Modeling Covid 19 Vaccine Hesitancy Based on CDC County Level Estimates

Chad Ramos

November 26, 2021

Abstract

A multiple regression analysis was conducted to evaluate the ability of metrics relating to healthcare access, race, education, income, political leaning, sex, and age, to predict COVID-19 vaccine hesitancy as estimated at the county level by the CDC, based on a Census Bureau Household Pulse Survey conducted between March 3rd and March 15th, 2021 (Centers for Disease Control and Prevention, 2021). 16 variables were evaluated and 6 were found to be valid for use in the multiple regression analysis: median household income, percent of population in poverty, percent of population identifying as non-Hispanic black, percent of population with less than a bachelor's degree, percent of population without health insurance, and the CDC's Social Vulnerability Index. A best-subsets regression evaluation was conducted on the possible combinations of these 6 predictors and the best model was found to be the model making use of all 6 input variables. Together, these predictors accounted for 37 percent of the variance in COVID-19 vaccine hesitancy across counties, and all of the variables were significant predictors. The model was significant overall ($F_{(7, 3094)} = 306.02$, $p<0.001$) and the strongest predictors by scaled regression coefficients were percent of population with less than a bachelor's degree ($B=0.24$, $p<0.001$), and median household income ($B=-0.21$, $p<0.001$).

1. Introduction

It has been 20 months since the World Health Organization declared COVID-19 a global pandemic. Since then, nearly 50 million cases have been reported in the US and at the time of this report, more than 750,000 US deaths have been attributed to the disease. Vaccines for COVID-19 were approved for emergency use in December of 2020 and have been readily available for all adults since June of 2021. Regarding the effectiveness of the vaccines, a study conducted by the State of Texas Department of State Health Services shows that between January 15th and October 1st of 2021, the age adjusted COVID-19 death rate per 100,000 in unvaccinated people is 40 times higher than that of vaccinated people (AJMC Staff, 2021, Centers for Disease Control and Prevention, Texas State Health Services, 2021). Despite the availability and effectiveness of vaccines, there is still widespread COVID-19 vaccine hesitancy in the US. The US Centers for Disease Control and Prevention (CDC) estimated the percent vaccine hesitancy

for each US county based on a Household Pulse Survey conducted by the US Census Bureau for the collection period of March 3rd to March 15th, 2021. The CDC estimate is based on the survey question, "Once a vaccine to prevent COVID-19 is available to you, would you get a vaccine?", which listed 4 response options indicating varying degrees of likelihood ranging from "definitely get a vaccine" to "definitely not get a vaccine". The mean estimated COVID-19 vaccine hesitancy across US counties is 19% (Centers for Disease Control and Prevention, 2021).

Taking a public health point of view, the purpose of this research is to evaluate the possible underlying factors related to the CDC estimated vaccine hesitancy. A multiple regression analysis will be performed to discern the ability of common population characteristics to predict the vaccine hesitancy, and an emphasis will be placed on understanding conceptually how these factors may relate to the hesitancy. 16 metrics relating to healthcare access, race, education, income, political leaning, sex, and age will be considered and those which strongly relate to the dependent variable will be entered into the multiple regression. The implications of which variables create the best model and their relative strength to the prediction will be discussed. Finally, the spatial distribution of model error will be assessed and further research will be considered.

2. Sources and Methods

2.1 Sources

The independent variables evaluated in this analysis were chosen to in an attempt to represent a wide range of subject matter (for more detailed variable descriptions and sources see Table 1 in Appendix A). There are two variables serving as proxies for access to healthcare: percent uninsured and primary care physicians per 100k people. Three measures of income or poverty: percent unemployment, percent of population in poverty, and median household income. Three measures of race and ethnicity: percent White, percent Black, and percent Hispanic. Two measures of education: percent of population with less than a high school diploma, and percent of population with less than a bachelor's degree. Two measures of age: percent of population under 29 and percent of population over 65. And there are single measures of sex, political leaning, rural vs. urban, and social vulnerability (as defined by the CDC).

2.2 Methods

To examine the estimated Covid-19 vaccine hesitancy across US counties, a multiple regression analysis was performed. 16 variables representing a range of social, economic, demographic, political, education, and healthcare measures (See Table 1 in Appendix A) were compiled into a dataset for US counties. These variables were assessed on the strength of their correlation and degree of linear relationship with the independent variable. The multiple regression analysis proceeded with 6 variables that were found to be adequate ($r > 0.3$, linear relationship with dependent variable).

With 6 predictor variables, there are 63 possible combinations of variables from which to create a model. A two-part best-subsets approach was used to identify the strongest model. First, the two best 2-variable, 3-variable, 4-variable, and 5-variable models were selected programmatically by choosing the models with the highest Mallow's Cp for each number of variables. These 8 models were then assessed alongside all 6 1-variable models and the 6-variable model for a total of 15 models to evaluate. The resulting models were first checked for validity (VIF < 5, overall significance, containing only variables that contribute significantly to the model) and then compared on adjusted-R^2 and Standard Error. Model assumptions were evaluated for the selected model using residual vs. fitted, scale-location, quantile-quantile, and cook's distance plots.

The scaled regression coefficients of the best model were then compared on their relative strength to the regression and the non-standardized equivalents were evaluated in terms of their relevance to the study.

Finally, the standardized residuals were mapped and a Global Moran's I test was conducted to assess whether the residual errors exhibited spatial autocorrelation across US counties.

