How Pre-existing Socio-economic Factors of U.S. Counties Relate to the Spread of COVID-19

Paras Ahuja
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD, USA
pahuja2@jh.edu

***Abstract*—On October 13, 2020, the World Health Organization (WHO) along with other organizations published a joint statement about the COVID-19 pandemic, "tens of millions of people are at risk of falling into extreme poverty while the number of undernourished people, currently estimated at nearly 690 million, could increase by up to 132 million by the end of the year." In addition, nearly 3.3 billion people are at risk of losing their livelihoods [1]. Although the statement by the WHO and others discusses what could happen because of the pandemic, this paper examines pre-existing socio-economic factors and their relation to COVID-19. Multivariable linear regression models are used to study and understand select features from 2018 U.S. Census data and their relation to COVID-19 cases. Number of COVID-19 cases is treated as the dependent variable in all instances of analysis. We noticed that socio-economic factors play a part in explaining COVID-19 cases. Factors like age, race, and employment status were found to have the highest impact. Throughout the U.S. racial background of citizens can explain 94% of the cases. On a state specific level, age, employment status and poverty were key factors. In Arizona, population under the age of 18 and number of COVID-19 cases have a correlation greater than 0.98 and simple linear regression can be used to explain 99% of the cases.**

***Index Terms*—COVID-19, Linear Regression, Socio-economic factors, Pandemic**

I. Introduction

The U.S. has the highest number of COVID-19 cases in the world. At the time of writing, there are 16,074,511 active COVID-19 cases in the U.S., and 297,864 Americans have lost their lives because of the virus [2]. Cost of the pandemic to the U.S. economy is estimated to top $16 trillion [3]. Due to the pandemic, approximately 42% of the U.S. labor force is now working from home full-time; 33% of those who were in the workforce prior to the lockdown are no longer working; and 26% of the people are essential service workers and continue to work [4]. Since the pandemic started, 46% of lower-income adults have had trouble paying bills, and 33% of lower-income adults say that they are facing difficulties making rent or mortgage payments. However, only 20% of the middle-income adults state that they have experienced financial hardships [5]. This raises some key questions, which this work aims to answer:

– RQ1: Is there are correlation between pre-existing socio-economic factors and COVID-19 cases, and which factors have the highest correlation?
– RQ2: Do socio-economic factors differ by state and how might this information aid state and local government responses?

II. Background

A. Related Work

Existing work studies the impact of COVID-19 on a much broader scale, and utilizes different statistical methods and datasets. Work by R.B. Hawkins, E.J. Charles, and J.H. Mehaffey studies similar topics, however their research utilizes hierarchical linear modeling to study the data. Hawkins, Charles, and Mehaffey conclude in their research that: "lower education levels and greater percentages of black residents are strongly associated with higher rates of both COVID-19 cases and fatalities. Socio-economic factors should be considered when implementing public health interventions to ameliorate the disparities in the impact of COVID-19 on distressed communities" [6].

B. Multiple Linear Regression

This research uses multiple linear regression methods to study relationships in the data to:

1) Identify the strength of the effect that socio-economic factors have on COVID-19 cases.
2) Understand how the COVID-19 cases will change when socio-economic factors change.
3) Get point estimates.

Socio-economic factors are treated as explanatory variables, and the number of COVID-19 cases is considered to be the dependent variable. Linear model should help us define:

$$\text{cases} = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n + \epsilon \quad (1)$$

where x_i is a socio-economic factor, β_i is a parameter to estimate, and ϵ is the error term.

The fit of the multiple linear regression model is measured by R^2. Generally, the higher the number of variables we have, the higher the amount of variance we can explain. The goal for this research is to select a few important socio-economic factors to ascertain if correlation is present. This is because even if each variable doesn't explain much, adding a large number of variables can result in very high values of R^2. Therefore, adjusted R^2 is used because it provides the ability to compare regression models.

III. Dataset

The dataset is obtained from the Johns Hopkins University, and it was made available by Dr. Ian McCulloh, and Dr. Anthony Johnson. It consists of Census county level data.

A. Addition to the Data and Data Pre-processing

More data was collected and this was added the original dataset. Specifically, County Population by Characteristics: 2010-2019 dataset was obtained from the Census Bureau website. From this dataset, total population, total number of males, and the total number of females were extracted for each county and added to the original dataset.

