Indian Statistical Institute, Kolkata

Project for 'Statistical Methods -III'

Bachelor of Statistics (Hons.), 2023-26

COVID-19 Hospitalizations: Impact of Vaccination Amid Rising Cases

Aditya Aryan

BS2305

Advisor/Instructor

Dr. AYANENDRANATH BASU

Professor, Interdisciplinary Statistical Research Unit

Contents

1 Introduction

The COVID-19 pandemic has had a global impact since its emergence. As the virus spread rapidly, it led to widespread illness, death, and significant disruptions to daily life. Governments and healthcare systems struggled to contain the virus, implementing various measures to slow its transmission and which burdened the healthcare infrastructure.

Hospitalizations have been one of the most important metrics in assessing the severity of the pandemic. The number of hospitalizations is directly tied to the number of severe cases, which in turn depends on preventive measures like vaccinations. High hospitalization rates place substantial strain on hospitals, affecting both the quality of care for COVID-19 patients and the treatment of non-COVID medical conditions.

Vaccinations have emerged as one of the most effective tools in reducing the spread of COVID-19 and preventing severe illness, including hospitalization. Vaccines have been shown to reduce the severity of infections and lower the risk of hospitalization, but their impact may not be immediate. It is likely that vaccination rates have a delayed protective effect, particularly as immunity builds up in the population over time.

This project aims to investigate the relationship between new confirmed COVID-19 cases, vaccination rates, and hospitalizations in the United States. Specifically, we focus on the delayed effect of vaccination rates, hypothesizing that higher vaccination rates reduce hospitalizations over time, with the effect becoming more pronounced after a certain delay.

2 Data

The dataset utilized for this analysis is sourced from the Google COVID-19 Open Data Repository and specifically covers the United States over a two-year period. It includes daily records of various metrics such as new confirmed cases, new deaths, new tests, and other relevant indicators, providing a detailed view of the pandemic's progression and impact within the country. More info about the data can be found at *https://health.google.com/covid-19/open-data/*.

The US COVID-19 dataset has over 560 columns of which we will focus on columns like `new_confirmed`, `new_hospitalized_patients`, `new_persons_vaccinated`, `new_deceased` and

other other public health measure factors.

2.1 Data exploration

Starting with the descriptive statistics for new COVID-19 cases, daily hospitalizations and vaccinations. These statistics highlight the variation of these metrics over a time period of two years.

Table 1: Descriptive Statistics for Daily New COVID-19 Cases

Count	Mean	Std	Min	25%	50%	75%	Max
988.0	92905.46	126435.4	0.0	26993.75	55838.5	119412.2	1235521.0

Table 2: Descriptive Statistics for Daily New Hospitalizations

Count	Mean	Std	Min	25%	50%	75%	Max
977.0	5816.72	4753.37	-2858.0	2216.0	4859.0	7372.0	23477.0

Table 3: Descriptive Statistics for Daily New individuals vaccinated

Count	Mean	Std	Min	25%	50%	75%	Max
641.0	411674.5	507879.9	0.0	61056.0	252837.0	492242.0	2556148.0

The scatter plot helps visualize fluctuations in new COVID-19 cases over time. Periods of sharp increases in COVID cases representing waves of infection.