3. Results

3.1 Input Variables and Descriptive Statistics

The dependent variable being examined is the percent of the population for each county that has been estimated to be hesitant or unsure about the Covid-19 vaccine. The estimate was created by the US Centers for Disease Control and Prevention and the estimate is based on a Census Bureau Household Pulse Survey conducted between March 3 to March 15, 2021 (Centers for Disease Control and Prevention, 2021). The timeframe for this survey is 2 weeks before the first states started to open

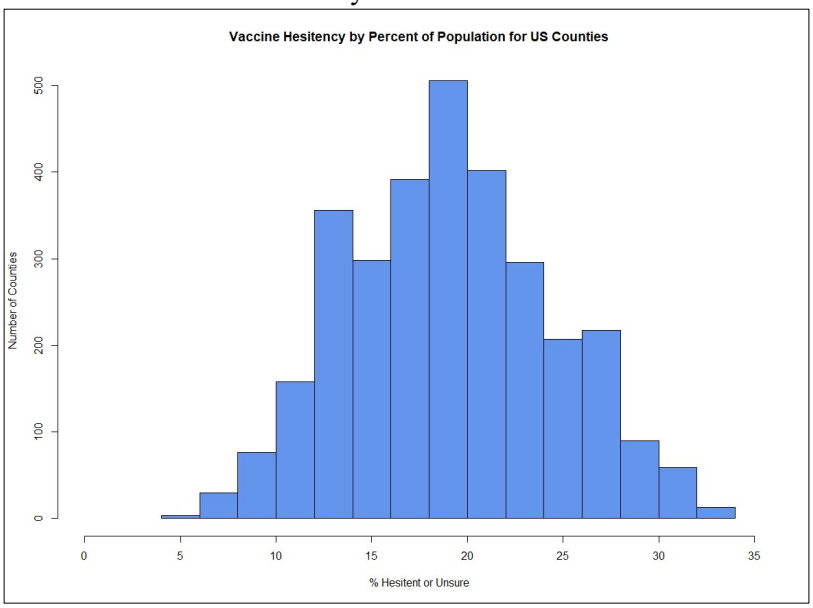

Figure 1. Histogram of Vaccine Hesitancy Across US Counties

vaccinations to all adults and approximately 2 months before the vaccine was widely and readily available across the US (AJMC Staff, 2021). The estimated vaccine hesitancy ranges from 4.99% to 32.33% with a mean of 19.08% and a standard deviation of 5.32% (Figure 1, Table 2).

The 16 independent variables were chosen to represent a wide range of population characteristics that may be related to the vaccine hesitancy. Aside from traditional population parameters such as race/ethnicity and income, variables relating to healthcare access, education, political leaning, sex, and age were also included. See Table 1 in Appendix A for variable descriptions, sources, and variable dates. See Table 2 below for descriptive statistics for all variables.

Table 2. Descriptive Statistics

	Mean	SD	Min	Max	Range	SE
estHesUns	19.08	5.32	4.99	32.33	27.34	0.10
pcp_100	19.24	5.51	5.52	32.33	26.81	0.10
pcUnIns	11.91	5.12	2.40	35.80	33.40	0.09
SVI	0.50	0.29	0.00	1.00	1.00	0.01
pcWhite	78.70	18.54	7.94	100.00	92.06	0.35
pcBlack	9.01	14.50	0.00	87.23	87.23	0.26
pcHisp	7.12	9.89	0.00	84.21	84.21	0.19
pcFemale	50.02	2.21	21.51	58.50	36.99	0.04
pcLess29	37.05	5.24	12.48	70.98	58.50	0.10
pcOver65	17.74	4.34	3.85	53.11	49.25	0.08
medHHinc	47829.61	12492.03	18972.00	125672.00	106700.00	224.29
pcUnemply	7.11	3.26	0.00	29.93	29.93	0.06
pcPovAll	14.46	5.78	2.70	47.70	45.00	0.10
pcLessHS	13.65	6.08	1.28	41.53	40.25	0.11
pcLessClg	79.20	9.14	19.79	97.01	77.23	0.16
pcRural	59.56	30.90	0.00	100.00	100.00	0.58
pctrump16	28.37	8.11	1.93	62.50	60.57	0.15

3.2 Variable Selection

From the 16 independent variables compiled for the analysis, only those with an absolute value of Pearson's r correlation of 0.3 or higher and displaying a linear relationship with the dependent variable were selected for use in the regression. The linearity of the relationship between these selected variables and the dependent variable was assessed through a series of scatterplots.

Seven variables meet the requirement of a correlation of $r > 0.3$ with the dependent variable: pcUnIns, SVI, pcBlack, medHHinc, pcPovAll, pcLessHS, and pcLessClg (see Table 3 in Appendix A).

Scatterplots of these 7 variables (along with a selection of 5 other variables deemed of interest) show linear relationships with the dependent variable in varying degrees of strength (Figure 2 below). Percent of the population without health insurance (pcUnIns) and median household income (medHHinc) appear to express the strongest linear relationships. The scatterplot for percent of population considered rural (pcRural) shows a large number of counties at 100% rural, which may weaken the correlation. Of the two measures of education, percent of the population with less than a high school diploma (pcLessHS) and percent of the population with less than a bachelor's degree (pcLessClg), only one can be used in the regression. These variables are distinctly non-independent as any person counted towards the population with less than a high school diploma is also counted towards the population with less than a bachelor's degree. Because of this, only the stronger correlated of the two, pcLessClg (r = 0.47) is used in the regression. The 6 variables to be used in the regression then, are pcUnIns, SVI, pcBlack, medHHinc,

