Analyzing effects of educational attainment on COVID-19 mortality rates

Taarika Hegde

***Abstract*—** Studies on COVID-19 mortality experienced by different demographics have focused on inequities among different ethnic, race, and age populations in combination with educational attainment. It is well-documented that there exists significant differences in absolute COVID-19 mortality between college educated and non-college educated groups. The objective of this study was to therefore understand whether the relative COVID-19 mortality, as calculated as a ratio of COVID-19 deaths to Total Deaths per each combination of age, sex, race is also dependent on educational attainment level. The study analyzed data on the U.S. adult population for the 2020 calendar year stratified by educational attainment levels (college educated and non-college educated), using a Two-Sample T-test and multiple linear regression model. The conclusions from the analysis showed no significant differences between the means of COVID-19 attributable deaths to Total Deaths between educational attainment groups and also showed that educational attainment was not a meaningful indicator of predicting relative mortality rates attributed to COVID-19.

***Keywords*—** COVID-19, educational attainment, health outcomes and disparities

I. Introduction

The coronavirus disease 2019 (COVID-19), a disease caused by the novel coronavirus SARS-CoV-2, was the third leading cause of death in the United States for most of 2020 and briefly became the leading cause of death in the United States in early 2021. Early observational studies showed that the severity of COVID-19 across patient populations was not only dependent on the physical characteristics of patients including age, gender, and health, but also on the socioeconomic conditions of patients including educational attainment levels [1].

The purpose of this analysis was to further investigate the relationship between educational attainment levels and COVID-19 severity, as measured by mortality, in the U.S. adult population. The hypothesis assessed in this study was that college educated populations would experience lower relative mortality rates attributed to COVID-19 than non-college educated populations.

The following sections of this paper provide background on related studies, information on the dataset used in analysis, an overview of the statistical methods applied, and lastly, the results and conclusions of the findings.

II. Background

Prior studies regarding COVID-19 mortality inequities have focused on understanding the relationship between demographic factors, like race and sex, in combination with socioeconomic factors, like educational attainment. One study, concluded that ethnic minorities, including within the same educational attainment levels, had higher COVID-19 mortality rates than non-Hispanic White populations [2]. Another study looked at mortality rate changes by bachelor's degree status specifically between 2019 and 2020 and concluded that the effects of having a bachelor's degree was more profound on absolute risk and had minimal effect on relative risk within different demographics. The study documents that those with BA degrees are more likely to work in jobs that allow for telework as opposed to non-BA holders where jobs are primarily in in-person service industries. As the chance of infection was greater among occupations with high in-person contact, differences in absolute risk could be easily explained among college educated and non-college educated groups [3].

Since absolute differences in COVID-19 mortality between college educated and non-college educated populations was well documented, this analysis took a different approach by looking specifically at differences between COVID-19 deaths to Total Deaths for each demographic to determine the effects educational attainment had on COVID-19 attributable deaths as a whole.

III. Data

The dataset used for this analysis was taken from the Centers for Disease Control and Prevention (CDC) provided by the National Center for Health Statistics [4]. The open access dataset provides cumulative provisional counts for COVID-19 deaths and Total Deaths in the U.S. by educational attainment, race, sex, and age group for the full calendar year of 2020. The three categories reported for educational attainment include: high school or General Educational Development (GED) certification or less, associate's degree or some college, and bachelor's degree or more. It is noted that the dataset is provisional because there could be undercounts in later weeks due to reporting delays. Since the data was current as of May 2021, the effect of reporting delays was considered minimal.

IV. Method

Two different methods were used to analyze the effect of educational attainment on mortality rates attributed to COVID-19, a two sample T-test and a multiple linear regression model.

A. Two Sample T-Test

In the Two Sample T-Test, the dataset was first split into two sets, Set 1 consisted of college educated mortality rates and Set 2 consisted of non-college educated mortality rates. College educated was defined as being the educational attainment categories of associate's degree or some college and bachelor's degree or more. Non-college educated was defined as being the educational attainment category of high

school, GED, or less. The age range of 0 to 17 was removed from the dataset to account only for the U.S. adult population. In addition, to determine what percent of mortality for each demographic (age, race, and sex) could be attributed to COVID-19, a ratio was taken of COVID-19 deaths to Total Death as provided in the data source. This was done to account for differences in population size and mortality rates between different demographics. These ratios for each set were then used in the Two Sample T-test.

