The Impact of Socioeconomic and Vaccination Statuses on COVID-19 Cases in Maryland

Jay Jung, Ana Kuri, and Zhaoxu Zhang

April 17, 2023

Theory

Motivation

Statistical methods, specifically regression models, have important applications within the field of epidemiology. Within epidemiology, regression models are used to examine the effect of various explanatory variables (i.e., exposures, subject characteristics, and risk factors) on a response variable such as mortality or disease. Adjusted effect estimates can be derived from multiple regression models that take into account the effect of potential confounders on the effect estimates.

Weighted least squares is a type of weighted linear regression that fits a linear model, weighting the observations by their variances. Weighted least squares is useful in epidemiological studies as the variance of the response variable may differ across different subgroups of the study population, so assigning greater weights to observations with smaller variances results in more precise estimates of the model parameters. As such, weighted least squares is a powerful tool in epidemiological research that can inform us about relationships between predictor and response variables associated with a disease and help identify subpopulations that are heavily affected by the disease in question, motivating targeted interventions to improve their health outcomes. Weighted least squares has been applied in epidemiological research on COVID-19 to investigate case and hospitalization rates, revealing important predictors of the disease along with factors and subpopulation characteristics that contribute to higher COVID-19 rates.

Literature Review

Reviewing the literature on hospitalization rates for COVID-19 and case incidence in the United States, studies show that the hospitalization rates and case incidence for vaccinated individuals are lower than that for unvaccinated

individuals. A cross-sectional study of U.S. adults hospitalized with COVID-19 from January 2022 to April 2022 (the period of Omicron variant predominance) revealed that compared to vaccinated persons who received a booster dose, the COVID-19-associated hospitalization rates among unvaccinated persons is 10.5 times higher and for vaccinated persons with no booster it is 2.5 times higher. In this study, compared to unvaccinated hospitalized persons, vaccinated hospitalized individuals were more likely to be older and have more underlying medical conditions, which are the subsets of the population most vulnerable to COVID-19, suggesting a difference in the severity of COVID-19 illness among unvaccinated and vaccinated individuals where unvaccinated younger non-immunocompromised individuals experience greater severity, requiring hospitalization. As such, COVID-19 vaccines are strongly associated with prevention of serious COVID-19 illness.

Another study done April 4 to December 25, 2021 looked at the COVID-19 incidence and death rates in 25 U.S. jurisdictions (Alabama, Arkansas, California, Colorado, District of Columbia, Florida, Georgia, Idaho, Indiana, Kansas, Louisiana, Massachusetts, Michigan, Minnesota, Nebraska, New Jersey, New Mexico, New York, New York City (New York), Rhode Island, Seattle/King County (Washington), Tennessee, Texas, Utah, and Wisconsin) and had similar findings. The study found that there were more COVID-19 cases among unvaccinated persons aged 18 years (n=6,812,040) compared to fully vaccinated persons (n=2,866,517). Additionally, the average weekly, age-standardized rates of cases and deaths (events per 100,000 population) were consistently higher in all COVID-19 strain periods (pre-Delta, Delta emergence, Delta predominance, Omicron emergence) among unvaccinated persons. More specifically, the decrease in averaged weekly, age-standardized case incidence rate ratios (IRRs) among unvaccinated persons compared with fully vaccinated persons in 2021 was from 13.9 pre-Delta to 8.7 as Delta emerged, and to 5.1 during the period of Delta predominance. In October and November, compared with fully vaccinated persons who received booster doses, unvaccinated persons had 13.9 times the risk for infection and 53.2 times the risk for COVID-19-associated death. In October and November, compared with fully vaccinated persons without booster doses, unvaccinated persons had 4.0 and 12.7 times the risks, respectively. In December 2021 (when the Omicron variant emerged), case IRRs decreased to 4.9 for fully vaccinated persons with booster doses and 2.8 for those without booster doses, relative to October-November 2021. Additionally, the impact of booster doses against infection and death compared with full vaccination without booster doses was highest among persons aged 50-64 and 65 years. Ultimately, fully vaccinated persons with a booster dose had lower rates of COVID-19 cases (25.0 per 100,000 population) compared to fully vaccinated persons without a booster dose (87.7 per 100,000 population) and much lower rates compared to unvaccinated persons (347.8 per 100,000 population) from October to November, and in December (148.6, 254.8, and 725.6 per 100,000 population, respectively). Similar trends were noted for differences in the mortality rates among these three groups (0.1, 0.6, and 7.8 per 100,000 population,

respectively) during the months of October and November.