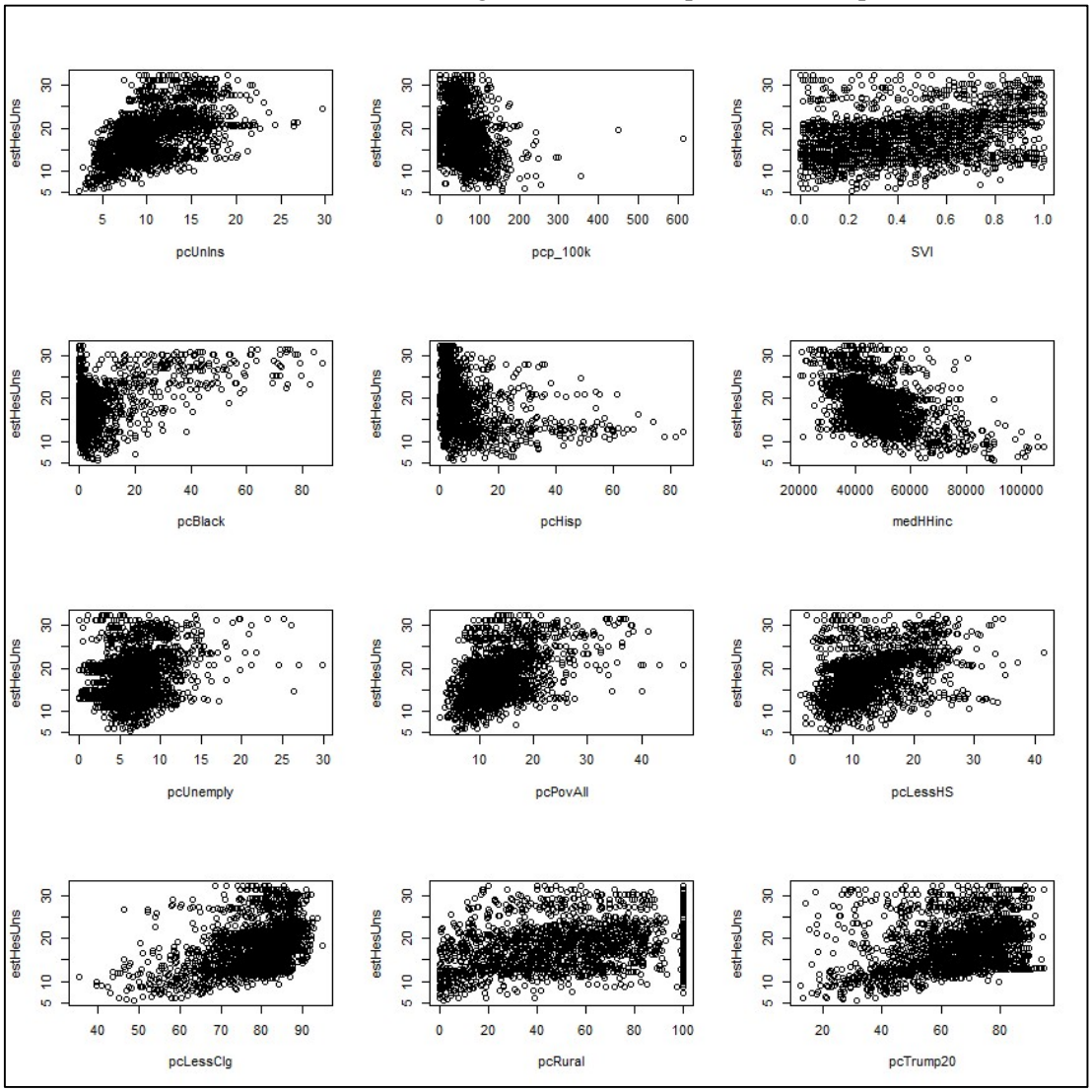

Figure 2. Selected Scatterplots

pcPovAll, and pcLessClg. It should be noted that some of these independent variables exhibit moderate to high correlation with one another, and that multicollinearity will be later assessed by calculating the Variable Inflation Factors for all variables in multi-variable models.

3.3 Regression Results and Residuals Plots

Given that there are 63 possible combinations of the 6 selected variables, it was imperative to programmatically reduce the total number of models to assess. This was done by allowing the statistical software (R via R Studio) to evaluate the possible 2-variable, 3-variable, 4-variable, and 5-variable models. The software returned the two models from each subset with the highest Mallow's Cp. These 8 models were further assessed alongside all 6 1-variable models and the 6-variable model. These 15 models were first checked for validity by removing from consideration any model with a VIF greater than 5, by removing any model that was not overall significant, and by removing any model that had an input variable that did not significantly contribute to the model. This proved unnecessary as all 15 models meet these requirements (see Table 4 below). The best model then is the 6-variable model, which had both the highest R^2-adjusted (0.371) and the lowest Standard Error (0.793).

Table 4. Regression Results

n	F	p-value	S	R	R^2 (Adj)	Variables	Max VIF
6	306.02	<0.0001	0.793	0.610	0.371	pcUnIns*, SVI*, pcBlack*, medHHinc*, pcPovAll*, pcLessClg*	3.82
5	354.83	<0.0001	0.798	0.604	0.363	pcUnIns*, SVI*, pcBlack*, medHHinc*, pcLessClg*	2.46
5	349.27	<0.0001	0.800	0.601	0.360	pcUnIns*, pcPovAll*, pcBlack*, medHHinc*, pcLessClg*	3.09
4	432.43	<0.0001	0.802	0.599	0.358	pcUnIns*, pcBlack*, medHHinc*, pcLessClg*	2.15
4	414.14	<0.0001	0.808	0.590	0.348	pcUnIns*, pcBlack*, pcPovAll*, pcLessClg*	1.84
3	521.69	<0.0001	0.815	0.579	0.335	pcUnIns*, pcBlack*, medHHinc*	1.21
3	516.49	<0.0001	0.817	0.577	0.333	pcUnIns*, pcPovAll*, pcLessClg*	1.38
2	702.79	<0.0001	0.830	0.559	0.312	pcUnIns*, medHHinc*	1.14
2	690.51	<0.0001	0.832	0.555	0.308	pcPovAll*, pcLessClg*	1.28
1	1186.7	<0.0001	0.851	0.526	0.277	medHHinc*	n/a
1	952.32	<0.0001	0.875	0.485	0.235	pcPovAll*	n/a
1	472.72	<0.0001	0.932	0.364	0.132	pcUnIns*	n/a
1	869.18	<0.0001	0.884	0.468	0.219	pcLessClg*	n/a
1	544.36	<0.0001	0.922	0.387	0.149	SVI*	n/a
1	303.75	<0.0001	0.954	0.299	0.089	pcBlack*	n/a
6**	424.3	<0.0001	0.702	0.682	0.465	pcUnIns*, SVI*, pcBlack*, medHHinc*, pcPovAll*, pcLessClg*	4.31

