This is an example report conducting fundamental statistical analysis.
You may find the formatting and analysis beneficial for our final project.

Project Report

1. Introduction

The source of our data is coming from a government website (Health Data) in which the COVID-19 Team and faculty from the White House create a Community Profile Report (CPR). The CPR is where data is provided and updated for it to be current. This data is useful because COVID-19 pandemic trends could be visualized and understood more easily. The source of the dataset is valid because it not only focuses on COVID-19 data and outcomes in the last seven days but it also reflects the most recent changes to the data. This dataset includes variables ranging from the county of interest to cases per 100k in the last 7 days and even extends to individuals who have the income status 'in poverty.'

Our group was very interested in seeing whether the deaths in the United States had a positive linear relationship with the American people who have the income status 'in poverty.' Since there is a global pandemic going on, this information intrigued us because it is important to be aware of the Social Vulnerability Index (SVI) and the 'in poverty' status in the United States. It is expected for people with the income status of 'in poverty' to have a higher SVI Score as they do not have the basic resources compared to the individuals not 'in poverty,' especially during a situation like the COVID-19 pandemic.

The specific variables that we included in our analysis are 'SVI Score' and 'In Poverty,' with the individuals being 'In Poverty' representing the explanatory variable and 'SVI Score' representing the dependent variable. As these variables are numerical, they represent the value of individuals who are more vulnerable, analyzed across the nation, which allows us to compare that data to the percentage of individuals in poverty. This will help us answer our posed question: whether or not there is a positive linear relationship between the SVI score and the individuals 'in poverty.'

Since we are in a global pandemic, it is crucial that we are aware of the communities that are more vulnerable. Our group expects that the individuals who have the income status 'in poverty' are more likely to have a higher SVI score. This could be due to the lack of resources, basic needs or insurance. The people who are living 'in poverty' are financially struggling thus, they will have an even harder time during the COVID-19 pandemic. As we progress as a nation, an important piece of information is whether or not the amount of vulnerability of certain groups significantly increases when compared against the number of individuals 'in poverty.' The way to receive a remedy for these people, and increase the chances of others helping those below the poverty line the ability to remove themselves from the never ending cycle of poverty, is to help visualize this clear cut data showing a significant correlation. Hence, everyone will be more educated and knowledgeable of the immense struggles that poverty coincides with, and the data is very straightforward representing a strong relationship between the two.

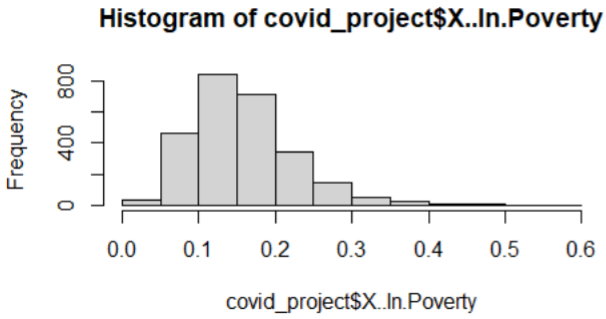
The linear regression model is appropriate to answer the research question because it will clearly indicate the correlation between the 'SVI Score' and the individuals living below the poverty line. Finding a strong linear regression model is important for determining the amount of residuals and the average of those values to explain the correlation between both variables. Once a linear regression model

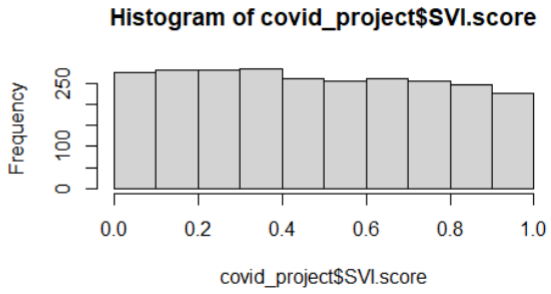
is found, we are more clearly able to see how affected individuals 'in poverty' become more socially vulnerable during a climate such as the current one. Between the correlation coefficient and the linear regression model, this allows anyone, even those not really interested in statistics or knowledgeable in the subject to understand how the two are associated. Although association is not the same as causation, methods to alleviate the increasing vulnerability can be implemented and if the COVID-19 pandemic ever recurs or something similar to this pandemic occurs, the numbers will hopefully be different, and those 'in poverty' will not be as greatly affected.

2. Data description

The first numerical variable our group chose was the SVI score and the second is the individuals in poverty (in percentage). The SVI score is the social vulnerability index or a score that is between zero and one demonstrating a community's level of vulnerability. A zero SVI score is the goal and an SVI score of one is the most severe. Having the highest vulnerability means that the community is at a very high risk for negative effects caused by natural disasters, diseases, or man-made disasters. The lowest vulnerability is what we are looking to achieve, and it represents a very strong community that could make it through the aforementioned issues if they occurred. The second variable, 'in poverty' indicates the individuals who have a very low income status, less than standard living conditions and facilities. The 'in poverty' variable is seen in percentage format which makes it easy to analyze.

