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

Jay Jung, Ana Kuri, and Zhaoxu Zhang

April 21, 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

COVID-19 Hospitalization Rates and Case Incidence

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. Havers et al. (2022) perform a cross-sectional study of U.S. adults hospitalized with COVID-19 from January 2022 to April 2022 (the period of Omicron variant predominance). Havers et al. (2022) reveal 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.

From April 4 to December 25, 2021, Johnson (2022) 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 $\geq$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 $\geq$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 mor-

tality 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, Scobie et al. (2021) examines vaccination records in 13 US jurisdictions with COVID-19 cases. The dates range from April 4th to July 17th, 2021. Depending on when the Delta variant is most prevalent, the analysis is broken up into 2 periods: April 4 to June 19 and June 20 to July 17. As part of the algorithm, the researchers examined age-standardized incidence rate ratios (IRRs),

$$IRR = \frac{cases\ in\ people\ \text{“}not\ fully\ vaccinated\text{”}}{cases\ in\ people\ \text{“}fully\ vaccinated\text{”}}$$

They also adopted the formula

$$PVC = \frac{[PPV\text{–}(PPV\text{–}VE)]}{[1\text{–}(PPV\text{–}VE)]},$$

where PVC is the percentage of COVID-19 cases from vaccinated individuals out of all cases, PPV is the proportion of the population that is vaccinated, and VE is vaccine effectiveness. To obtain the results, age-standardized crude VE was estimated as $(1 - \frac{incidence\ in\ vaccinated}{incidence\ in\ unvaccinated})$. 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. IRRs are in direct relationship with VE. Cases in the unvaccinated population have IRRs of 11.1 (95% confidence interval: 7.8–15.8). IRRs of cases in the fully vaccinated population are only 4.6 (95% confidence interval: 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. It was observed that vaccinated people account for 18% of the cases, which would happen if VE were to be 80%. In general, 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 whether 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.

Socioeconomic Status' Effect on COVID-19 Rates

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

Karmakar et al. (2021) 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.

The study uses mixed-effects negative binomial regression to approximate COVID-19 cases, an initial bivariate analyses to assess relationships between cases and SES, and serial cross-sectional models to 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.

There's evidence that COVID-19 cases and death rate are correlated with population density and urbanicity. Also, there's a strong correlation between SVI and COVID-19 cases, such as a 0.1 increase in SVI means a 14.3% increase in infection rates. 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 rates. Factors that induce COVID-19 cases include percentage of the population living in crowded housing, low English proficiency, single parent households, obesity rate, high percentage of racial/ethnic minorities.

Hawkins et al. (2020) examines SES's relationship with COVID infections on 50 states of the US. The primary tool in analysis is Distressed Communities Index (DCI), which ranges from 0 – 100, with 100 being the most distressed. A DCI score >75 means the community is severely distressed. Counties are separated based on DCI into 2 groups, one >75, the other ≤ 75. They are all under univariate analysis using the Mann-Whitney U test. Adjusted rate ratios are made by regression. Programs in visualization are SAS and Prism 8.

In the exploration, Hawkins et al. (2020) 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." "Covariates with significant associations with cases per 100,000 persons" are the percentage of adults without high school diploma, percentage of blacks, income, and poverty levels. Two strongest factors are adults without high school diploma and percentage of blacks.

Hatef et al. (2020) finds the association of COVID-19 cases and socioeconomic status in 7 states for April 20 to May 30, 2020. Area Deprivation Index (ADI) ranks communities by socioeconomic status. Similar to SVI and DCI, a higher

ranking on ADI means a low socioeconomic community. The study uses descriptive analyses and correlation coefficients to analyze each community.

Communities with higher ADI in IL and MD had higher COVID-19 cases than communities across the US. While less-socioeconomically-stable communities in all states (except VA) have more COVID-19 cases than high-socioeconomic 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.

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-19 cases. In the exploration, we hope to further analyze the association of COVID-19 with SES in Maryland.

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:

$$\begin{aligned} \epsilon^T \epsilon &= y^T y - \beta y X - \beta X^T y + \beta^2 X^T X \\ &= y^T y - 2\beta y X + \beta^2 X^T X \end{aligned}$$

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

$$X^T X \beta = X^T y$$

Then, the equation results in

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

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 to 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.