**Regression analysis after removal of 172 influential points identified by Cook's Distance > 4/n
*p<0.001

Regression assumptions were checked for the selected model using residual vs. fitted, scale-location, quantile-quantile, and cook's distance plots (Figure 3). Both the residual vs. fitted and scale-location plots suggest that the residuals are randomly distributed along the regression axis, albeit displaying a slight deviation from perfect homoscedasticity. It was regarded that this slight heteroscedasticity is not

enough to bring into question the coefficient values. The normal QQ plot indicates that the residuals deviate slightly from normal, however the histogram of the standardized residuals (Figure 4) shows them to be sufficiently close to normally distributed.

Finally, the Cook's Distance (Residuals vs. Leverage) plot in Figure 3 and the more detailed Cook's Distance plot in Figure 5 suggest that there are 172 observations exerting an inflated influence on the slopes of the regression line. There are several general rules by which an observation is determined to be influential by Cook's Distance. For these plots, observations with a Cook's Distance greater than 4/n where n is the number of observations, were regarded as influential. It is of course poor practice to remove outliers for the sake of improving a model, however doing so can provide insight into how much

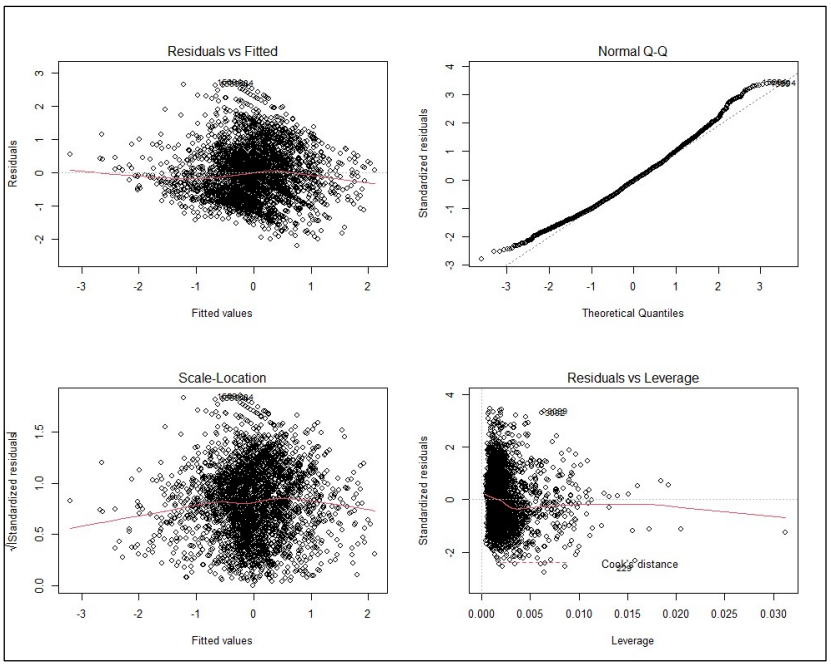

Figure 3. Residuals Plots

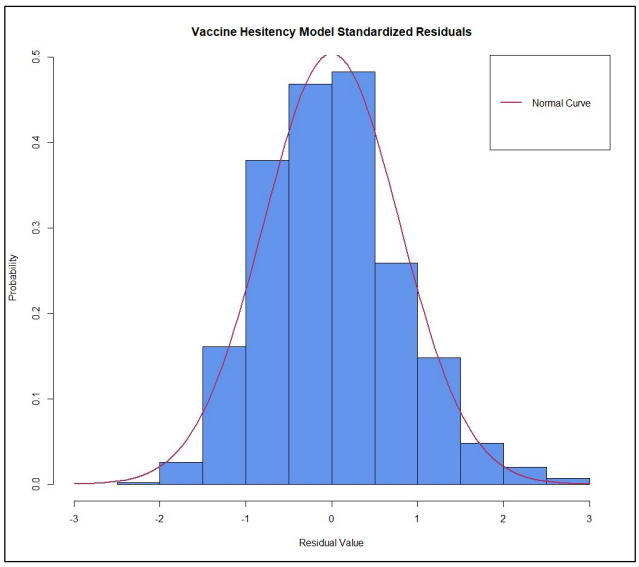

Figure 4. Probaility Density Plot of Standardized Residuals

the model *could* be improved if these outliers are properly investigated and accounted for. In this case, the removal of these 172 influential outliers improved the R^2-Adjuted from 0.371 to 0.465, a considerable increase, and lowered the Standard Error from 0.793 to 0.702.

