore these relationships, I will construct multiple regression models for the top positive and negative factors. These models will help us determine whether the identified factors remain significant predictors of COVID-19 incidence rates when controlling for the effects of other variables.

```
pos_model <- lm(IncidenceRate ~ pct_e_age17 + pct_e_aian + pct_e_uninsur ,data = covid_states)
neg_model <- lm(IncidenceRate ~ pct_e_hburd + pct_e_asian + pct_e_twomore ,data = covid_states)
```

Table: Regression Coefficients for Positive Factors

```
(R-squared = 0.384, Adj R-squared = 0.345)
```

Term	Estimate	Std. Error	P-value
(Intercept)	-7.8042012	3.0009731	0.0124026
pct_e_age17	0.6485414	0.1500432	0.0000796
pct_e_aian	0.1255903	0.0982241	0.2073124
pct_e_uninsur	-0.0653885	0.1046446	0.5350831

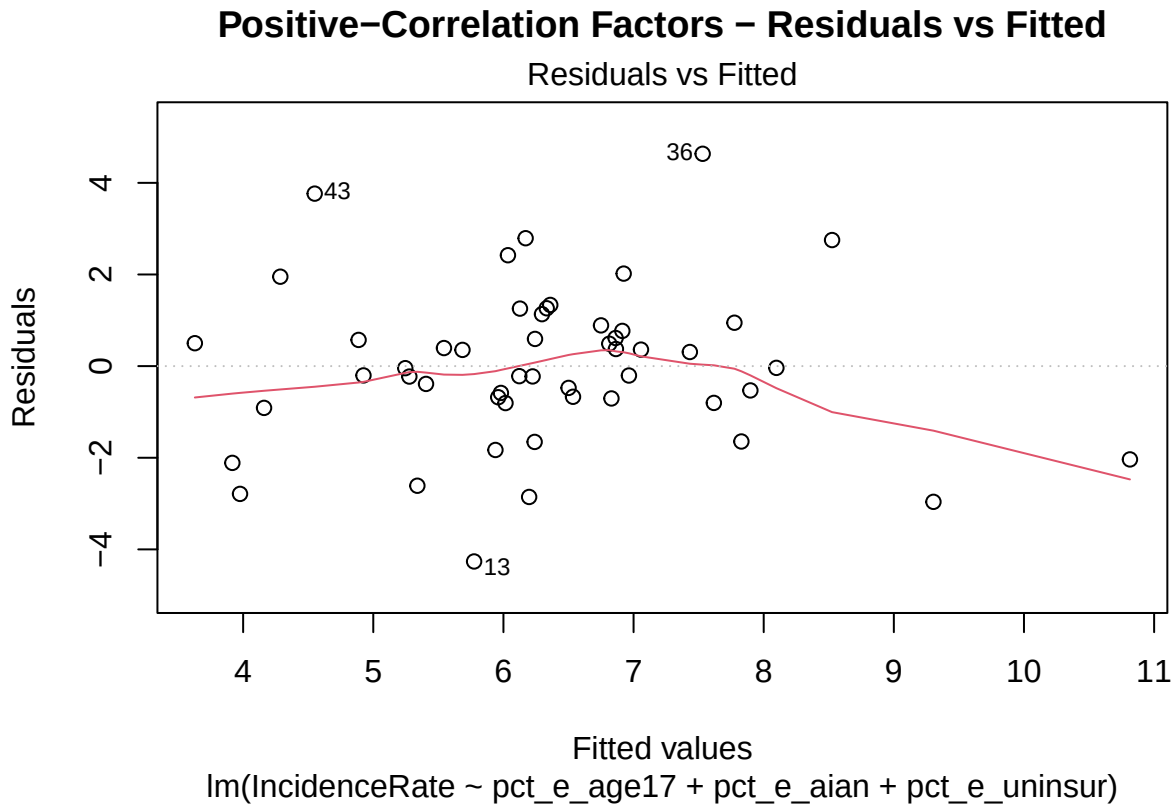
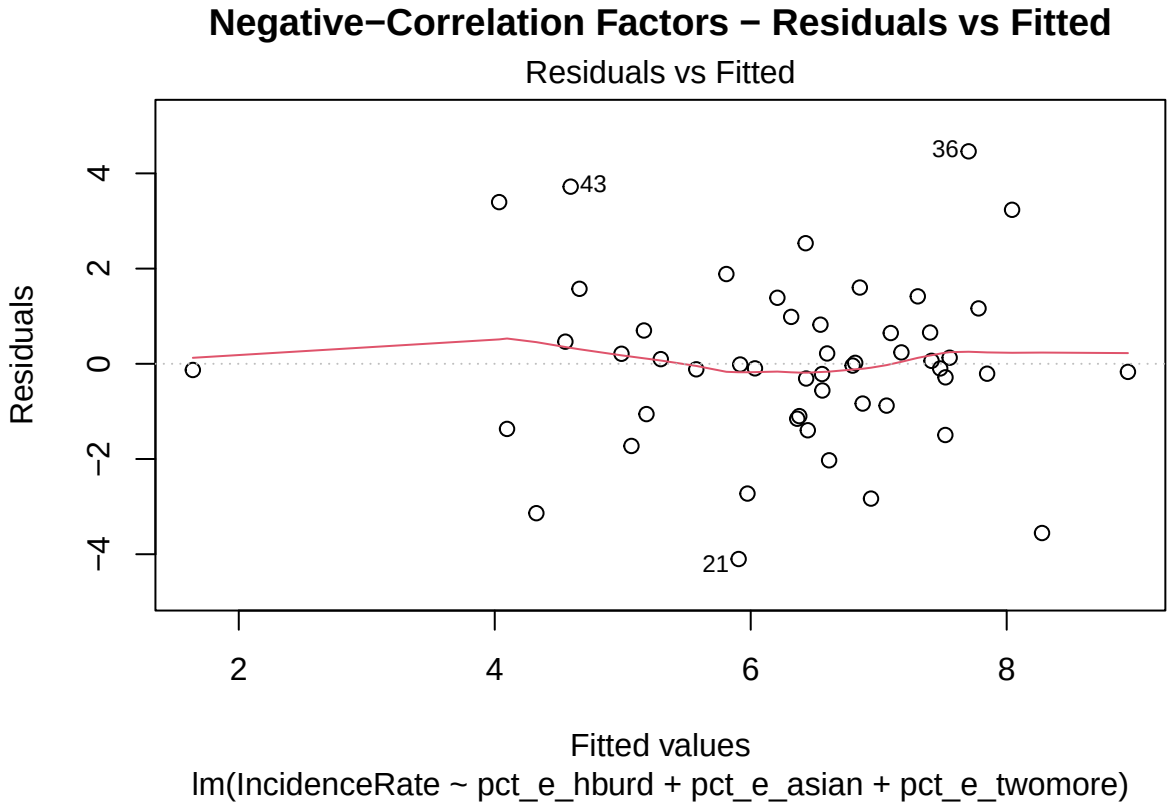