To measure the effectiveness of the regression models, we utilize the R-squared values. The R-squared values represent the proportion of the variations in the dependent variable explained by the independent variables in the regression. Mathematically, the R-squared value can be computed through the following equation:

$$R^2 = 1 - \frac{\Sigma_i (y_i - \hat{y_i})}{\Sigma_i (y_i - \bar{y})}$$

The second part of the equation represents the proportion of the sum of residuals to the total sum of the squared errors. One of the concerns with the R-squared value is the issue of overfitting. The overfitting occurs when the number of independent variables in the regression is high relative to the number of observations. Overfitting can be attributed to the fact that the R-square values does not decrease as more independent variables are added to the regression model. Such phenomena is problematic due to how it reduces the generality of the regression models. To account for the issue of overfitting, we compute the adjusted R-squared value that has the following equation:

$$R^2 \text{ Adjusted} = \frac{\Sigma_i (y_i - \hat{y_i})/df_\epsilon}{\Sigma_i (y_i - \bar{y})/df_y} = \frac{(1 - R^2)(n-1)}{n-k-1}$$

The adjusted R-squared value in the formula is based on the regression model with k independent variables and n observations. As such df_ϵ that represents the degree of freedom for the error variance is equal to $n - k - 1$ and df_y that indicates the degree of freedom for the variance of the dependent variable y being equal to $n - 1$ in the equation.

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:

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

More details regarding the datasets and how we implemented the regression analysis will be provided in the data exploration section.

Data Exploration

Data

We utilized the datasets from the Center for Disease Control (CDC) and the Maryland state database on the COVID–19 Pandemic. Specifically, we employed MDCOVID19 CaseByCounty data from the State of Maryland (2023b) that records cases from March, 2020 to March, 2023 across 23 Maryland counties and Baltimore City. The data on the socioeconomic status are extracted from the Open Data Portal provided State of Maryland (2023a) that provides information ranging from educational levels to racial compositions to population numbers. We employed COVID–19 Vaccinations in the United States, County dataset from Center for Disease Control (2023) that includes data such as the percentage of population that has completed the series of COVID–19 vaccines. The date range for the CDC dataset is from December, 2020 to March, 2023. Within the CDC dataset, we only utilized the Maryland counties, consistent with our analysis.

In order to utilize these datasets for regression analysis, we merged the three datasets based on the county and dates available from each dataset. Following the data merge, the final dataset has dates ranging from October, 2021 to March, 2023 for all Maryland counties and Baltimore City. The dataset has 1,776 observations with 79 variables available for the regression analysis.

Descriptive Statistics

Socioeconomic Status of Maryland Residents

Referring to Table 1 to examine the socioeconomic status of Maryland residents at the state level, the highest level of education attained by residents varies. The most common level is a high school diploma (24.16%) with a bachelor's degree being the second most common (21.79%). 19.10% of residents earned a graduate or professional degree and 18.72% completed some college. A very small proportion of Maryland residents have an associates degree (6.79%) or did not complete high school (5.57%) and even fewer completed less than 9th grade (3.87%). Employment rates in Maryland are high, with 94.83% of residents being employed. Additionally, there is a relatively equal split between male (47.51%) and female (52.49%) residents. In terms of race, Maryland is predominantly inhabited by White residents, who make up 50.16% of the population while 29.38% of residents are Black, 10.26% are Hispanic or Latino of any race, 6.33% are Asian, and the remaining 3.88% are American Indian/Alaskan Native, Native Hawaiian/Pacific Islander, some other race, or two or more races.