In addition to the above two articles, a third study examines vaccination records in 13 US jurisdictions ("Alabama, Arizona, Colorado, Indiana, Los Angeles County (California), Louisiana, Maryland, Minnesota, New Mexico, New York City (New York), North Carolina, Seattle/King County (Washington), and Utah") with COVID-19 cases. The dates range from April 4th to July 17th, 2021. Depending on when the Delta variant is most influential, the analysis is broken up into 2 periods: April 4 to June 19 and June 20 to July 17. The statistics are from 2019 U.S. intercensal population estimates and 2000 U.S. Census standard population. As part of the algorithm, the researchers examined age-standardized incidence rate ratios (IRRs), IRR = cases in people "not fully vaccinated"/cases in people "fully vaccinated". They also adopted the formula $PVC = \frac{[PPV–(PPV–VE)]}{[1–(PPV–VE)]}$, where PVC is the percentage of vaccinated persons occurring among outcomes, PPV is the proportion of the population that is vaccinated, and VE is vaccine effectiveness (using estimates of 80%, 90%, and 95%). To obtain the results, age-standardized crude VE was estimated as (1 – [incidence in vaccinated/incidence in unvaccinated]). Furthermore, they conducted a sensitivity analysis on people who got vaccination but not fully vaccinated. The primary programs are SAS and R.

The researchers based their results on IRRs. PVC is inversely related to VE. As VE decreases (vaccination increases), PVC increases, which is comparatively complicated. While IRRs are in direct relationship with VE. Cases in the unvaccinated population have IRRs of 11.1. With a 95% confidence interval, it's 7.8–15.8. IRRs of cases in the fully vaccinated population are only 4.6. With a 95% confidence interval, it's 2.5–8.5. During April 4th to June 19th, there's a 37% vaccination coverage. With 90% VE, vaccinated individuals should be 6% of total cases, which is close to the observed data of 5%. During June 20th to July 17th, there's 53% coverage on vaccination. Vaccinated people should be 10% of the cases. However, the observed was that vaccinated people account for 18% of the cases, which would happen if VE were to be 80%. It is easy to observe that the IRRs are much lower in the fully vaccinated population and the vaccine effectiveness is high in both time periods.

After reviewing the literature, there is a clear consensus that COVID-19 case incidence and hospitalization rates in the U.S. for vaccinated persons is much lower than for unvaccinated persons. However, not much is known about if this relationship between vaccination status and COVID-19 case incidence is reflected in Maryland on the county level, so we plan to explore this relationship in this report.

Besides vaccinations, there are many other factors affecting COVID rates in different populations. According to studies, socioeconomic factors are one of them.

A literature titled "Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US" compares data from January 20 to July 29, 2020 on 50 US states, including Washington and District of Columbia. The study uses Social Vulnerability Index (SVI) to measure a community's susceptibility to catastrophes. It's based on "socioeconomic status, household composition and disability, racial/ethnic minority status and language, and housing type and transportation." The scale ranges from 1-10, where 10 means very sensitive to disasters, and 1 means relatively not sensitive.

For COVID infections, it uses mixed-effects negative binomial regression to approximate COVID-19 cases. Furthermore, an initial bivariate analyses was implemented to assess relationships between cases and "population density, urbanicity, and COVID-19 testing rate." Serial cross-sectional models analyze the relationship between socio-demographic variables and incidence by week. Incidence rate ratios (IRRs) and estimated probabilities find correlations, while sensitivity analyses prevent using counties in five states ("Arizona, Connecticut, Delaware, the District of Columbia, and Rhode Island") with the highest COVID-19 incidence rate. The researchers use programs such as R and Bonferroni adjustment.