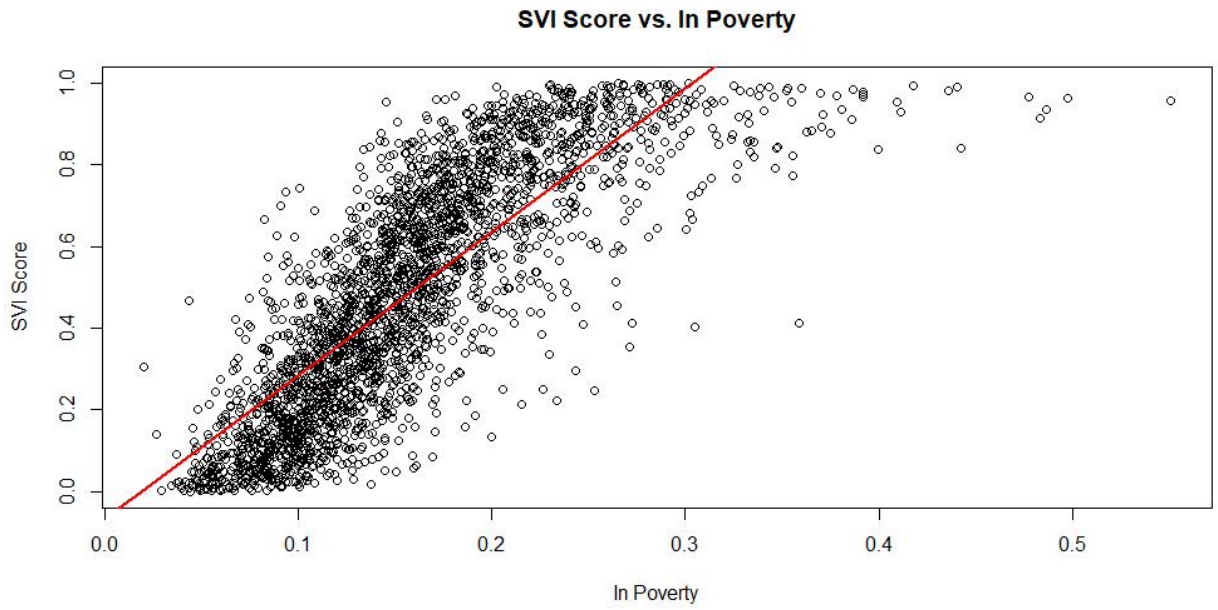
After determining what aspects of the data we were looking for and recognizing the meaning of their values, we input the data into two histograms, one for the 'in poverty' values and one for the "SVI scores." We expected to see a unimodal right skew for the percentage in poverty as one of our objectives for analyzing this data is that there is not enough attention paid to these impoverished communities, allowing the issues they face to escalate especially during a pandemic. By observing this histogram it is clear that the mean is less than 50% and generally if less than 50% of a population is dealing with something, it tends to go unnoticed and ignored. The histogram for SVI score is uniform with no significant modes or patterns. There is a very slight drop off along the latter half of the x-axis approaching the max SVI score. The fact that this histogram is uniform represents that there is a frequency for the most severe SVI score as there is for.





	In Poverty	SVI Score
Mean	0.1565772	0.4823273
Standard Deviation	0.06533502	0.2855045

Our data was plotted and the linear regression line is shown on the scatterplot:



The most efficient graphing method is a histogram because we do have continuous data. A barplot is too inclusive which makes it not the most efficient graphing method. Based on the graphs, we have a high frequency of counties with a poverty percentage between 5% and 25% with the highest number of counties having a poverty percentage between 10% to 15%. Moreover, the histogram for the SVI score indicates that each level of vulnerability analyzed in the data, between 0 and 1, has nearly the same number of counties correlated to it. The histogram for the SVI score is uniform.

