Examining COVID-19 Health Outcomes and Vaccine Uptake Predictors: Insights for HEW Policy and Outreach

Exploratory Data Analysis

To investigate public health factors during COVID-19, I conducted a regression analysis with two main goals. First, I assessed the relationship between in-person schooling and COVID-19 health metrics to inform school attendance policies during pandemics. Second, I identified key predictors of vaccine uptake by analyzing early 2021 behavioral and attitudinal data to forecast vaccination acceptance as eligibility expanded.

Data Summary:

The HEW dataset, collected by Carnegie Mellon's Delphi group via the COVID-19 Trends and Impact Survey on Facebook, provides county-level COVID-19 indicators in the U.S. for two periods in early 2021. This data captures behaviors, health outcomes, and public sentiment during a critical pandemic phase, supporting analysis of in-person schooling's relationship to COVID-19 incidence and the predictive impact of behaviors and beliefs on vaccine uptake. Key variables include:

- **time value**: Date of observation (format: "YYYY-mm-dd"), marking the week ending on that day.
- **geo value**: County identifier using a five-digit FIPS code.
- **cli**: Estimated % of people with COVID-like symptoms in households.
- **tested 14d**: % of people tested for COVID-19 in the past 14 days.
- **tested positive 14d**: Estimated test positivity rate over the past 14 days.
- **confirmed 7dav incidence prop**: New confirmed COVID-19 cases per 100,000, representing incidence.

- **inperson school fulltime**: % of households with children attending school full-time in person.
- **inperson school parttime**: % of households with children attending school part-time in person.
- **covid vaccinated or accept**: % of people vaccinated or willing to be vaccinated if offered.
- **covid vaccinated**: % of people who have received a COVID-19 vaccine.
- **wearing mask 7d**: % of people who wore masks most or all the time in public in the past 7 days.
- **others masked**: % of people reporting that most or all others in public wore masks.
- **public transit 1d**: % of people who used public transit in the past 24 hours.
- **work outside home 1d**: % of people who worked or attended school outside their home in the past 24 hours.
- **shop 1d**: % of people who visited a market or pharmacy in the past day.
- **restaurant 1d**: % of people who visited a restaurant or café in the past day.
- **spent time 1d**: % of people who spent time with non-household members in the past day.
- **large event 1d**: % of people who attended a large event (10+ people) in the past day.
- **worried become ill**: % of people somewhat or very worried about severe COVID-19 illness.
- **vaccine likely friends**: % of people likely to vaccinate if friends or family recommend it.
- **vaccine likely who**: % of people likely to vaccinate if recommended by the WHO.
- **vaccine likely govt health**: % of people likely to vaccinate if advised by government health officials.
- **vaccine likely politicians**: % of people likely to vaccinate if recommended by politicians.

For my analysis, I chose (confirmed 7dav incidence prop) the 7-day COVID-19 incidence rate as the response variable for assessing health outcomes related to in-person schooling. This metric—new cases per 100,000 people—provides a clear measure of disease prevalence. For predicting vaccine uptake, I selected the percentage of respondents either vaccinated or willing to be vaccinated, as it aligns with understanding factors that drive vaccine acceptance. While the dataset contains other indicators on testing, beliefs, and symptoms, I focused on essential

variables to simplify and enhance model interpretability. The excluded variables and the rationale behind each decision are outlined below:

Variable	Reason for Exclusion
cli (COVID-like illness)	While informative, this variable was excluded due to incorrect calculations that resulted in inaccurate values.
tested_14d and tested_positive_14d	Testing rates and positivity rates, though relevant, are indirectly related to school reopening decisions.
various belief indicators	Belief indicators like 'vaccine likely politicians' offer insight into vaccine attitudes but are less directly tied to COVID-19 health outcomes.

Data Exploration

January 30 acts as a baseline disease load or starting point to help explain the COVID-19 incidence rate in March. This approach helps to account for pre-existing conditions when evaluating how variables like in-person schooling relate to COVID-19 outcomes. So basically, to understand the relationship between in-person schooling in March and COVID-19 outcomes in March, we are using January data as the starting point or baseline. More details have been discussed in the Methods section.

In the analysis, to ensure the integrity of our analysis, I addressed missing values across key variables. Specifically, I removed records with missing values for critical predictors. This approach maintained consistency in the dataset, enabling a robust analysis without imputing potentially inaccurate values.