Initial bivariate analyses suggest that "population density and urbanicity, but not COVID-19 testing rate, were significantly associated with COVID-19 incidence and mortality". There's a strong correlation between SVI and COVID-19 cases and deaths, such as a 0.1 increase in SVI means a 14.3% increase in infection rates, 13.7% increase in death rates. Also, there's a 0.9% increase in weekly cumulative increase in infection rate per 0.1 increase in SVI. A high SVI county also increases faster in weekly cumulative incidence and mortality rates.

Factors that induce COVID-19 cases include percentage of the population living in crowded housing, low English proficiency, or single parent households, obesity rate, which are correlated to socioeconomic status. Communities with more racial/ethnic minority populations also have higher COVID-19 cases.

A second study titled "Socio-economic status and COVID-19-related cases and fatalities" examines effects of social factors on COVID infections. The study includes 50 states of the US, or 3143 counties. The primary tool in analysis is Distressed Communities Index (DCI), which is based on "unemployment, education level, poverty rate, median income, business growth, and housing vacancies." It ranges from $0 - 100$, with 100 being most distressed. A DCI score greater than 75 means the community is severely distressed. Counties are separated based on DCI into 2 groups, one over 75, the other equal or below 75. They are all under univariate analysis using the Mann-Whitney U test. Socioeconomic status is determined using "hierarchical linear mixed models with Laplace approximation and a negative binomial distribution." Adjusted rate ratios are made by regression. Programs in visualization are SAS and Prism 8.

In the exploration, researchers found there was "no difference in median cases per 100,000 persons" between 2 groups, but a "higher median fatalities per 100,000 persons in severely distressed counties." Furthermore, the percentage of African Americans, elderly residents, uninsured residents, and people with chronic disorders and heart diseases are high in severely distressed counties.

Socioeconomic factors ("lower education level, higher proportion of black Americans, higher income, and lower poverty rate") of counties are related to cases of COVID-19, but there's no significant evidence to say that health comorbidities relate to COVID-19 cases. "Covariates with significant associations with cases per 100,000 persons" are the percentage of adults without graduate from high school, percentage of blacks, median income, and poverty rates. Two strongest factors are adults without graduate from high school and percentage of blacks.

The third study on "Assessing the Impact of Neighborhood Socioeconomic Characteristics on COVID-19 Prevalence Across Seven States in the United States" finds the association of COVID-19 cases and socioeconomic status on 7 states (Arizona, Florida, Illinois, Maryland, North Carolina, South Carolina, and Virginia) for April 20 to May 30, 2020. Area Deprivation Index (ADI) ranks communities by socioeconomic status. Similar to SVI and DCI, a higher rank on ADI means a worse-socioeconomically community. The study uses descriptive analyses, correlation coefficients for age, gender, and the square mileage in each community.

Communities with higher ADI in IL and MD had higher COVID-19 cases than communities across the US or in the same state with lower ADI. While less-socioeconomically-stable communities in all states (except VA) had more COVID-19 cases than socioeconomically-better neighborhoods. There also are cases when the inability of COVID-19 testing due to socioeconomic factors may hide the actual number of COVID cases in sensitive communities. Furthermore, the researchers observe a negative correlation between the percentage of white in communities and COVID-19 prevalence.

In general, the studies arrive at a similar conclusion that socioeconomic factors such as urbanicity, African American proportions, poverty rates, etc., are in positive correlation to COVID cases.

Algorithm

The mathematical model for weighted least squares originally comes from the ordinary least squares regression that has the following formula:

$$y = \beta X + \epsilon$$

The y represents the dependent variable with the X indicating a vector of independent variables and ϵ representing the error term. The β coefficients represent

the coefficients of the independent variables in the regression specification. The goal of the ordinary least squares method is to minimize the square of the error term

$$\epsilon = y - \beta X$$

that may be expressed as the matrix $e^T e = [e_1 * e_1, ..., e_n * e_n]$ and hence the term least squares.