3.4 Coefficient comparison

Having checked the model's validity and statistical significance and having further confirmed regression assumptions with the residuals plots, the regression coefficients for each input variable can be compared. The scaled and non-scaled regression coefficients, ranked in order of the absolute value of the scaled coefficients, can be found in Table 5 below. According to the model, the strongest predictor of vaccine hesitancy is the percent of each county's population with less than a bachelor's degree (B = 0.237, $p<0.001$). This coefficient is positive and by the inverse we can say that an increase in the number of people with *at least* a bachelor's degree is related to a decrease in vaccine hesitancy. The second

strongest predictor is median household income (B = -0.213, p<0.001) indicating that an increase in median household income is related to a decrease in vaccine hesitancy. Further implications of the relative strengths of these coefficients will be considered in the Discussion section below.

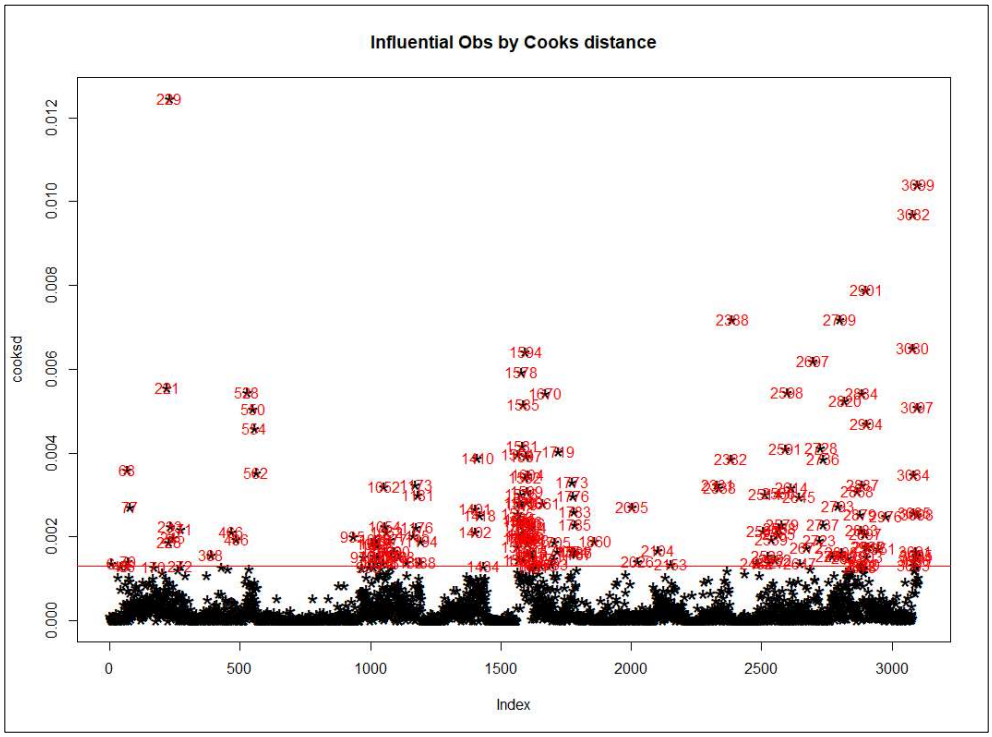

Figure 5. Influential Observations by Cook's Distance

Table 5. Regression Coefficients

	Scaled Coefficients	t-value	p-value	Non-Scaled Coefficients	Description
(Intercept)	0.00063	0.044	0.96	8.576	
pcLessClg	0.237	11.71	<0.001	0.137	Increase in population with a bachelor's degree related to a decrease in vaccine hesitancy
medHHinc	-0.213	-7.86	<0.001	-9.1E-05	Increase in median household income related to a decrease in vaccine hesitancy
pcUnIns	0.209	12.71	<0.001	0.217	Increase in population without health insurance related to an increase in vaccine hesitancy
pcBlack	0.188	10.86	<0.001	0.068	Increase in non-Hispanic Black population related to an increase in vaccine hesitancy
SVI	-0.177	-7.42	<0.001	-3.262	Increase in SVI related to a decrease in vaccine hesitancy
pcPovAll	0.176	6.31	<0.001	0.161	Increase in poverty rate related to an increase in vaccine hesitancy

3.5 Residuals map and Moran's I

Finally, mapping the standardized model residuals gives insight into the spatial discrepancies in the model's predictive abilities (Figure 6). Portions of Texas along the Rio Grande Valley, inland California, parts of Colorado, Virginia, Nebraska, Minnesota, and the far North East have residuals more than 1 standard deviation below the mean. Based on the 6 input variables, the model predicts that these areas should have a higher vaccine hesitancy than is observed. Large areas in Louisiana, Mississippi, Arkansas, Oklahoma, and Arizona show values between 1 and 2 standard deviations above the mean, and most of Wyoming and Montana have residuals more than 2 standard deviations above the mean. These large positive residuals indicate that the model is underpredicting the vaccine hesitancy in these areas.

A visual inspection of the map suggests that the distribution of the large and small residuals, i.e. the areas where the model grossly over or underpredicts the CDC estimated vaccine hesitancy, is not random. This is conformed with a Moran's I test for spatial autocorrelation (Moran's I = 0.429, $p < 0.01$). The clustered nature of model error provides a spatial basis to search for anomalies that detract from model performance, and which may give better insight into the underlying causes of vaccine hesitancy.