The data was tested for normality by plotting a QQ-plot. After confirming the data was approximately normally distributed, the T-test was conducted. The *t.test* function in R was used to run a one-sided Welch's Two Sample T-test where the alternative Hypothesis was set to 'less'. The T-test allows one to test the null hypothesis that the means for two sets are equal. In this case, the null hypothesis was that the means for the COVID-19 deaths to Total Deaths ratios was equal for the college educated and non-college educated sets. The alternate hypothesis was that the college educated set had a mean COVID-19 death to Total Deaths ratio less than the non-college educated set.

B. Multiple Linear Regression Model

In the multiple linear regression method, the dataset was split into college educated and non-college educated categories as described in the Two Sample T-test method. The educational levels were treated as independent variables and the COVID-19 deaths to Total Deaths ratio was treated as the dependent variable. The *lm* function in R was used to run a multiple linear regression of the data. A diagnostic test of the residuals was performed to confirm the validity of the results.

V. Results and Findings

A. Results of Two Sample T-test

First, the normality of the data was assessed using a QQ-normal plot. The results for the college educated and non-college educated QQ-normal plots are shown in Fig.1 below.

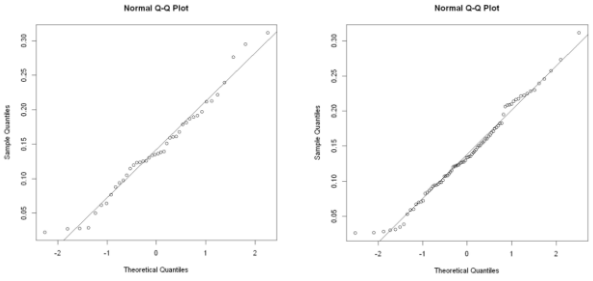

Fig. 1 Normality plots for college (left) and non-college (right)

Both plots show the mortality ratios for both groups are approximately normally distributed since they follow straight lines. Due to the normality of the data, hypothesis testing was determined to be an appropriate method for analysis.

The two sets were then compared using a Two Sample T-test. The summary of results from the output are shown in Table 1.

```
        Welch Two Sample t-test

data:  set1 and set2
t = -0.30173, df = 74.697, p-value = 0.3818
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.01742937
sample estimates:
mean of x mean of y
0.1384994 0.1423556
```

Table 1. Output of Two Sample T-test

The mean of the college educated set was .1385 and the mean for the non-college educated set was .1424 indicating a lower ratio of deaths attributed to COVID-19 among the college educated set. However, the p-value from the results is 0.3818, which is not statistically significant. Additionally, the 95th percent confidence interval contains the null value, 0, showing there is no statistically meaningful difference between the groups.

It can be concluded that at the 95th percent confidence level, we accept the Null Hypothesis that the difference in means between the mortality ratios of the two sets (college educated and non-college educated) is not statistically significant. We can reject the Alternate Hypothesis that Set 1 (college educated population) has a mean COVID-19 to Total Deaths mortality ratio less than Set 2 (non-college educated population).

B. Results of Multiple Linear Regression

A multiple linear regression was run on the college educated and non-college educated groups from the dataset, denoted as 'Status'. The formula for the linear regression was as follows: formula = COVID-19 to Total Deaths ratio ~ Status. The results are shown in Table 2.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.120300 -0.043688 -0.005472  0.040356  0.173114