In terms of X and Y, the squared error would be expressed as the following:

$$e^T e = (y - \beta X)^T (y - \beta X)$$

Expanding the equation, the error squared term would be the following:

$$\epsilon^T \epsilon = y^T y - \beta y X - \beta X^T y + \beta^2 X^T X$$

$$= y^T y - 2\beta y X + \beta^2 X^T X$$

Minimizing the error term with respect to β would be equivalent to

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} = -2yX + 2\beta X^T X = 0$$

that allows for the normalizing equation

$$X^T X \beta = yX$$

Then, the equation results in

$$\beta = (X^T X)^{-1} yX$$

Similar to most mathematical models, the ordinary least squares model has some assumptions, which includes how the variance of the error term is constant. More specifically, the model assumes that the error term is normally distributed with mean zero and a constant variance σ. Under many applications, including the epidemiological datasets, a constant error term is a strong assumption that is not always satisfied. To address the concern of nonconstant error term, many have implemented the weighted least squares method of performing regression analysis.

The fundamental idea behind the weighted least squares is similar to the general ordinary least squares regression analysis. The main difference between the two methods originate from how the weighted least squares method assigns certain weight to each observation. Such weighting would allow the observations to have proper influence over the parameter estimation or the β coefficients for the regression analysis.

Suppose that $y = \beta X + \epsilon$, similar to the ordinary least squares. Due to how the weighted least squares assigns certain weights, the squared error is expressed

by multiplying the existing $e^T e$ with a weight matrix W. Indeed, the ordinary least squares is the special case in which the weights are all equal with one. As such, the weighted squared error is the following:

$$We^T e = W(y - \beta X)^T (y - \beta X)$$

Implementing the identical method to finding the β coefficients in the ordinary least squares method, the β in the weighted least square is

$$\beta = (X^T W X)^{-1} X^T W y$$

The weight matrix W is a diagonal matrix with the diagonals being $\frac{1}{\sigma_i^2}$ and the other elements being zero within the matrix. The σ_i^2 is the variance of the i^{th} observation. The other elements in the weight matrix represent the covariance between the variables. The following is the matrix expression of the weight matrix for n observations:

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix}$$

The weighted least squares method is particularly useful when the sample size of the dataset is relative small. Another instance in which the weighted least squares method is applicable is when one can conjecture that the variance of the error term would vary across different values of the explanatory variables or the X matrix in the regression specification. For example, when attempting to regress age and income over the net worth, one may conjecture that the variance of the error term would increase as age increases due to how net worth would diverge as age increases.

In the case of the COVID-19 dataset, the weighted least squares may be useful due to how regions in Maryland with different populations may have higher variance in the error term. To implement the weighted least squares method with the COVID-19 dataset, we have the following specification as the general specification for the analysis:

$$\text{COVID–19 Cases} = \beta_1 \text{SES} + \beta_2 \text{Vaccination} + \beta_3 \text{Year} + \beta_4 \text{Other Factors} + \epsilon$$

In order to perform the regression analyses, we will utilize the datasets from the Maryland state database on COVID–19 data, Maryland state data on county–level socioeconomic variables, and the Center of Disease Control dataset on vaccination rate. The datasets are time series data from 2020 to 2023 that have been cleaned to include the relevant variables across the common time frame

recorded within both datasets.[1]

In our analysis, we will be performing the ordinary least squares first that would allow for a comparison with the weighted least squares method. Not only that, but the ordinary least squares model would allow for the computation of the error variance necessary for the construction of the weight matrix in the weighted least squares method. To test how the results change, the analysis will demonstrate the results from other specifications to assess which model would be most effective in analyzing the COVID–19 cases at the county level in Maryland from 2020 to 2023.

[1]Please refer to the code for details on how the data cleaning has been performed