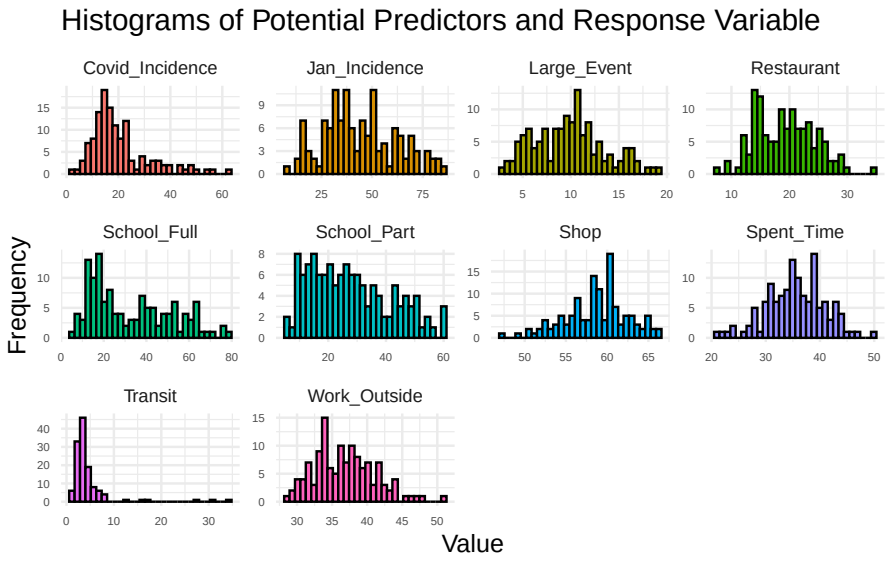

Figure 1: Distribution of predictor variables including the response variable

In Figure 1, the covid incidence rate variable shows a right-skewed distribution, which is indicated by a longer tail on the right side.

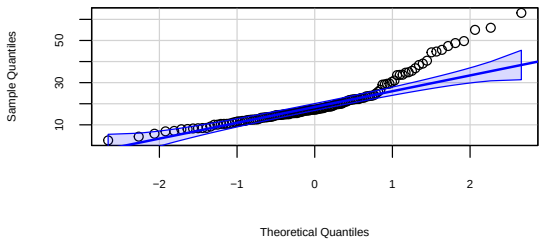

Figure 2: Q-Q Plot of Response Variable

To further assess the normality of the response variable, i generated a QQ plot with an envelope (Figure 2). The QQ plot corroborated our observations from the histograms, revealing that the right tail of the data significantly deviates from the expected theoretical quantiles and the leverage points are addressed will be discussed in the Methods section.

Methods

Applying a log transformation on the confirmed_7dav_incidence_prop (Covid Incidence Rate) variable and helped fix the problem of varying spread (heteroscedasticity), making the variance more consistent across different levels of the independent variables.

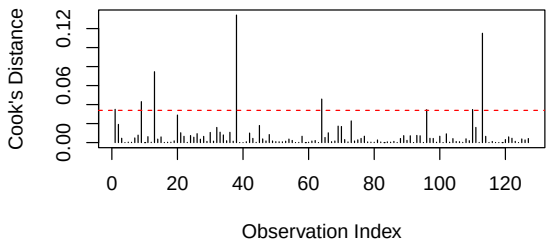

Figure 3: Cook's Distance

I ran the initial model, and the diagnostics indicated four influential points (observation numbers 13, 39, 114, 65). After identifying these influential observations, we removed them from the dataset and re-ran the model, subsequently performing diagnostics on the updated results.

Model - Part a (after all required changes):

$\log(\text{Covid Incidence Rate}) = \beta_0 + \beta_1 \cdot \text{Fulltime School} + \beta_2 \cdot \text{Parttime School} + \beta_3 \cdot \text{Jan Incidence Rate} + \beta_4 \cdot \text{Public Transit} + \beta_5 \cdot \text{Work Outside} + \beta_6 \cdot \text{shop} + \beta_7 \cdot \text{restaurant} + \beta_8 \cdot \text{spent Time} + \beta_9 \cdot \text{large Event} + \varepsilon$

In this part a) analysis, I used the sandwich estimator (heteroskedasticity-consistent standard errors) to address potential heteroskedasticity in the residuals of my linear regression model. Heteroskedasticity occurs when the variability of the residuals differs across levels of an independent variable, which can lead to biased standard errors if left uncorrected.