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.138499   0.007116  19.462   <2e-16 ***
StatusNon-College Educated 0.003856   0.012326   0.313    0.755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06522 on 124 degrees of freedom
Multiple R-squared:  0.0007887, Adjusted R-squared:  -0.007269
F-statistic: 0.09788 on 1 and 124 DF,  p-value: 0.7549
```

Table 2. Output of multiple linear regression

Additionally, the linear regression was plotted, and the output is shown in Figure 4 below.

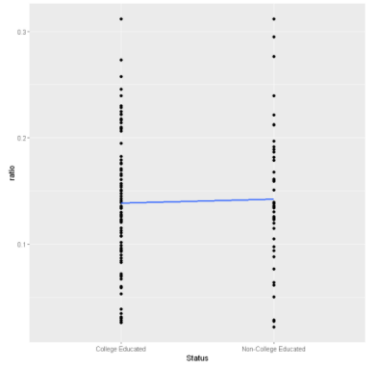

Fig. 4 Linear regression of educational attainment

The p-value from the regression was 0.7549, which is not less than the statistically significant value of 0.05. This indicates that changes in the predictor, educational attainment status, are not associated with changes in the response, COVID-19 mortality ratio. Additionally, referring to Fig. 4, the line for the linear regression only shows a slight increase between college educated and non-college educated groups' predictors. The R-squared value is also close to 0 which indicates that the explanatory variables are not good predictors of the response variable.

Diagnostic plots of the residuals were plotted in order to check normality assumptions. The diagnostic plots are shown below in Figure 5 and, overall, they show the model meets normality assumptions.

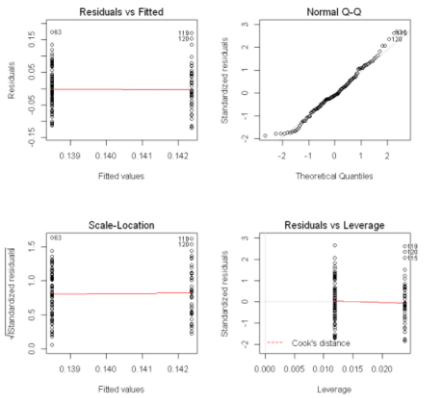

Figure 5. Diagnostic plot of residuals

VI. Conclusion

Overall, the results from the two analytic methods conducted do not show that there is a statistically significant difference between the means of Covid-19 deaths to Total Deaths of college educated and non-college educated US adult populations.

The Covid-19 deaths to Total Deaths ratio was used in order to account for the differences between the overall population sizes of college educated and non-college educated groups as well as the overall mortality inequalities between the two groups. The ratio was calculated using the COVID-19 death counts divided by the Total Deaths for each demographic. However, this may not have been the best measure of COVID-19 severity. This measure shows the attributability of COVID-19 to the overall death counts for each population group, but does not show the standalone COVID-19 rate based on overall population sizes. This study also looked at the entire adult population (18 years or older) including all races, genders, and age groups. As older populations tend to have higher mortality rates regardless of socioeconomic factors, more specific comparisons between younger and middle-aged populations may have been more meaningful. Similarly, specific comparisons between combinations of race, gender, and age would have likely shown greater differences between educational attainment and COVID-19 attributed mortality.

As the COVID-19 pandemic is still ongoing and research continues to be conducted on the effects of educational attainment on COVID-19, these preliminary results suggest that educational attainment alone is not a strong predictor for COVID-19 deaths to Total Deaths per demographic. These results do not negate educational attainment as a significant predictor of absolute COVID-19 mortality rates, but rather emphasize that relative to the total number of deaths experienced by both groups, college and non-college educated groups experienced approximately the same ratio of COVID-19 attributed deaths to Total Deaths per group.

References

[1] Ortaliza, Orgera, Amin, Cox. (2021, October 13). *COVID-19 continues to be a leading cause of death in the U.S. in September 2021*. Peterson-KFF Health System Tracker. https://www.healthsystemtracker.org/brief/covid19-and-other-leading-causes-of-death-in-the-us/

[2] Feldman JM, Bassett MT. Variation in COVID-19 Mortality in the US by Race and Ethnicity and Educational Attainment. *JAMA Netw Open*. 2021;4(11):e2135967. doi:10.1001/jamanetworkopen.2021.35967

[3] Chen JT, Testa C, Waterman P, Krieger N. Intersectional inequities in COVID-19 mortality by race/ethnicity and education in the United States, January 1, 2020–January 31, 2021. Harvard Center for Population and Development Studies Working Paper Volume 21, Number 3, February 23, 2021.

[4] Centers for Disease Control and Prevention. AH deaths by educational attainment, 2019-2020. Published February 24, 2021. Accessed December 5, 2021. https://data.cdc.gov/NCHS/AH-Deaths-by Educational-Attainment-2019-2020/4ueh-89p9