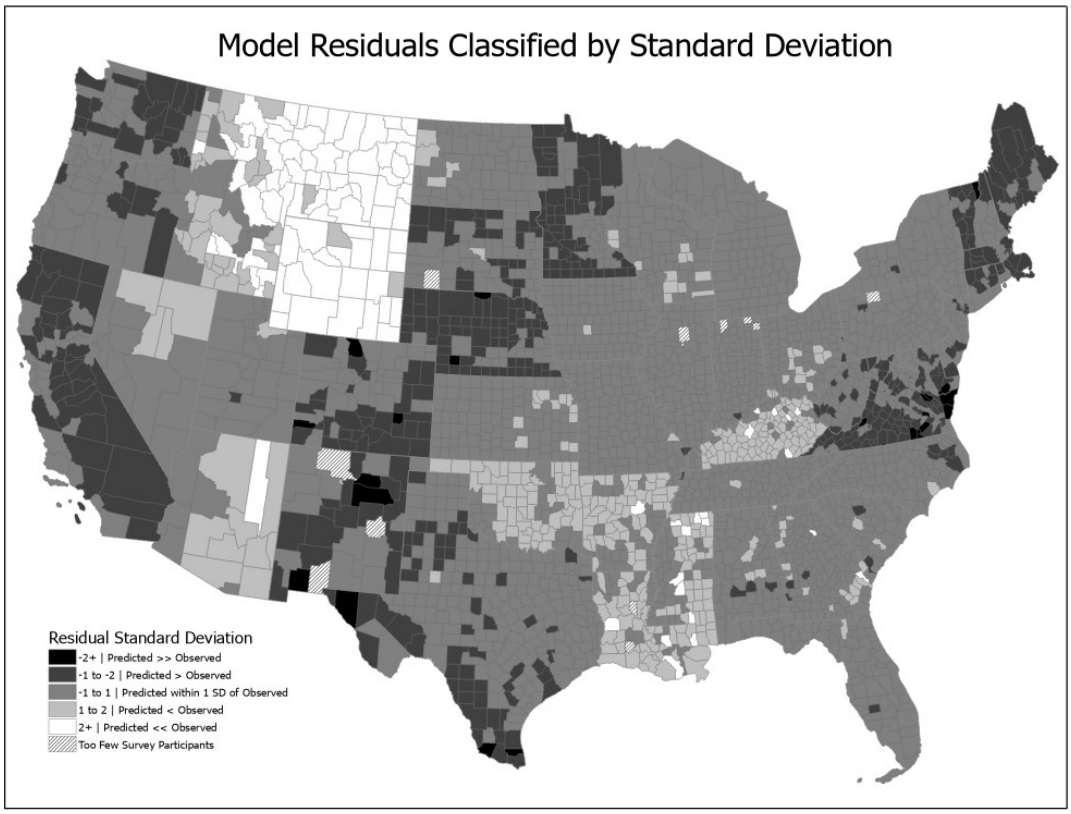

Figure 6. Spatial Distribution of Model Error Classified by Residual Standard Deviation

4. Discussion

4.1 Variable selection

If the goal from a public health point of view is to lower the vaccine hesitancy, it is important to understand the factors that relate to and may contribute to the hesitancy. In this way, it is worthwhile to remark on both the variables that correlate with and were eventually used in the regression, as well as the variables that did not correlate with the dependent variable.

For the latter, it is of interest that percent of the population without health insurance correlated strongly enough with vaccine hesitancy to be of use in the regression ($r = 0.5$) while primary care physicians per 100k people did not ($r = -0.28$). Both are common measures of access to healthcare, and in both cases *more* healthcare (i.e. a *lower* percent of population uninsured and a *higher* number of primary care physicians per capita) equated to a lower estimated vaccine hesitancy, but only percent uninsured had a strong enough correlation and a distinct linear relationship with the dependent variable. Similarly, both median household income ($r = -0.54$) and percent of population in poverty ($r = 0.51$) correlated strongly enough while percent unemployment did not ($r = 0.28$). Though again, for all three measures of income or poverty, *more* income (higher median household income, lower percent poverty, and lower percent unemployment) equated to a lower vaccine hesitancy.

It is of no surprise that percent female did not correlate ($r = -0.01$) as the male/female ratio across counties exhibited little variation from which to draw a distinction. Similarly, neither percent of the population over 65 or under 29 was related to the dependent variable ($r = -0.08$ and $r = 0.14$, respectively). It is surprising however, given the politicization of Covid-19 vaccines, that neither the percent of the population that voted for Donald Trump in 2016 ($r = 0.08$) or percent of the population considered rural ($r = 0.24$) were related strongly enough to the dependent variable. It should be noted however that many counties were 100% rural which may have affected the correlation. And at the time of this report the 2020 election returns, which would likely be a better predictor, are not readily available for the majority of the counties in the southern U.S.

Regarding the race and ethnicity variables, percent White, percent Hispanic and percent Black, only percent of the population identifying as non-Hispanic Black met the requirements for regression ($r = 0.32$). And finally, remarking upon the two education indicators, percent of the population with less than a high school diploma and percent of the population with less than a bachelor's degree, both were strong indicators relative to other variables in the study ($r = 0.44$ and $r = 0.47$, respectively). Both of these education indicators are positively correlated with vaccine hesitancy indicating by the inverse, an increase in education (i.e. a *decrease* in the population with *less* than a high school diploma or bachelor's degree) correlates to a decrease in vaccine hesitancy

Taking the position that the larger goal of this research is to eventually decrease vaccine hesitancy for

the COVID-19 vaccine or future vaccines as needed, it is worthwhile to remark upon both the variables that do and do not correlate with estimated vaccine hesitancy. The highest correlated variables relate to income, education, and access to healthcare. There are a number of non-correlating variables and the implications of these, as well as the correlating variables, go well beyond this brief discussion.