Since the dataset had neither the number of COVID-19 cases nor the number of COVID-19 related deaths for each of the counties, this information was obtained from USAFacts, and added to the dataset for analysis.

Rows that contained missing information were excluded from analysis. For this reason, this research does not include information from the State of Alaska.

IV. Socio-economic Factors and COVID-19 Cases

With regards to RQ1, evidence suggests that there is a U.S.–wide correlation between socio-economic factors and COVID-19 cases. This is not to say that one is solely responsible for the other, but simply that there is strong association between the two. There is no uniformity in socio-economic factors. Furthermore, a factor that may be associated with the rise in COVID-19 cases in one county, may not do the same for the other counties.

A. Correlation

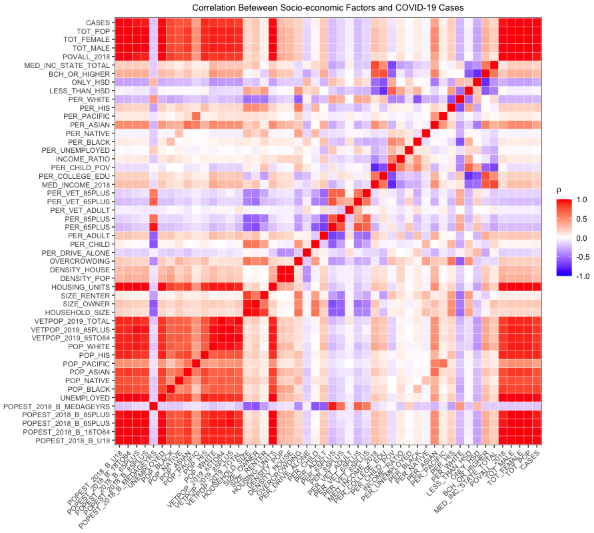


Fig. 1. COVID-19 cases and socio-economic factors correlation heatmap

Looking at the correlation heatmap above, in the very first row, we notice some strong correlation between socio-economic factors and COVID-19 cases. In particular we notice that there is a strong correlation with age but not the category of age (e.g. middle-aged). There is also correlation with race, and poverty levels.

Overall this research found the following socio-economic correlation with the COVID-19 cases:

TABLE I
Socio-economic Factors and Correlation with COVID-19

Factor	Avg. ρ
Age	0.9440
Race	0.7323
Income	0.2089
Family Size	0.1658
Employment Status	0.9517
Education	0.0351

Research suggests that gender may play a role in socio-economic well being of an individual [8]. There are other such factors like poverty levels in an area. Therefore, these factors are included as socio-economically associated factors. Associated factors are as follows:

TABLE II
Associated Factors and Correlation with COVID-19

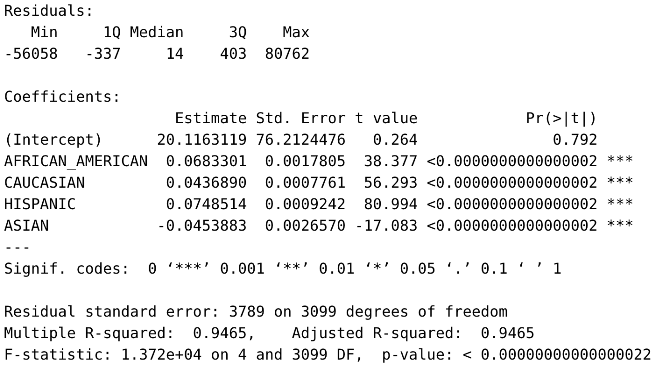
Associated Factors	Avg. ρ
Gender	0.9565
Poverty Levels	0.9380
Housing	0.9532

We notice that factors like age, employment status, gender, poverty levels, and housing have the highest association with COVID-19 cases. Within each of these factors, highest correlation is amongst the following:

TABLE III
Associated Factors and Correlation with COVID-19

Factors	Categories	ρ
Age	Under 18	0.9560
	18 to 64	0.9550
	65 to 84	0.9405
	85 and above	0.9243
Race	African American	0.7900
	Native American	0.7248
	Asian	0.7299
	Pacific	0.3986
	Hispanic	0.8927
Income	Median Income	0.2089
Family Size	Household Size	0.1658
Employment Status	Unemployed	0.9517
Education	Less than High School Diploma	-0.0283
	Only High School Diploma	-0.2793
	College or Associates Degree	0.1673
	Bachelors or Higher	0.2807
Gender	Male	0.9557
	Female	0.9574
Poverty Levels	Number of People in Poverty	0.9380
Housing	Number of Housing Units	0.9532

Research shows that at a macro-level racial factors can explain approximately 94% of the data as seen in Fig. 2. We notice here that the Adjusted R^2 is 0.9465 with a $P < .001$ indicating a highly statistically significant model.



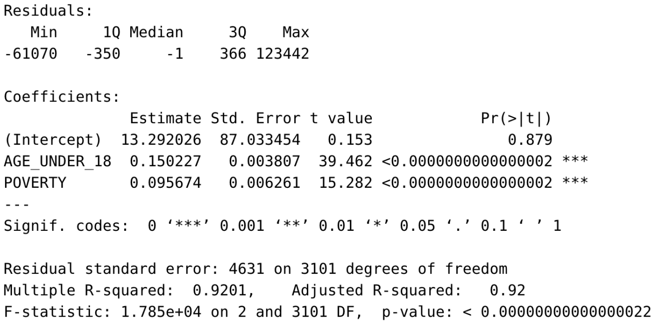
```
Residuals:
   Min     1Q Median     3Q    Max
-56058   -337     14    403  80762

Coefficients:
                   Estimate Std. Error t value            Pr(>|t|)
(Intercept)      20.1163119 76.2124476   0.264               0.792
AFRICAN_AMERICAN  0.0683301  0.0017805  38.377 <0.0000000000000002 ***
CAUCASIAN         0.0436890  0.0007761  56.293 <0.0000000000000002 ***
HISPANIC          0.0748514  0.0009242  80.994 <0.0000000000000002 ***
ASIAN            -0.0453883  0.0026570 -17.083 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3789 on 3099 degrees of freedom
Multiple R-squared:  0.9465,    Adjusted R-squared:  0.9465
F-statistic: 1.372e+04 on 4 and 3099 DF,  p-value: < 0.00000000000000022
```

Fig. 2. Race based linear model

Race is not the only factor of concern here. Even though racial background of citizens may have an impact on COVID-19 cases, other factors like age, gender, poverty, employment status have a higher correlation to the COVID-19 cases and are better studied as seen below in Fig. 3 and Fig. 4.



```
Residuals:
   Min     1Q Median     3Q    Max
-61070   -350     -1    366 123442

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept)  13.292026  87.033454   0.153               0.879
AGE_UNDER_18  0.150227   0.003807  39.462 <0.0000000000000002 ***
POVERTY       0.095674   0.006261  15.282 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4631 on 3101 degrees of freedom
Multiple R-squared:  0.9201,    Adjusted R-squared:   0.92
F-statistic: 1.785e+04 on 2 and 3101 DF,  p-value: < 0.00000000000000022
```