Referring to Table 2 to examine the socioeconomic status of Maryland residents at the county level, there are significant differences in the highest level of education attained by residents across Maryland. Prince George's county has the

highest percentage of residents who have attained less than 9th grade level education (6.78%), while Calvert County has the lowest (1.17%). Somerset County has the highest percentage of residents with a high school diploma as their highest earned degree (12.26%), while Howard County has the lowest (2.38%). The counties with the highest percentage of its residents who have earned a bachelor's degree as their highest degree are Howard County (30.38%) and Montgomery County (27.29%), while the lowest are Somerset County (10.81%), Caroline County (11.37%), Allegany County (11.43%), and Dorchester County (11.65%). Similarly, Howard County (32.29%) and Montgomery County (31.96%) have the highest percentage of residents with a graduate or professional degree, while Somerset County has the lowest (4.96%). Across all counties, over 90% of residents are employed, but Somerset County has the highest percentage of unemployed residents (9.23%). The gender distribution of residents is relatively equal across all counties, with most counties having a population consisting of 45-50% males and 50-55% females. When it comes to race, Garret County has an exceptionally high percentage of White residents (96.22%), while Baltimore City and Prince George's County have the highest proportion of Black residents (61.56% and 61.23%, respectively). Howard County has the highest proportion of Asian residents (18.41%), and Montgomery County (19.53%) and Prince George's County (18.79%) have the highest proportion of residents who are Hispanic or Latino of any race.

Table 1: Maryland Socioeconomic Status

	Percentage (%)
Education	
Less than 9th Grade	3.87
High School No Diploma	5.57
High School Diploma	24.16
Some College No Degree	18.72
Associates Degree	6.79
Bachelor's Degree	21.79
Graduate or Professional	19.10
Employment	
Employed	94.83
Unemployed	5.17
Gender	
Male	47.51
Female	52.49
Race	
White	50.16
Black	29.38
Asian	6.33
American Indian/Alaska Native	0.19
Native Hawaiian/Pacific Islander	0.03
Some Other Race	0.40
Two or More Races	3.26
Hispanic or Latino (of any race)	10.26

Table 1: County–level Socioeconomic Status

Jurisdictions	Below HS	HS no Diploma	HS Diploma	College No Degree	AS Deg.	BA/BS Deg.	Grad Deg.	Employed	Unemployed
Allegany	2.30	7.87	40.54	20.38	9.56	11.43	7.91	92.54	7.46
Anne Arundel	2.10	4.70	22.71	19.93	7.57	24.74	18.27	95.75	4.25
Baltimore	3.14	5.29	25.20	19.34	7.28	22.75	17.01	94.96	5.04
Baltimore City	4.56	9.91	28.49	19.25	4.93	16.74	16.13	92.44	7.56
Calvert	1.17	4.55	29.10	22.77	8.22	18.72	15.48	96.20	3.80
Caroline	5.13	10.61	39.23	18.81	7.27	11.37	7.59	95.14	4.86
Carroll	1.74	5.12	28.76	19.40	8.03	23.02	13.93	96.56	3.44
Cecil	3.07	6.89	35.22	21.99	7.55	15.32	9.95	94.79	5.21
Charles	2.39	4.05	31.08	23.99	8.49	17.63	12.37	95.63	4.37
Dorchester	3.46	9.54	38.38	22.04	6.71	11.65	8.23	92.75	7.25
Frederick	3.25	4.19	23.79	18.60	8.49	23.21	18.47	95.92	4.08
Garrett	2.82	7.03	42.69	15.66	8.32	13.59	9.89	95.85	4.15
Harford	2.16	4.78	26.01	21.70	8.45	21.38	15.52	95.85	4.15
Howard	2.11	2.38	13.35	14.11	5.37	30.38	32.29	96.13	3.87
Kent	4.10	6.63	31.19	15.79	5.99	21.11	15.18	96.69	3.31
Montgomery	4.97	3.67	13.38	13.16	5.57	27.29	31.96	95.36	4.64
Prince George's	6.78	6.03	25.30	21.13	6.36	19.16	15.23	93.62	6.38
Queen Anne's	1.58	5.17	29.02	19.57	8.16	21.55	14.94	96.88	3.12
Somerset	4.22	12.26	37.06	22.98	7.71	10.81	4.96	90.77	9.23
St. Mary's	3.47	6.16	30.52	19.43	8.36	18.55	13.49	96.51	3.49
Talbot	2.79	6.06	23.13	20.05	8.27	21.25	18.46	97.08	2.92
Washington	3.17	8.71	35.97	22.00	7.78	13.24	9.14	94.51	5.49
Wicomico	4.14	8.20	33.05	19.31	7.30	16.28	11.71	92.10	7.90
Worcester	2.23	5.52	30.84	23.30	8.32	18.68	11.12	94.11	5.89