Table: Regression Coefficients for Negative Factors

```
(R-squared = 0.360, Adj R-squared = 0.319)
```

Term	Estimate	Std. Error	P-value
(Intercept)	17.0074084	2.6570271	0.0000001
pct_e_hburd	-0.9758394	0.2544291	0.0003724
pct_e_asian	-0.0469812	0.0935917	0.6180250
pct_e_twomore	-0.2489712	0.1983596	0.2156291



Positive Correlations Regression

`pct_e_age17`: This variable has a positive coefficient of 0.65, meaning that for every 1% increase in the population under 17, the incidence rate increases by 0.65%. This factor is the only statistically significant one from our analysis (p-value < 0.001), indicating it is a strong predictor.

The residuals graph shows the difference between observed and predicted incidence rates. A random scatter of residuals around zero suggests the model is appropriate for the data. Here, the residuals do not show strong patterns, indicating that the model assumptions (linearity, homoscedasticity) are mostly satisfied. However, some outliers call for further investigation.

These results suggest that younger populations likely contribute to higher COVID-19 incidence rates, potentially due to higher mobility, group interactions, or larger household sizes.

Negative Correlations Regression

`pct_e_hburd`: This variable has a significant negative coefficient (-0.98), meaning that for every 1% increase in households with high housing burdens, the incidence rate decreases by nearly 1%. This is the only statistically significant (p-value < 0.001) variable from our regression analysis.

The residuals appear randomly distributed around zero, supporting the appropriateness of the model. However, there is slight clustering of residuals near certain predicted values, which might indicate areas where the model could be refined.

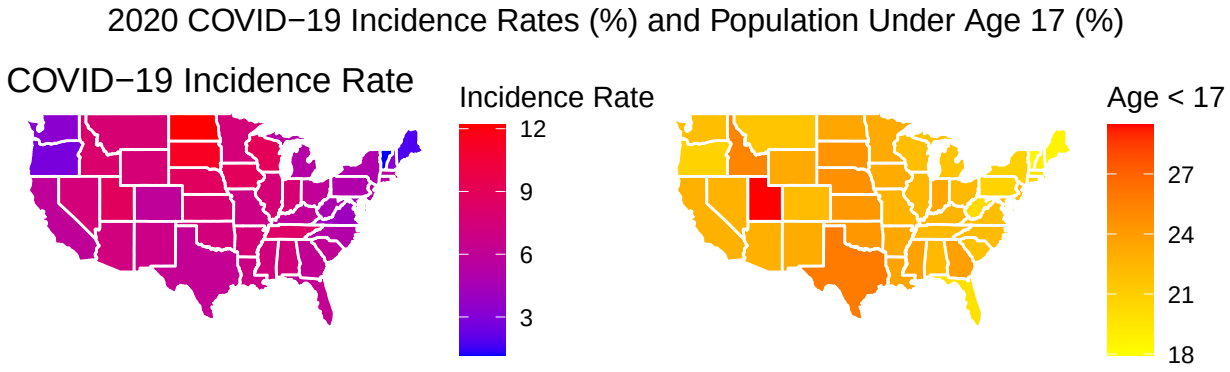
This suggests that areas with a higher proportion of households facing high housing cost burdens experienced lower COVID-19 incidence rates. This could reflect the reduced mobility or interaction of people living under economic struggles, or possibly that such areas already adopted stricter public health measures.

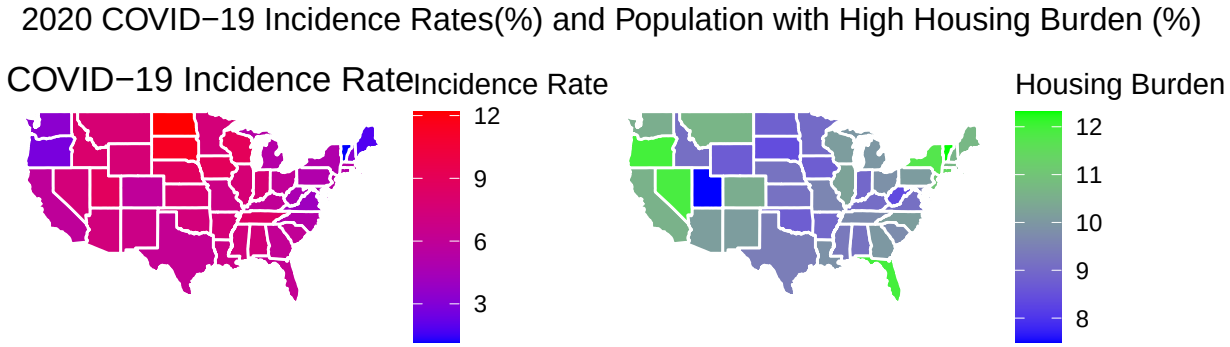
Low R-Squared Values

Both models have quite low R-squared values (~0.35), meaning only about 35% of the variability in incidence rates is explained by the predictors. This implies other variables, such as policy measures, healthcare access, or behavioral factors were causing bigger impacts on COVID incidence. However, among our dataset, we can reasonably say we've identified the dominant social factors influencing COVID rates.

Mapping Our Data

To visualize how these findings compare to the incidence rates, here are some choropleth maps.