4.2 Regression Model Practical Considerations

The best model was the 6-variable model, which had both the highest R^2-adjusted (0.371) and the lowest Standard Error (0.793). However, an argument could be made that for the practical use of the model, the best choice was instead the 5-variable model without the SVI variable (R^2-adjusted = 0.360, S = 0.800). Already having an estimate for Covid-19 vaccine hesitancy for all counties, it seems the most practical use for a model then is to predict vaccine hesitancy on other spatial designations, namely cities and other municipalities. Using a model with the CDC's Social Vulnerability Index as a predictor limits the usefulness of the model as the smallest geography the SVI is computed for is Census Tracts. It would be imprecise to attempt to measure SVI for a city based on the Census Tracts contained fully or partially within the city limits. It is conceivable however that city administration would have access to values for the other 5 variables: percent uninsured, percent in poverty, percent Black, median household income, and percent with less than a bachelor's degree.

4.3 Coefficient comparison and other implications

The scaled regression coefficients of the best model (Table 4) begin to paint a picture of what may be some of the underlying causes of vaccine hesitancy. The strongest predictor is the percent of the of each county with less than a bachelor's degree ($B = 0.237$, $p<0.001$). By the inverse of this, we can say that more people with a bachelor's degree is related to a lower vaccine hesitancy. This is perhaps the least surprising result however that does not make it any easier to alleviate, particularly in the current political climate. It seems that a segment of the population would be averse to any educational outreach were it to come from a government institution. That is all that will be said about education level being the strongest predictor, it is left to the reader to further consider the implications.

The next strongest predictor was median household income ($B = -0.213$, $p<0.001$). This along with poverty rate ($B = 0.176$, $p<0.001$) and the percent of population with less than a bachelor's degree are all intuitively related to what we might consider being affluent or well-off. These 3 predictors related to one's affluence suggest that perhaps vaccine hesitancy has less to do with ideology and more with one's economic well-being. However, the final 3 variables, percent without health insurance ($B = 0.209$, $p<0.001$), percent Black ($B = 0.188$, $p<0.001$), and SVI ($B = -0.177$, $p<0.001$), detract from this economic well-being theory. A negative coefficient for SVI contradicts the idea that vaccine hesitancy is related to affluence, and both percent Black and percent without health insurance seem unrelated.

The positive correlation between percent Black and vaccine hesitancy suggests that the Black population is more averse to vaccines than the other races and ethnicities assessed in this research. This is supported by a recent study conducted among 10,871 healthcare workers at two large academic hospitals which found that Black healthcare workers were 5 times as likely to be vaccine hesitant compared to White healthcare workers (Momplaisir et. al., 2021). Knowing this to be true does not offer a solution, however it does provide a starting point for more targeted research into the reasons behind vaccine hesitancy in the Black population.

Finally, the percent of population without health insurance being positively correlated with vaccine hesitancy suggests that those without health insurance are more likely to be vaccine hesitant. At first this seems counterintuitive as it would seem that those without health insurance would be more likely to want to protections offered by the vaccine. However, if health insurance is considered to be measure of access to healthcare, then *more* access to healthcare, i.e. a *lower* percent of population without health insurance, would equate to a decrease in vaccine hesitancy.

4.4 Insights from Standardized Residuals Map

The map of residuals (Figure 6) shows that the model error is clustered. This provides an excellent opportunity to improve upon the model as we can look to these areas where the model grossly over or under predicts the observed values for clues. Take for instance the counties with residuals more than 1 standard deviation above the mean in Louisiana, Mississippi, Arkansas, and Oklahoma. It is not that the values for the 6 input variables are drastically extreme here for if they were, then the model would account for the high values and predict a higher vaccine hesitancy. No, in these areas the 6 input variables do not explain vaccine hesitancy as well as they do in other areas of the country, or even other areas in the South (see Alabama, Georgia, South Carolina etc.). This allows us to target certain areas to look for different input variables and to question what is different here from similar counties.

4.5 Further Research

There are many avenues of further research that could be taken from this initial analysis. Four will be discussed here. First, more variables related to the strongest predictors could be evaluated. More measures of education, income, and access to healthcare such as test scores, median home values etc. could be assessed for their ability to increase the explained variance. Second, this search for variables could be improved by using the standardized residuals map to target counties that are not well explained by the current model. Might there be some other underlying variables that better explain the vaccine hesitancy in these areas? This could be done in conjunction with examining the 172 counties identified by Cook's Distance as having an inflated influence on the model, the removal of which increased the adjusted-R^2 from 0.371 to 0.465. And finally, having shown both visibly and with a Global Moran's I (I

= 0.429, $p<0.01$) that the residuals are spatially autocorrelated, this research is a prime candidate for a geographically weighted regression.

5. Conclusion

A multiple regression analysis was conducted to better understand and predict Covid-19 vaccine hesitancy across US counties. It was shown that of 16 variables relating to income, race, healthcare, education, age, sex, politics, and urban vs. rural divide, 6 were valid for use in the multiple regression analysis: median household income, percent of population in poverty, percent of population identifying as non-Hispanic black, percent of population with less than a bachelor's degree, percent of population without health insurance, and the CDC's Social Vulnerability Index. A best-subsets regression evaluation found that the best model was the model making use of all 6 input variables. Together, these predictors accounted for 37 percent of the variance in Covid-19 vaccine hesitancy across counties, and all of the variables were significant predictors. The model was significant overall ($F_{(7, 3094)} = 306.02$, $p<0.001$) and the strongest predictors based on scaled regression coefficients were percent of population with less than a bachelor's degree ($B=0.24$, $p<0.001$), and median household income ($B=-0.21$, $p<0.001$). A map of the standardized residuals shows the model error to be clustered (Moran's $I = 0.429$, $p<0.01$) and this suggests that of the many avenues for further research, a geographically weighted regression would be advisable.

References

AJMC Staff. (2021, June 3). *A timeline of covid-19 vaccine developments in 2021*. AJMC. Retrieved November 29, 2021, from https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021.

AJMC Staff. (2021, January 1). *A timeline of covid-19 developments in 2020*. AJMC. Retrieved November 30, 2021, from https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020.