Table 2: County–level Socioeconomic Status Continued

Jurisdictions	Male	Female	White	Black	Asian	AI/AN	NHPI	Other Race	2+ Races	Hispanic or Latino
Allegany	52.42	47.58	86.60	8.19	0.93	0.15	0.01	0.11	2.15	1.87
Anne Arundel	49.24	50.76	67.06	16.41	3.79	0.14	0.04	0.29	4.24	8.03
Baltimore	46.14	53.86	56.15	28.94	6.03	0.21	0.04	0.35	2.70	5.58
Baltimore City	45.46	54.54	27.30	61.56	2.46	0.26	0.03	0.39	2.58	5.42
Calvert	49.02	50.98	77.63	12.23	1.84	0.14	0.04	0.15	3.83	4.15
Caroline	48.98	51.02	75.17	13.15	0.51	0.12	0.00	0.19	3.31	7.55
Carroll	48.89	51.11	88.34	3.54	2.04	0.25	0.01	0.21	1.92	3.70
Cecil	49.33	50.67	84.36	6.42	1.50	0.08	0.01	0.05	3.02	4.57
Charles	47.36	52.64	38.28	46.86	3.08	0.57	0.02	0.40	4.74	6.06
Dorchester	46.04	53.96	62.56	25.35	1.18	0.04	0.00	0.02	5.10	5.75
Frederick	48.93	51.07	72.40	9.50	4.42	0.17	0.06	0.15	3.32	9.97
Garrett	49.32	50.68	96.22	1.19	0.44	0.05	0.04	0.00	0.87	1.18
Harford	48.31	51.69	75.30	13.61	2.63	0.12	0.00	0.41	3.25	4.67
Howard	48.21	51.79	50.68	18.76	18.41	0.17	0.03	0.54	4.36	7.04
Kent	47.03	52.97	77.40	13.91	1.14	0.08	0.01	0.20	2.79	4.49
Montgomery	47.33	52.67	43.10	17.97	14.89	0.14	0.04	0.66	3.68	19.53
Prince George's	46.18	53.82	12.33	61.23	4.22	0.22	0.03	0.52	2.66	18.79
Queen Anne's	49.12	50.88	85.42	5.84	1.09	0.06	0.00	0.27	3.19	4.12
Somerset	54.66	45.34	51.49	40.60	0.83	0.27	0.00	0.02	3.11	3.68
St. Mary's	49.80	50.20	73.62	14.28	2.57	0.06	0.02	0.17	3.98	5.31
Talbot	46.51	53.49	77.51	10.63	1.31	0.16	0.00	0.17	3.37	6.86
Washington	50.68	49.32	78.01	10.72	1.69	0.14	0.09	0.17	3.74	5.43
Wicomico	46.29	53.71	62.10	25.91	3.01	0.15	0.02	0.38	3.05	5.38
Worcester	47.95	52.05	79.92	12.53	1.17	0.16	0.00	0.20	2.38	3.62

Maryland Vaccination Rates (2021 to 2023)

Table 4 is the descriptive statistics of the vaccination rate within Maryland from 2021 to 2023. As expected, the vaccination coverage is increasing for the entire Maryland state. The statistics on vaccination and cases start from November 9, 2021 to March 15, 2023. The three separate calculations are based on 23 counties of Maryland and the Baltimore City of dates 2023-03-15, 2022-05-15, and 2021-09-18. Compared to 2021, the mean vaccination coverage increased by 13.36%. From 2021 to 2023, the maximum coverage in counties increased by 19.1%; however, the minimum coverage in counties only changed by 9.7%. The standard deviation of the statistics also increased from 8.673% in 2021 to 10.92% in 2023, meaning the variation in different counties increased.

From the percentiles, it was calculated that there are 11 counties falling below the 50% percentile for all three dates−−Allegany County, Baltimore city, Caroline County, Cecil County, Dorchester County, Garrett County, Queen Anne's County, Somerset County, St. Mary's County, Washington County, Wicomico County.

There are 4 counties falling below 25% percentile for all three dates−−Allegany County, Garrett County, Somerset County, Wicomico County.

