A7 - Project Report

Preston Stringham

DATA 512

A. Introduction

COVID-19 has impacted us all in various ways, but I feel there has been no group of people more negatively impacted by COVID-19 than the hospital and healthcare workers who risk their lives to take care of people with severe cases of the virus. My problem statement addresses the impact of hospitalization rates in Middlesex County, New Jersey and how the introduction of vaccination in the United States impacted this. I intend to investigate the change of hospitalization rate in Middlesex County, New Jersey. In particular, I am interested in the correlation in the change of hospitalization rate and vaccination rates. In addition, I am also interested in how the hospitalization rates of JFK Medical Center in Middlesex compare to the county as a whole. The reason this is a human-centered problem is that it focuses on the impact of both healthcare and non-healthcare workers. From the perspective of non-healthcare workers, more hospital beds and resources could be used to support those suffering from non-COVID related illnesses. From the perspective of the healthcare worker, their work would be less risky as they would interact with less COVID patients than usual. During my research, I was unable to find analysis of this problem so I felt that these topics would be particularly interesting to look into and present in a reproducible manner. I will accomplish this by providing my code and data sources in a public repository, so any person could observe what I did to reach their own conclusions or make additional changes to my work in order to make it more valid.

B. Background/Related Work

There are many questions that I have about the data. In particular, I want to incorporate the work I did in A4 into my research in A5. So, I will be interested in the relationships between vaccination, virus spread, and hospitalization rate in Middlesex County. Again, I feel that the primary research conducted will be about hospitalizations and will be from a human-centered perspective. I want to look at the vaccination rate over time and see if any significant change occurs and how it correlates with the hospitalization rate. I believe that there will be significant changes to the number of hospitalizations before and after the beginning of vaccination. However, considering that vaccinations did not completely resolve the pandemic, I imagine I will see an uptick in the hospitalization rate even after the introduction of vaccinations. A formal hypothesis for this problem might be, "There is no relationship between vaccination rate and hospitalization rate."

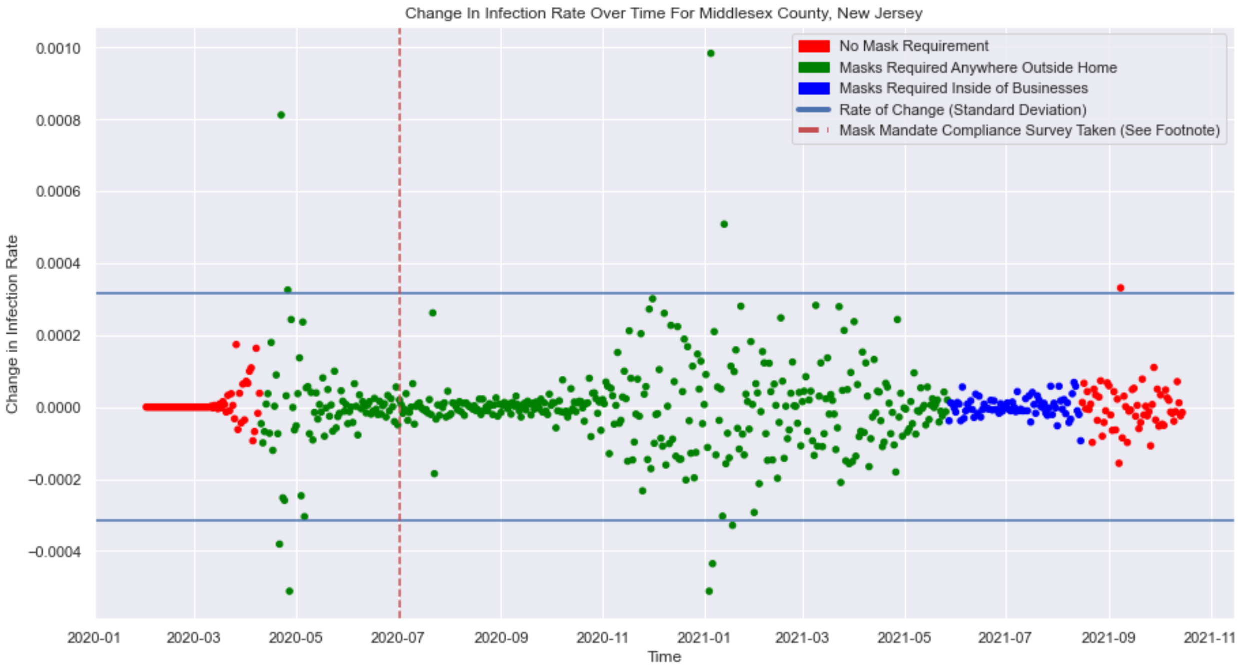
Furthermore, I believe that it would be interesting to compare the mean hospitalization rate in particular cities in Middlesex County, such as Edison city to the rest of the county. A formal hypothesis could be, "There is a significant change in the mean hospitalization rate in Woodbridge Township than the rest of Middlesex County." I think that this is an interesting question since Woodbridge Township is the most densely populated area in the county, so I feel that the hospitalization there would be much greater. It will be interesting to see if that is actually the case based on my statistical testing.