Fig. 3. COVID-19 cases and socio-economic factors correlation heatmap

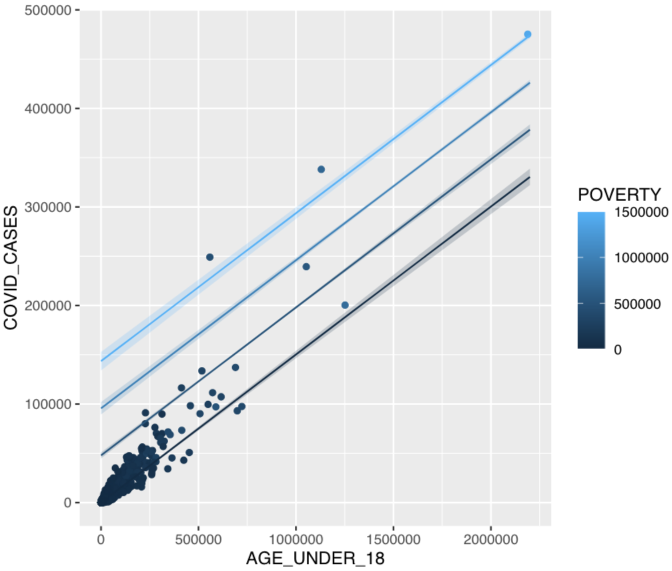


Fig. 4. U.S. Multiple linear regression model prediction

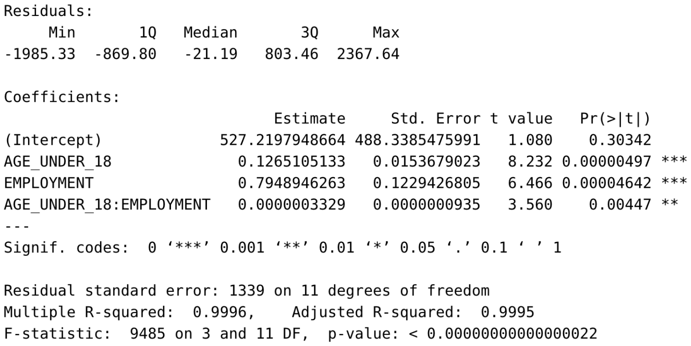
Research shows that 92% of the data can be explained by the poverty levels and population under the age of 18. We notice an Adjusted R^2 of 0.92 with statistical significance similar to race.

V. Socio-economic Variation Across the U.S.

With regards to RQ2, there is evidence that socio-economic factors impact residents of different states differently. However, there are some common elements. Consider the data and model for the states of Arizona and Kansas.

A. *Arizona*

In Arizona the model is as follows:



```
Residuals:
     Min       1Q   Median       3Q      Max
-1985.33  -869.80   -21.19   803.46  2367.64

Coefficients:
                         Estimate     Std. Error t value   Pr(>|t|)
(Intercept)          527.2197948664 488.3385475991   1.080    0.30342
AGE_UNDER_18           0.1265105133   0.0153679023   8.232 0.00000497 ***
EMPLOYMENT             0.7948946263   0.1229426805   6.466 0.00004642 ***
AGE_UNDER_18:EMPLOYMENT  0.0000003329   0.0000000935   3.560    0.00447 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1339 on 11 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic:  9485 on 3 and 11 DF,  p-value: < 0.00000000000000022
```

Fig. 5. Multiple linear regression model

99.96% of the cases in Arizona can be explained by employment status of residents and population under the age of 18. Average county population under the age of 18 in Arizona is 109,510 people. While the average number of people unemployed is 11,080, and an average of 66,019 are in poverty. Research shows that Maricopa County is the hardest hit county in the state of Arizona.

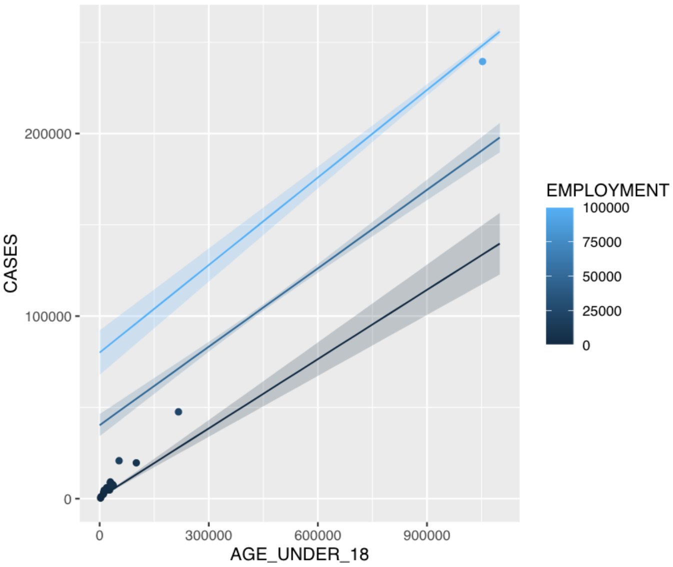
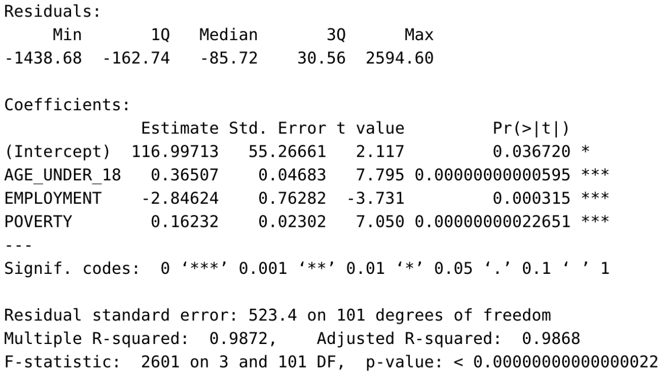