Table 4: Vaccination Descriptive Statistics

	Mean	STD	Min	25%	50%	75%	Max
2021 Vaccination Coverage	57.12	8.67	42.4	49.50	58.00	62.70	74.7
2022 Vaccination Coverage	67.60	9.99	50.5	59.48	68.85	73.15	87.5
2023 Vaccination Coverage	70.48	10.92	52.1	61.80	71.45	76.18	93.8

Ordinary Least Squares

Table 5 is the result from the regression specification that focuses on race, economic factors, and educations levels across the Maryland counties. The main difference between the Model I and Model II in the table is how the second model accounts for the vaccination rate in the Maryland counties. Based on the first column of the table, one percentage increase in the black population is associated with the reduction of around 6,048 COVID–19 cases. In addition, a higher rates of unemployment, number of households, and poverty rate lead to a higher number of cases. For instance, one percentage increase in the unemployment rate is associated with around 167.44 increase in cases. The result for the educational levels is mixed. For instance, while having an bachelor's degree is associated with 0.12 case reduction, having a graduate degree is associated with an increase of around 0.13 cases. However, based on the standard error of the education coefficients, most of the educational variables are statistically insignificant at the 5% or 10% level, except for those with below high school

education and those with some college without the degree. The lack of significant difference between the R-squared value and the adjusted R-squared value may indicate that overfitting is not present in the model. The R-squared values of around 0.82 implies that around 82% of the variations in case numbers are explained by the independent variables.

The second OLS model from the third column of the Table 5 that includes the vaccination rate has mostly the similar interpretations to the first column of the table. However, the number of household is negatively associated with the number of cases in the second regression model with the coefficient of around -0.75, which indicates that having one more household is associated with 0.75 reduction of cases. The result for the educational levels remain mixed, similar to the first OLS model. Intriguingly, the higher levels of vaccination is associated with a higher number of cases. Indeed, one percentage increase in COVID–19 vaccination is associated with around 1,402 increase in cases. Such phenomena may be due to how the number of cases have been overall increasing throughout the period in which the vaccines were administered. The R-squared value of the second OLS regression is around 0.860 with the adjusted R-squared value of around 0.859 that indicates around 86% of the variations in case numbers are explained by the independent variables. One may infer from the low difference between the R-squared and adjusted R-squared values that there may not be overfitting in the regression model.

Weighted Least Squares

The WLS models are presented in the second and fourth columns of Table 5 with same independent variables as the first and third columns, respectively. The WLS model presented in the second column has mostly the similar interpretations as the OLS model presented in the first column. The main difference between the two models originate from how the African–American population has a different effect on the number of cases. Specifically, in the WLS model, we observe that one percentage increase in the African–American population is associated with around 3,065 increase in number of cases. The effects of education remain mixed with how the magnitude of coefficients decrease as the education levels increase with most of the coefficients being statistically insignificant.[1]

The weighted least square for Model II in column 4 has similar interpretations for the racial and economic factors, except how the increase in one unit of total household is associated with around 0.44 increase in case numbers. The WLS model indicates that as educational levels increase, the number of cases decrease. For example, while having one more person with below high school education increases the case by around 1.09, presence of one more person with an associate degree decreases 0.59 cases. The model also demonstrates one percent increase in the vaccination rate is associated with around 190.63 increase in cases. How-

[1]Please refer to the code for the confidence interval and p-values of the coefficients

ever, this do not necessarily imply that vaccination helps to spread COVID-19. Since COVID-19 cases are cumulative, the case number will increase with time, but vaccination coverage is also cumulative and it is increasing as people become more aware of the pandemic, which suggest a possible explanation for the correlation.

The R squared value of the weighted least squares for Models I and II are 0.8179 and 0.8856, respectively. Based on the R-squared values, Model II has a stronger explanatory power with the socioeconomic factors and vaccination explain around 88% of the variations of the case numbers. Due to how the adjusted R-squared value is not significantly different from the R-squared value, it may be plausible that both models do not have the issue of overfitting.

