Stats 10 Final Project Report

Group #10: Antwan Adams, Sarah Wang, Damien Ha, Anita Wong

I. Introduction

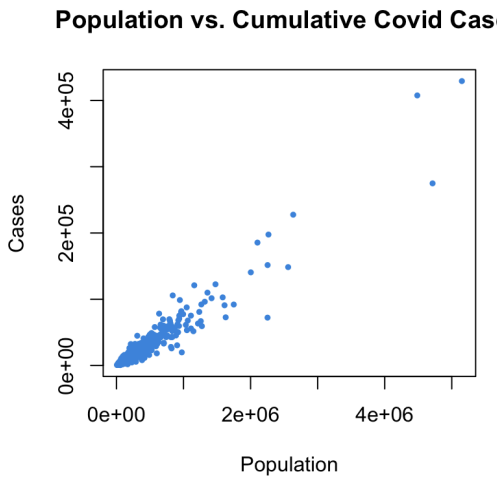
The dataset was obtained from U.S. COVID-19 Community Profile Report (CPR), sourced from the White House COVID-19 team and last updated on March 17, 2021. The CPR provides information for all regions, states, core-based statistical areas (CBSAs), and counties in the United States mainly framed in the context of COVID-19 outcomes in the last seven days at each geographical level. The variables in the data include county type, core-based statistical area type, deaths per 100k in the last seven days, cases per 100k in the last seven days, cases as a percentage of the national total in the last seven days, population total per county, and cumulative case totals per county. Although the data is calculated using standard metrics, the collected data may differ from data displayed on state and local websites due to differences in data reporting or metric calculations.

Our objective is to determine whether or not the population size of a county is correlated with the cumulative number of COVID-19 cases in that county. Therefore, the independent explanatory variable is the population size of the county and the dependent response variable is the cumulative number of cases per county. More specifically, we believe that our analysis of the data will reflect a positive correlation between population size and the number of cumulative cases in each county. This is a reasonable hypothesis because greater population size, especially in areas labeled “large central metro” or “large fringe metro” where density is greater, will often result in more frequent person-to-person contact and potential opportunities for the spread of COVID-19. By answering this research question, public health officials can learn more about what causes COVID-19 case numbers to increase and then take preventative measures to slow the spread of the virus.

A linear regression model is appropriate to answer the research question because it will determine whether or not there is a positive correlation between the cumulative number of cases and population size per county. We hypothesize that as the population size (explanatory variable) increases, the cumulative number of cases (response variable) increases. If we overlay the linear regression line over the scatter plot graph, we can visualize how closely the data points align with the positive slope.

II. Data description

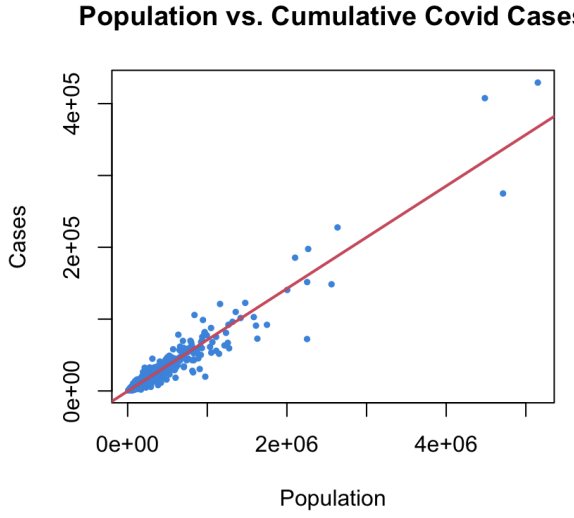
The first numerical value is the population count per county. The population mean is about 265678 people per county and the standard deviation is about 440168.4. The second numerical value is the number of cumulative COVID-19 cases per county at the time of reporting. The cumulative cases mean is about 18622 cases per county and the standard deviation is about 32740.5 cases.



The distribution of the variables are correlated in a strong positive association. It is roughly symmetric on both sides of the linear regression line and has strong linearity. The majority of the data points are clustered close to the origin and become more scattered as x and y increase. Toward the higher end of the graph, the linearity decreases and we observe a few minor outliers in the data.

The scatterplot above displays the relationship between the cumulative number of cases per county, the response (dependent) y-variable, and the population per county, the explanatory (independent) x-variable. The correlation coefficient between these two variables was found to be 0.9602773, indicating that they have a relatively strong positive linear relationship.

In this case, we can interpret this to mean that the cumulative number of cases is strongly correlated with the population of each county, though we cannot infer causation. This graph demonstrates that the number of COVID-19 cases increases proportionally with the county's population.