Fig. 6. Arizona multiple linear regression model prediction

Furthermore, how different factors interact is also of importance. Employment status is used as an interaction term in the model for the State of Arizona.

B. Kansas

It was enough in Arizona to explain the data using age category and employment status. This is not the case in Kansas. In Kansas, an added factor better explains the model.



```
Residuals:
     Min       1Q   Median       3Q      Max
-1438.68  -162.74   -85.72    30.56  2594.60

Coefficients:
              Estimate Std. Error t value         Pr(>|t|)
(Intercept)  116.99713   55.26661   2.117         0.036720 *
AGE_UNDER_18   0.36507    0.04683   7.795 0.00000000000595 ***
EMPLOYMENT    -2.84624    0.76282  -3.731         0.000315 ***
POVERTY        0.16232    0.02302   7.050 0.00000000022651 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 523.4 on 101 degrees of freedom
Multiple R-squared:  0.9872,    Adjusted R-squared:  0.9868
F-statistic:  2601 on 3 and 101 DF,  p-value: < 0.00000000000000022
```

Fig. 7. Kansas multiple linear regression model

We notice that in Kansas, population under the age of 18 along with employment status and poverty, have the greatest impact in predicting COVID-19 cases. It may very well be the case that as other states are studied, a trend will emerge in population age group, employment status of individuals, poverty levels and their relation to COVID-19 cases.

VI. Conclusion

In this research we studied socio-economic factors like age, race, gender, income, employment status, and poverty levels of U.S. counties using multiple linear regression models. Research found that for the U.S., race of citizens could explain approximately 94% of the data.

Race of citizens is not the only factor to have an impact on COVID-19 cases. Other socio-economic factors like, age category, employment status, and poverty levels were found to have an even greater power to explain the data at a state-level.

In Arizona, factors like age, and employment status explained 99.96% of the data. Furthermore, we see a similar trend in Kansas where age, employment status, and poverty levels explained 98% of the data.

Research leads to the assumption that educating the population under the age of 18, and working to curb poverty levels may potentially see a decline in COVID-19 related cases. However, more research is needed to fully understand the impact of pre-existing socio-economic factors at a much deeper level. This type of research will assist local and state governments in creating programs that specifically target COVID-19 related socio-economic factors.

References

[1] The World Health Organization. (2020, October 13). Impact of COVID-19 on people’s livelihoods, their health and our food systems. Retrieved December 13, 2020, from https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people’s-livelihoods-their-health-and-our-food-systems

[2] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1.

[3] Cutler, D. M., &; Summers, L. H. (2020). The COVID-19 Pandemic and the $16 Trillion Virus. Jama, 324(15), 1495. doi:10.1001/jama.2020.19759.

[4] Wong, M. (2020, June 26). A snapshot of a new working-from-home economy. Retrieved December 13, 2020, from https://news.stanford.edu/2020/06/29/snapshot-new-working-home-economy/

[5] Parker, K., Minkin, R., &; Bennett, J. (2020). Economic Fallout From COVID-19 Continues To Hit Lower-Income Americans the Hardest. Pew Research Center.

[6] Hawkins, R., Charles, E., &; Mehaffey, J. (2020). Socio-economic status and COVID-19–related cases and fatalities. Public Health, 189, 129-134. doi:10.1016/j.puhe.2020.09.016

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] Falkenberg, H., Lindfors, P., Chandola, T., & Head, J. (2020). Do gender and socioeconomic status matter when combining work and family: Could control at work and at home help? Results from the Whitehall II study. Economic and Industrial Democracy, 41(1), 29-54.