Table 5: Regression Results

	OLS Model I	WLS Model I	OLS Model II	WLS Model II
const	-2949.2601	-1862.8509	-105997.1300	-16036.9457
	(1508.7164)	(224.8217)	(4918.9549)	(1236.1342)
Race				
Black Percentage	-6048.2906	3065.6137	-19116.5158	-10841.1877
	(5769.3798)	(1421.4674)	(5158.2539)	(1743.2935)
Economic Factors				
Unemployment Rate	167.4391	120.5153	1277.5523	287.2615
	(498.8692)	(91.7215)	(445.9208)	(134.4560)
Total Households	0.0434	0.3118	-0.7487	0.4388
	(0.2504)	(0.0669)	(0.2253)	(0.0865)
Percent Families in Poverty	393.0239	22.6758	2428.4846	621.2946
	(353.9020)	(75.4149)	(327.8762)	(98.3143)
Educational Levels				
Less than 9th Grade	0.6364	0.8329	0.8499	1.0963
	(0.1805)	(0.1475)	(0.1606)	(0.1009)
High School No Diploma	-0.0053	-0.1807	-0.5324	-1.3459
	(0.5967)	(0.1946)	(0.5304)	(0.2324)
High School Diploma	0.0370	0.2492	1.5518	0.4788
	(0.2535)	(0.0589)	(0.2356)	(0.1015)
Some College No degree	0.5426	-0.4038	0.0752	-0.2671
	(0.2602)	(0.1036)	(0.2321)	(0.1113)
Associates Degree	0.1341	0.3726	0.4552	-0.5916
	(0.5525)	(0.2329)	(0.4909)	(0.2490)
Bachelor's Degree	-0.1201	-0.1715	0.0100	-0.2167
	(0.2977)	(0.0799)	(0.2645)	(0.1047)
Graduate or Professional	0.1299	0.0082	0.7257	-0.0993
	(0.1716)	(0.0693)	(0.1548)	(0.0656)
Vaccination Rate				
Pop with Complete Vaccination			1402.1862	190.6280
			(64.4025)	(16.7825)
R-squared	0.8223	0.8179	0.8600	0.8856
R-squared Adj.	0.8212	0.8167	0.8590	0.8849

Note:
[1] The standard errors are in parenthesis
[2] The confidence interval and p–values are provided in the code

References

Center for Disease Control (2023). COVID-19 Vaccinations in the United States,County . https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh.

Hatef, E., H.-Y. Chang, C. Kitchen, J. Weiner, and H. Kharrazi (2020). Assessing the impact of neighborhood socioeconomic characteristics on covid-19 prevalence across seven states in the united states. *Frontiers in Public Health 8*.

Havers, F. P., H. Pham, C. A. Taylor, M. Whitaker, K. Patel, O. Anglin, A. K. Kambhampati, J. Milucky, E. Zell, H. L. Moline, et al. (2022). Covid-19-associated hospitalizations among vaccinated and unvaccinated adults 18 years or older in 13 us states, january 2021 to april 2022. *JAMA Internal Medicine 182*(10), 1071–1081.

Hawkins, R. B., E. J. Charles, and M. J. H (2020). Socio-economic status and covid-19-related cases and fatalities. *Public health 189*.

Johnson, A. G. (2022). Covid-19 incidence and death rates among unvaccinated and fully vaccinated adults with and without booster doses during periods of delta and omicron variant emergence—25 us jurisdictions, april 4–december 25, 2021. *MMWR. Morbidity and mortality weekly report 71*.

Karmakar, M., P. M. Lantz, and T. Renuka (2021). Association of social and demographic factors with covid-19 incidence and death rates in the us. *JAMA network open 4*.

Scobie, H. M., A. G. Johnson, A. B. Suthar, R. Severson, N. B. Alden, S. Balter, et al. (2021). Monitoring incidence of covid-19 cases, hospitalizations, and deaths, by vaccination status — 13 u.s. jurisdictions, april 4–july 17, 2021. *MMWR. Morbidity and mortality weekly report 70*.

State of Maryland (2023a). Maryland Counties Socioeconomic Characteristics. https://opendata.maryland.gov/Demographic/Maryland-Counties-Socioeconomic-Characteristics/is7h-kp6x/data?pane=feed.

State of Maryland (2023b). MDCOVID19 CasesByCounty . https://coronavirus.maryland.gov/datasets/maryland::mdcovid19-casesbycounty/explore.