Centers for Disease Control and Prevention. (2021, June 17). *Covid-19 County Hesitancy*. Centers for Disease Control and Prevention. Retrieved November 9, 2021, from https://data.cdc.gov/Vaccinations/COVID-19-County-Hesitancy/c4bi-8ytd.

Centers for Disease Control and Prevention. (n.d.). *CDC Covid Data tracker*. COVID Data Tracker. Retrieved November 25, 2021, from https://covid.cdc.gov/covid-data-tracker/#datatracker-home.

MIT Election Data + Science Lab. (2018, August 15). *U.S. Primary Elections 2018*. Data | MIT Election Lab. Retrieved November 27, 2021, from https://electionlab.mit.edu/data.

Momplaisir, F. M., Kuter, B. J., Ghadimi, F., Browne, S., Nkwihoreze, H., Feemster, K. A., Frank, I., Faig, W., Shen, A. K., Offit, P. A., & Green-McKenzie, J. (2021). Racial/ethnic differences in covid-19 vaccine hesitancy among health care workers in 2 large academic hospitals. *JAMA Network Open*, *4*(8). https://doi.org/10.1001/jamanetworkopen.2021.21931

Texas State Health Services, COVID-19 Cases And Deaths by Vaccination Status (2021). Retrieved November 25, 2021, from https://www.dshs.texas.gov/immunize/covid19/data/Cases-and-Deaths-by-Vaccination-Status-11082021.pdf.

U.S. Department of Agriculture. (2021, January 5). *Poverty Estimates for the U.S., States, and Counties, 2019*. USDA ERS. Retrieved November 27, 2021, from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/.

U.S. Health Resources & Services Administration. (n.d.). *Area Health Resources Files*. Workforce Data. Retrieved November 7, 2021, from https://data.hrsa.gov/topics/health-workforce/ahrf.

Appendix A: Large Tables and Figures

Table 1 Summary of Variables and Sources

Variable	Description	Variable Date	Source
estHesUns	Estimated percent of population that is hesitant or unsure about receiving the Covid-19 vaccine	2021	US Centers for Disease Control and Prevention (2021)
pcUnIns	Percent of population without health insurance	2019	Census Bureau Small Area Health Insurance Estimates 2019
pcp_100k	Primary care physicians per 100k people	2019	US Health Resources & Service Administration Area Health Resource Files
SVI	Social Vulnerability Index	2021	US Centers for Disease Control and Prevention (2021)
pcWhite	Percent of population identifying as White	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcBlack	Percent of population identifying as Black	2012-2016 ACS 5-Year estimates	MIT Elections Lab 2018 US General Elections Analysis Dataset
pcHisp	Percent of the population identifying as Hispanic	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcFemale	Percent of population Female	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcLess29	Percent of population less than 29 years of age	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcOver65	Percent of population over 65 years of age	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
medHHinc	Median household income in the past 12 months	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcUnemply	Unemployed population in labor force as a percentage of total population in civilian labor force	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcPovAll	Estimate of people of all ages in poverty 2019	2019	US Department of Agriculture Economic Research Service 2019 Poverty estimates for the U.S., States, and counties.
pcLessHS	Population with an education of less than a regular high school diploma as a percentage of total population	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcLessClg	Population with an education of less than a bachelor's degree as a percentage of total population	2012-2016 ACS 5-Year estimates	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pcRural	Rural population as a percentage of total population	2010	MIT Elections and Data lab 2018 US General Elections Analysis Dataset
pctrump16	Presidential candidate votes as a percentage of total votes	2017	MIT Elections and Data lab 2018 US General Elections Analysis Dataset

Table 3. Pearson's r correlation matrix

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1. estHesUns	-															
2. pcUnIns	**.50***	-														
3. pcp_100k	-.28*	-.22*	-													
4. SVI	**.42***	.48*	-.12*	-												
5. pcWhite	-.12*	-.36*	-.11*	-.61*	-											
6. pcBlack	**.32***	.25*	.02	.51*	-.71*	-										
7. pcHisp	-.22*	.18*	.09*	.26*	-.51*	-.09*	-									
8. pcFemale	-.01	0	.20*	0	-.04	.10*	-.09*	-								
9. pcLess29	.14*	.10*	.12*	.29*	-.40*	.18*	.27*	.07*	-							
10. pcOver65	-.08*	.03	-.13*	-.21*	.38*	-.22*	-.22*	.05*	-.83*	-						
11. medHHinc	**-.54***	-.45*	.31*	-.63*	.12*	-.28*	.12*	.03	.08*	-.27*	-					
12. pcUnemply	.28*	.23*	-.11*	.68*	-.49*	.48*	.06*	.06*	.14*	-.13*	-.46*	-				
13. pcPovAll	**.51***	.41*	-.19*	.76*	-.46*	.52*	-.03	-.08*	.16*	-.05*	-.76*	.64*	-			
14. pcLessHS	**.44***	.49*	-.30*	.76*	-.42*	.42*	.21*	-.07*	.10*	-.11*	-.60*	.54*	.68*	-		
15. pcLessClg	**.47***	.32*	-.58*	.45*	.04*	.11*	-.12*	-.16*	-.18*	.23*	-.70*	.31*	.48*	.61*	-	
16. pcRural	.24*	.23*	-.47*	-.06*	.30*	-.09*	-.33*	-.18*	-.44*	.48*	-.40*	-.03	.21*	.21*	.53*	-
17. pctrump16	.08*	.01	-.28*	-.47*	.68*	-.43*	-.38*	-.03	-.51*	.53*	-.02	-.41*	-.30*	-.24*	.24*	0.51*

* $p<0.05$