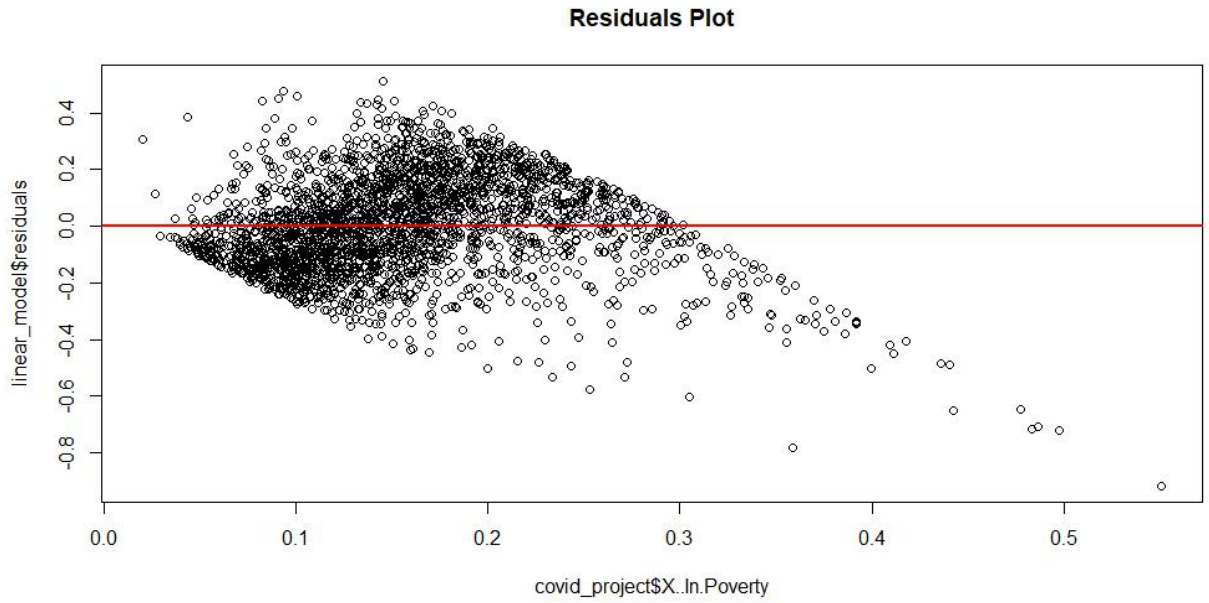
According to the scatter plot, there is a positive linear relationship between the SVI score and the percentage of people in poverty. The correlation coefficient is 0.806 which is a positive number hence, the direction of the graph is positive. The two variables are directly proportional and there is strong association.

3. Results and Interpretation

Our linear regression equation is:

$$\mathbf{y = 3.522558x - 0.069225}$$

The slope is 3.522558 and the intercept -0.069225. That is, as 1 unit increase in 'In Poverty' variable, we expect the SVI score to increase by 3.522558. The slope makes sense as there is a positive increase in the y value as x increases in the amount of our slope. The intercept fits the data and follows the trend, but when interpreted it does not make sense for there to be a negative SVI score when the in poverty value is 0. We checked our summary statistics and found out that our R square was 0.6498. In context, this means that 64.98% of the SVI score is explained by the percentage of people in poverty.



The residual plot above does show some notable pattern, which illustrates that the data does not nicely fit to the linear model.

4. Discussion

In our project we wanted to find out the relationship between the SVI scores in counties in the United States and the percentage of the population who have the status 'in poverty.' We found that 64.98% of the SVI score is explained by the percentage of people in poverty. Since it is a high association, our results show a strong positive linear relationship and we are able to conclude that the poverty level of a community is strongly related to the SVI score of that community. In conclusion, our project is significant because it provides us with information regarding the COVID-19 global pandemic. The more evidence-based knowledge we have, the more we will be able to aid in the overall health of people and our world.

Our results do make sense in the real world situation because studies have shown that low-income communities are more likely to contract COVID-19 than wealthier communities. A study also shows that impoverished communities face more adversity if they contract COVID-19 than others (Koma et. al, 2020). Wealthy communities have had greater opportunities to maintain a low SVI through means such as working from home, while low income communities make up the majority of essential workers who are consistently putting their lives at risk. These 'in poverty' communities have an overall more difficult

time during the pandemic, not to mention the lack of health resources available to them. Also, the few resources that you can find in these communities may not be available to everyone such as undoccumented individuals (Dott, 2021). All of these examples are a direct relation to the data we have observed. Down below we have listed a couple of articles, including one from the Center of Disease and Prevention (CDC), that support our results.

There are limitations in our project as there is always room for improvement. One possible limitation could be that all of our data is solely from observational studies and not experimental. This means that we have no insight as to whether there is a third variable present or not within our data, we have no control over the impact a third variable would have, and we are unsure of how great the confidence level is within these results versus the results we would get from having a control group.

Additionally, it would be helpful to know other information such as the safety practices that were/were not implemented in these two types of communities and the overall tendency of people to abide by the precautions. Being able to analyze this other data alongside the data for our two variables will further help us understand if the relationship we observe is as strongly correlated to our two variables or if a factor such as health practices has had an impact as well, altering our results.

Lastly, while the model still seems to be significant to explain the relationship between the two variables, the residual plot shows that there might be some of the patterns that are not fully explained by the linear model.

Overall, the data concludes that vulnerability for the portion of the population that is within the data grouping ‘In Poverty’ is highly likely to be much greater than the portion of the population living above the poverty line. This answers our question regarding the relationship between these two factors. From the data provided, and the observational and analytical measures taken, we can positively inform others of the relationship, and perhaps influence the changes necessary to bring some relief to these communities.

Reference

Dott, Mary. “Health Equity Considerations and Racial and Ethnic Minority Groups.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Apr. 2021, www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html.

Koma, Wyatt, et al. May 2020. “Low-Income and Communities of Color at Higher Risk of Serious Illness If Infected with Coronavirus.” KFF, 7 May 2020, www.kff.org/coronavirus-covid-19/issue-brief/low-income-and-communities-of-color-at-higher-risk-of-serious-illness-if-infected-with-coronavirus/.