County Level Analysis

To determine whether the social factors identified at the state level remain significant at more granular levels, I analyzed county-level data. By examining relationships at the county level, we can validate the bigger patterns observed and explore whether these factors maintain their influence when accounting for more local dynamics.

The analysis remains quite similar to the state-level analysis.

```
covid_counties <- read_csv("us-counties-all.csv")
svi <- read_csv("svi_interactive_map_2020.csv")

# Filter the COVID data for cases in 2020 only, then pick the totals (end of year).

covid_counties_2020 <- covid_counties %>%
  filter(year(date) == 2020)

covid_counties_2020 <- covid_counties_2020 %>%
  group_by(county) %>%
  filter(date == max(date)) %>%
  ungroup()

# Creating our combined dataset. Joining by both county and state as different
# states may have counties with the same name.
```

```r
colnames(svi) <- tolower(colnames(svi))

covid_counties_final <- covid_counties_2020 %>%
  left_join(svi, by = c("county", "state"))


# Creating a column for incidence rates by county.

covid_counties_final <- covid_counties_final %>%
  dplyr::mutate(County_IR = (cases / e_totpop) * 100)


# Selecting the percentages and our County_IR column to analyze.
# As mentioned earlier, the SVI data came with percentages of their variables by
# county, which we can now make use of.

county_correlation_data <- covid_counties_final %>%
  dplyr::select(County_IR, starts_with("ep_"))
```

```r
# Calculating the correlation matrix.

county_correlation_matrix <- cor(county_correlation_data, use = "pairwise.complete.obs")

# To see our strongest correlations, we'll now take the matrix data and turn it
# into a table which we can easily sort and display.

county_correlation_long <- as.data.frame(as.table(county_correlation_matrix))
top_county_correlations <- county_correlation_long %>%
  filter(Var1 == "County_IR" & Var2 != "County_IR") %>%
  arrange(desc(abs(Freq))) %>%
  slice_head(n = 10)
```

Table 5: Top 10 County-Level Correlations to Incidence Rates

Var1	Var2	Freq
County_IR	ep_age17	0.32
County_IR	ep_groupq	0.22
County_IR	ep_age65	-0.22
County_IR	ep_aian	0.20
County_IR	ep_hburd	-0.19
County_IR	ep_nohsdp	0.17
County_IR	ep_sngpnt	0.16
County_IR	ep_noint	0.15
County_IR	ep_asian	-0.14
County_IR	ep_pov150	0.14

County-level analysis supports the conclusions drawn at the state level, with overlapping factors such as younger populations (`ep_age17`) and housing burdens (`ep_hburd`) showing consistent correlations.

However, the relationships are generally weaker at the county level, signifying more variability coming from localized dynamics.

Also, the rise of new factors like `ep_groupq` highlights even more complexity added among finer levels of investigation.

Summary

This project focused on exploring how social factors influenced COVID-19 incidence rates during 2020, focusing on identifying the strongest predictors at both state and county levels.

To address this, the Social Vulnerability Index (SVI) data from the CDC was combined with COVID-19 case data at state and county levels. Correlation analyses identified the strongest relationships between social factors and incidence rates, while multiple regression models were used to control for confounding variables and refine the insights.

My analysis found that the percentage of the population under 17 (`pct_e_age17`) and the percentage of households with high housing burdens (`pct_e_hburd`) were the strongest social factors. Younger populations were associated with higher incidence rates, likely due to increased mobility and larger household sizes. High housing burdens correlated with lower rates, possibly due to lower mobility or stricter adherence to public health measures in poorer areas.

The findings showcase critical vulnerabilities that shaped the pandemic's spread, emphasizing the need for targeted public health strategies. Specific interventions focusing on demographic risk factors, like younger populations, can help mitigate future health crises.

The modest R-squared values indicate that many other factors, such as public health policies and healthcare infrastructure, were not accounted for in this analysis. Future work could incorporate these variables and explore alternative modeling techniques to better capture the complex dynamics at play. Additionally, temporal analyses could reveal how these factors evolved over the course of the pandemic.