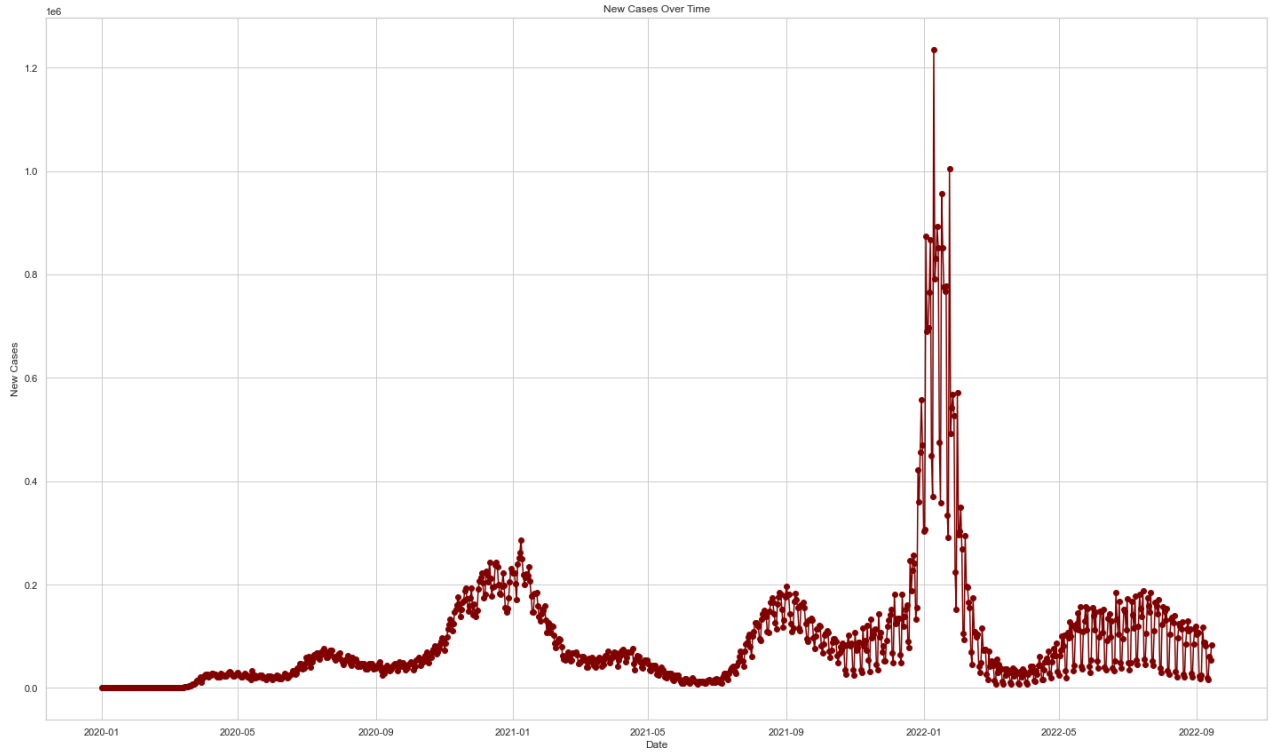

Figure 2.1: Scatter plot showing new cases Over Time

The frequency histogram for daily new COVID-19 cases illustrates the distribution of case counts over the analyzed period. Visually the frequency histogram seems to follow a half normal.