III. Results and interpretation

Below is a table of our intercept and slope values:

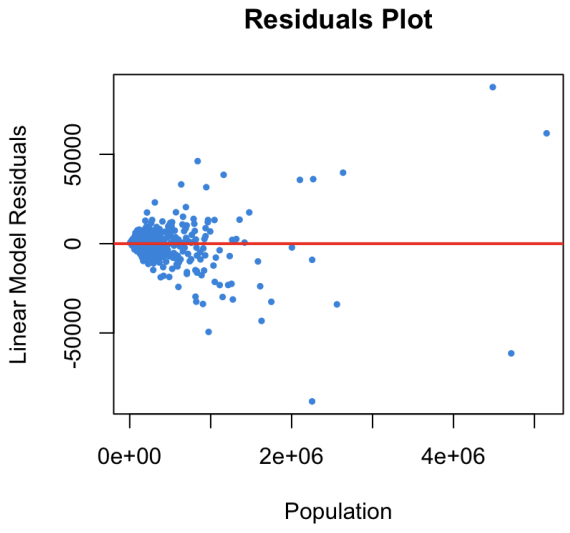
Parameter	Estimate
Intercept	-0.03548
Slope	0.07143

Our final regression equation is as follows:

$$y = 0.07143x + (-0.03548)$$

With y representing the cumulative number of COVID-19 cases of the county and x representing the population of the county. For every additional person in a given county, the number of COVID-19 cases increases by about 0.07143 cases. The negative intercept tells us that where the linear model predicts revenue (y) would be when subs (x) is 0. The negative intercept means that if the population were 0, the predicted number of covid cases would be -0.03548.

According to our linear regression analysis, there is a positive relationship between the number of people and cumulative cases in the county, with a slope of 0.07143. Therefore, for future observations, we can predict that with every increase in population, there would be an increase of roughly 0.07143 cases. For example, if a county had a population of 5000, the number of cumulative covid cases will be roughly 357.11452.



Based on the residual plot, there appears to be linearity and symmetry, but not equal variance. A good residual plot has values randomly clustered around zero, but for our residual plot, the points are clustered closer to zero for smaller populations and fan out for higher populations. So, our model has the concern of not meeting equal variance based on our residual plot.

From the data summary we computed, our standard error and R^2 values are as follows:

Residual standard error: 9142 on 768 degrees of freedom
Multiple R-squared: 0.9221, Adjusted R-squared: 0.922
F-statistic: 9095 on 1 and 768 DF, p-value: < 2.2 x 10^{-16}

Our R^2 value is about 0.922, meaning that approximately 92.2% of all variation in the cumulative number of cases/population ratio can be explained by our model. In this study, 92.2% of all variation in the number of cumulative COVID-19 cases can be explained by the population size per county.

IV. Conclusion

During this global pandemic, data analysis is crucial for scientists and public health leaders to identify and prevent causes of viral spread. Knowing that COVID-19 is most commonly spread from direct person-to-person contact as opposed to via surfaces or in-building ventilation systems, we predicted that areas with greater population sizes within the same relative geographic area range would have a greater number of COVID-19 cases. Therefore, among the fifty variables included in the Community Profile Report, we selected two variables that we thought to have great significance: the cumulative number of cases per county and the population size per county.

We were not surprised to find that there was a strong positive correlation between the number of cases and population size, which was confirmed by the dataset's correlation coefficient of 0.9602773. For an average county population of 265,678 individuals, there were approximately 18,622 confirmed cases of COVID-19.

The positive and strong correlation between the cumulative number of cases with the population of each county can directly be tied to an example of Los Angeles county in 2020 and 2021 so far. The bigger the county and the higher population means more dense communities and compact areas leading to an easy spread of the COVID-19 virus which then can increase the cumulative number of cases. This factual evidence from a real world situation helps reinforce our findings from this project study and so, our result makes sense in reality as well as hypothetically.

While the metrics used in this report were largely standardized, some limitations of this project include differences in reporting from county to county. For example, there may be inaccuracies in census counting of population sizes per county. Additionally, due to inconsistencies in the rollout of COVID-19 tests and faulty administration of tests, the confirmed cases per county could have been undercounted through lack of surveillance testing and underreported.

It is likely that there are a significant number of unconfirmed COVID-19 cases, especially for individuals who are asymptomatic, and therefore the cumulative number of COVID-19 cases per county could be underrepresented. Additionally, there are a few outliers in our residuals plot that could be potentially skewing our overall dataset and data analysis.

Lin, Rong-Gong & Smith, Hayley (2021, February 25). *L.A. County's grim discovery: 806 new winter COVID deaths*. Los Angeles Times. https://www.latimes.com/california/story/2021-02-25/la-county-discovers-806-new-winter-covid-deaths

Martinez, Lita & Pollack, Gina (2021, March 1). *LA County's Daily Coronavirus Numbers Are Finally Back Down To Pre-Surge Levels*. LAist. https://laist.com/latest/post/20210301/LA-county-daily-coronavirus-case-numbers-hit-pre-surge-levels-decline-vaccinations

COVID-19 Locations & Demographics - LA County Department of Public Health. (2021, March 19). L.A. County Public Health. http://publichealth.lacounty.gov/media/coronavirus/locations.htm