Diagnostic Plot:

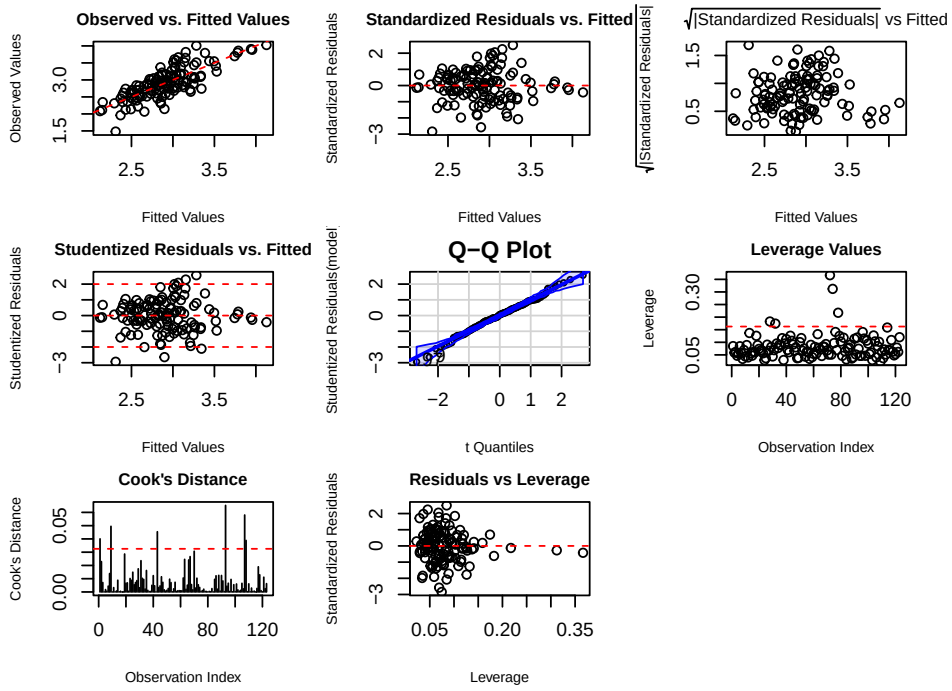

Figure 4: Diagnostic plot of the final model

Overall, this model demonstrates a solid adherence to the assumptions of linear regression.

Coefficient Significance and Effect Size:

Null Hypothesis: There is no relationship between in-person schooling and Covid outcomes.
$H_0 : \beta_1 = 0, \beta_2 = 0$
Alternate Hypothesis: There is relationship between in-person schooling and Covid outcomes.
$H_A : \beta_1 \neq 0, \beta_2 \neq 0$

Predictor Variable	Coefficient Estimate	Lower Bound (95% CI)	Upper Bound (95% CI)
Fulltime School	0.0067	0.0027	0.0107
Parttime School	0.0100	0.0057	0.0142

For vaccine uptake prediction, I split data into training and testing sets, focusing on predicting March vaccination rates. I used best subset, lasso, and ridge methods, choosing lasso for its balance of accuracy (lowest RMSE) and simplicity, retaining essential predictors. With *covid_vaccinated_or_accept* as the response variable, this approach leverages January data to highlight behavioral and belief factors that can guide HEW's targeted outreach in future pandemics.

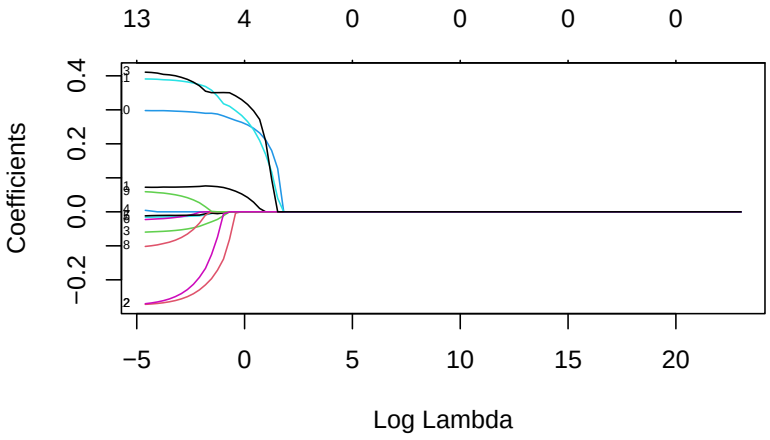

Figure 5: Lasso Regularazitation Path

Figure 5 highlights variables that remain significant as lambda decreases, helping identify stable predictors and an optimal lambda range. It guides feature selection and regularization to balance model complexity and predictive accuracy.