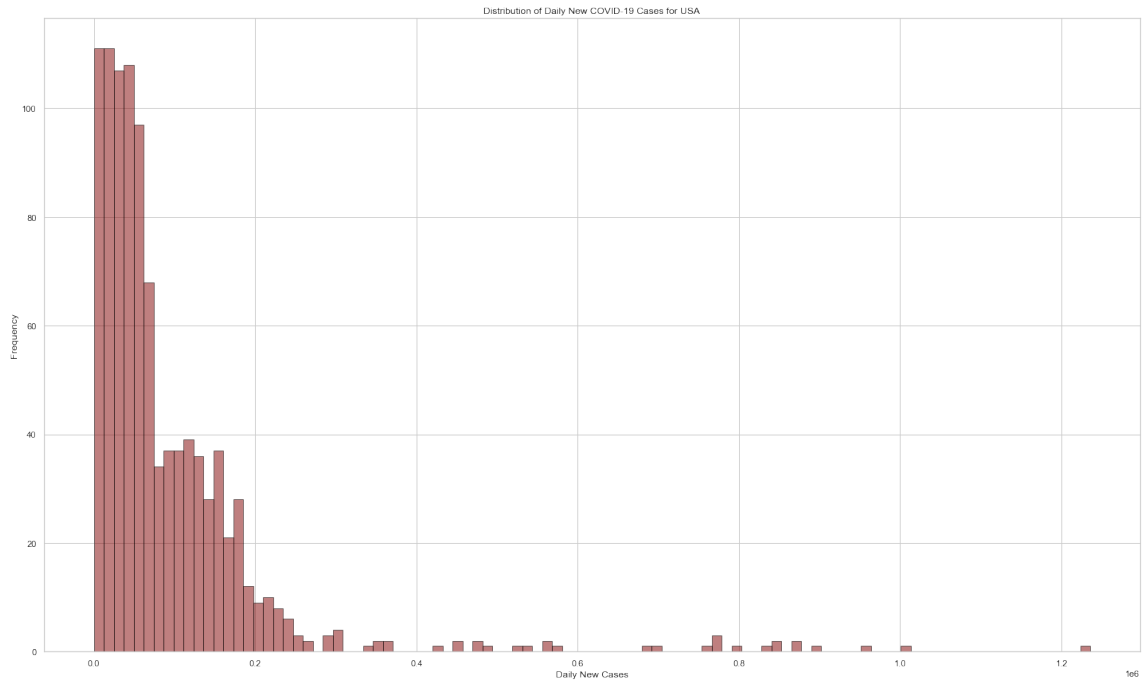

Figure 2.2: Frequency Histogram for new cases

3 Analysis of Factors Influencing New Hospitalizations

We wish to analyse influence of factors such as `new_confirmed` (new daily cases), `new_persons_fully_vaccinated` (daily number of individuals fully vaccinated with second dose) on `new_hospitalized_patients` (daily count of new hospitalizations) using statistical tools such as Pearsons correlation and multiple linear regression.

3.1 Association with number of daily vaccinations

To analyze the delayed effect of vaccination on new COVID-19 hospitalizations, we examine the linear association between `new_hospitalized_patients` (Y) and the vaccination rate X_{-L} lagged by L days. We compute the Pearson correlation coefficient between Y and X_{-L} for each lag $L \in \{0, 1, 2, \ldots, 29\}$ to assess how the vaccination rate at different lag periods correlates with hospitalizations.X_{-L} is created by introducing a lag of L days in `new_persons_fully_vaccinated`

3.1.1 Pearson Correlation Coefficient:

For each lag L, we calculate the Pearson correlation coefficient ρ_L between Y and X_{-L} as follows:

$$\rho_L = \frac{\text{Cov}(Y, X_{-L})}{\sigma_Y \sigma_X}$$

where:

- $\text{Cov}(Y, X_{-L})$ is the covariance between Y and X_{-L},
- σ_Y is the standard deviation of Y,
- σ_X is the standard deviation of X_{-L}.

A negative ρ_L value would suggest that as the vaccination rate increases, hospitalizations tend to decrease, indicating an inverse relationship.

Hypothesis Testing: For each lag L, we test the significance of the observed correlation between Y_t and X_{t-L} with the following hypotheses:

$$H_0 : \rho_L = 0 \quad \text{(No linear relationship between } Y_t \text{ and } X_{t-L})$$

$$H_A : \rho_L \neq 0 \quad \text{(A linear relationship exists between } Y_t \text{ and } X_{t-L})$$

where ρ_L denotes the correlation coefficient between hospitalizations and the lagged vaccination rate. Each computed ρ_L is evaluated for statistical significance using a two-tailed t-test, where the test statistic t is given by:

$$t = \frac{\rho_L \sqrt{n-2}}{\sqrt{1-\rho_L^2}}$$

Here, n is the number of observations. Under the null hypothesis H_0, the test statistic t follows a t-distribution with $n-2$ degrees of freedom. A low p-value (e.g., $p < 0.05$) indicates a significant linear association between Y and X_{-L}, allowing us to reject H_0 and conclude that the relationship is statistically significant at the specified lag period L.

3.1.2 Results Interpretation

A significant negative ρ_L would suggest an inverse relationship, implying that as the vaccination rate X_{-L} increases, the number of hospitalizations Y decreases. This relationship suggests that vaccinations have a delayed, but meaningful impact on reducing hospitalizations, with the optimal lag period representing the time delay at which this effect is strongest.
In this analysis, the most substantial correlation was found at a lag of 24 days, where:

$$\rho_{24} = -0.2019$$

with an associated p-value:

$$p\text{-value} = 4.368 \times 10^{-7}$$

These results indicate a statistically significant inverse association. The negative ρ_{24} value implies vaccination rates have most significant affect on hospitalizations approximately 24 days later. The extremely low p-value supports rejects the null hypothesis H_0.

3.2 Multiple linear regression

Now we try to evaluate the impact of independent variables: new confirmed COVID-19 cases and lagged vaccination rates on dependent variable: daily hospitalizations. It is selected based on the most optimal lag derived from Pearson's correlation with hospitalizations.

3.2.1 Model description

We specify a multiple linear regression model as follows:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon$$

where:

- Y denotes the dependent variable, `new_hospitalized_patients`.
- X_1 denotes the independent variable, `new_confirmed`.
- X_2 denotes the independent variable, `new_persons_fully_vaccinated_lagged`.
- α_0 is the intercept term
- α_1 and α_2 are the regression coefficients, where:
- ϵ is error term assumed to be normally distributed, $\epsilon \sim N(0, \sigma^2)$, capturing unobserved factors affecting hospitalizations.

Estimation and Hypothesis Testing: The parameters α_0, α_1, and α_2 are estimated using the Ordinary Least Squares (OLS) method, minimizing the sum of squared residuals:

$$\min_{\alpha_0,\alpha_1,\alpha_2} \sum_{i=1}^{n} (Y_i - (\alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i}))^2$$

The estimated coefficients, $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$, provide insights into the effect of each predictor on new hospitalizations.

To test the statistical significance of each predictor, we use the following hypotheses:

- **For** α_1 (impact of new confirmed cases):

- $H_0 : \alpha_1 = 0$ (no effect of new confirmed cases on hospitalizations)
- $H_1 : \alpha_1 \neq 0$ (new confirmed cases have a significant effect on hospitalizations)

- **For** α_2 (impact of lagged vaccination rates):
  - $H_0 : \alpha_2 = 0$ (no effect of lagged vaccination rates on hospitalizations)
  - $H_1 : \alpha_2 \neq 0$ (lagged vaccination rates have a significant effect on hospitalizations)

These hypotheses are tested using t-tests, where the t-statistic for each coefficient $\hat{\alpha}_j$ is given by:

$$t = \frac{\hat{\alpha}_j}{\text{SE}(\hat{\alpha}_j)}$$

where $\text{SE}(\hat{\alpha}_j)$ is the standard error of the coefficient estimate. A low p-value (typically $p < 0.05$) indicates that the corresponding predictor has a statistically significant effect on the dependent variable.

Model Evaluation: The goodness-of-fit of the model is assessed using the R-squared value, which represents the proportion of variance in hospitalizations explained by the predictors:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where $\hat{Y}_i$ is the predicted value of Y_i based on the model, and $\bar{Y}$ is the mean of Y. Additionally, the model's overall significance is evaluated using the F-statistic:

$$F = \frac{\text{Explained Mean Square}}{\text{Residual Mean Square}}$$

A high F-statistic with a low p-value indicates that the model as a whole is statistically significant.

3.2.2 Results Interpretation

The multiple linear regression model aims to predict new hospitalizations based on new confirmed cases and the lagged vaccination rate. The results provide insights into the model's performance and the statistical significance of these predictors.

Model Fit:

- **R-squared:** The model has an R-squared value of 0.657. This indicates a relatively strong model fit, with the predictors accounting for a significant portion of the variability in `new_hospitalized_patients`.
- **Adjusted R-squared:** The adjusted R-squared value of 0.656, accounts for predictors that are not significant in a regression model.

Statistical Significance

- **F-statistic:** The F-statistic is 587.4, with a p-value near zero ($p = 3.32 \times 10^{-143}$). This indicates that the regression model is statistically significant.

Coefficient Interpretation

- **Constant (α_0):** The intercept (α_0) is 4095.24.
- **New Confirmed Cases (α_1):** The coefficient for `new_confirmed` is 0.0242, with a highly significant t-value of 33.194 and a p-value of 0.000. This indicates significant positive relationship.
- **Lagged Vaccination Rate (α_2):** The coefficient for the lagged vaccination rate is -0.0005, with a t-value of -2.185 and a p-value of 0.029. This inverse relation is statistically significant between lagged vaccinations and `new_hospitalized_patients`.

4 Conclusion

This study investigates the relationship between `new_hospitalized_patients`, `new_confirmed` cases and `new_persons_fully_vaccinated`, using multiple linear regression analysis to model the impact of these factors. The results provide strong evidence that both `new_confirmed` cases, and `new_persons_fully_vaccinated_lagged` significantly influence hospitalizations. Moreover showing new confirmed cases are positively associated with hospitalizations, while higher vaccination rates are linked to a reduction in hospitalization a delayed effect.

The analysis reveals that vaccination rates have a delayed protective effect on hospitalizations, with the most pronounced impact observed around 24 days after vaccination. This finding underscores the importance of considering the time lag in the effectiveness of vaccination programs, which may not immediately translate to a reduction in hospitalizations.

In conclusion, the study emphasizes the critical role of vaccination in mitigating hospitalizations, particularly in the context of rising case numbers. Public health policies can benefit from this information by understanding the time frame within which vaccination rates are likely to have a significant impact on reducing hospitalizations, helping them in vaccination strategies and healthcare resource planning.