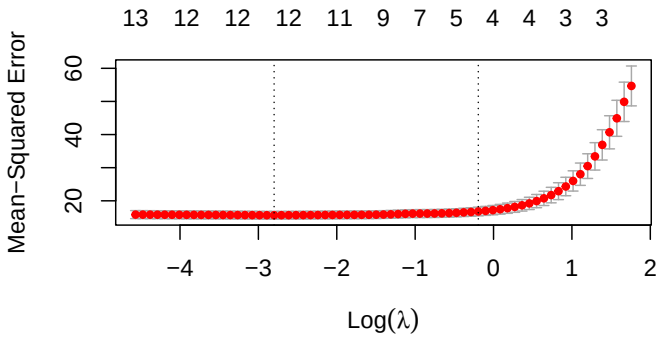

Figure 6: Cross-Validation Results for Lasso Regression

Figure 6 displays cross-validation results for the Lasso model, showing mean squared error (MSE) across a range of lambda values to identify the optimal regularization parameter.

Model (Part b)

$$\text{(Covid Vaccinated/Accept)} = \beta_0 + \beta_1 \cdot \text{Others Masked} + \beta_2 \cdot \text{Public Transit} + \beta_3 \cdot \text{Work Outside} + \beta_4 \cdot \text{Restaurant} + \beta_5 \cdot \text{Spent Time} + \beta_6 \cdot \text{Large Event} + \beta_7 \cdot \text{Worried Ill} + \beta_8 \cdot \text{Vaccine Friends} +$$

$\beta_9 \cdot \text{Vaccine WHO} + \beta_1 0 \cdot \text{Vaccine Govt} + \beta_1 1 \cdot \text{Vaccine Politicians} + \beta_1 2 \cdot \text{Wearing Mask} + \varepsilon$

Summary of Lasso Model Results

Metric	Value
Best Lambda (minimizing cross-validated error)	0.060839
Mean Cross-Validated Error (MSE)	15.650040
Standard Error of Cross-Validated Error	1.137761
Training RMSE	3.725851
Training R-Squared	0.746464

The optimal lambda value, $\lambda = 0.0608$, minimizes cross-validated error, balancing complexity with predictive accuracy. This Lasso model excludes less impactful predictors (zero coefficients) and retains key factors. Significant predictors—such as vaccine_likely_who, vaccine_likely_govt_health, and wearing_mask—highlight the importance of trusted health messaging and mask adherence in vaccine acceptance.

Results

Regression analysis showed a significant relationship between in-person schooling and COVID-19 incidence: full-time schooling had $\beta = 0.0067$, 95% CI $[0.0027, 0.0107]$, and part-time schooling had $\beta = 0.0100$, 95% CI $[0.0057, 0.0142]$, both significant at $p < .05$. Lasso regression for vaccine uptake prediction (optimal $\lambda = 0.0608$) explained 74.6% of variance, emphasizing the role of trusted health sources and safety practices in driving vaccine acceptance.

Discussion

The analyses conducted in this study provide insights into the factors influencing COVID-19 outcomes and vaccination behaviors, aiming to inform policy decisions and outreach strategies for managing public health initiatives during pandemics. For Problem A, which examined the relationship between in-person schooling and COVID-19 incidence rates, the findings indicate a statistically significant association between levels of in-person schooling and COVID-19 incidence rates. For Problem B, exploring the predictability of vaccination uptake based on behavior and beliefs, the analysis identified several behavioral and attitudinal predictors as significant predictors of vaccination willingness. Limitations include potential biases from self-reported data (survey) and limited generalizability, as findings are based on specific counties and a unique pandemic phase, possibly limiting relevance to other regions or future crises.