After researching for topics/work related to my problems, I was unable to find any work that directly addresses my questions. However, coincidentally, the state of Washington has published an article discussing the effects of hospitalization which did inspire me to pursue the question regarding the correlation between the change in hospitalization and vaccination rates in

New Jersey. In particular, the paper published by Washington state describes how the unvaccinated portion of the population had 16 times higher hospitalization rate than those who had received at least one dose in recent months [1]. This analysis, however, provides no information about how to reproduce the results they found. They did state their datasets in the paper, but no additional information is given about how the analysis was carried out and what tools were used to find the results presented in the paper. This inspired me to create analysis that is reproducible and available for the public to critique.

For some people who contract the virus, the effects can be deadly. For those who are lucky enough to be able to receive treatment for COVID in a hospital, these patients likely end up in Instant Care Unit (ICU) beds. It is for this reason that I focus on ICU bed usage in my analysis. According to News10, hospitalizations in New Jersey have risen 47 percent in the past two weeks [2], so I found that investigating hospitalizations in particular would give me a better understanding of how COVID is impacting New Jersey residents.

My work on A4, which was specifically regarding the change in COVID infection rate, was surely inspiration for continuing my research into hospitalizations and vaccinations in Middlesex County. The result of this analysis was the figure shown below.



Over 75% of respondants in a survey of mask mandate compliance
in Middlesex County, New Jersey said that they always wear masks.
The survey was taken on July 2, 2020, represented by the red dashed line.

What I found particularly interesting about this analysis was the fact that I could not particularly conclude the validity of mask use. Of course, I personally believe masks to be effective in preventing the spread of COVID-19, but, from the data sources provided, I was unable to provide a valid argument in favor of masks. It is for this reason that I chose to continue to investigate COVID data in particular, because I felt that there was more to the “story” of COVID other than change in infection rate. From this analysis, I also found that Middlesex County was open to mask mandates and that the people in the county were receptive to these mandates. I thought that might be indicative of a community willing to follow health advice and vaccination recommendations.

C. Methodology

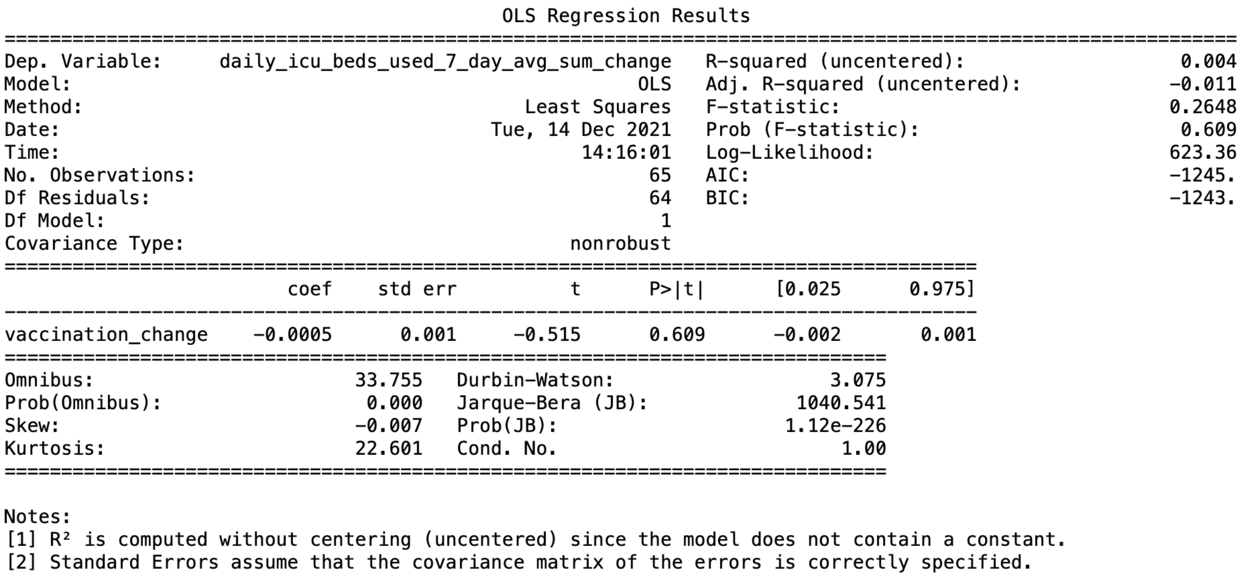
Much preprocessing was necessary to perform my statistical analysis. Hospitalization data was provided on a weekly basis and vaccination was provided on a daily basis, so some conversion was necessary there. In addition, I had to separate all of the hospitals so that I could compare JFK Medical Center to the rest of the county. In addition, all of the data are listed as counts, i.e. count of beds used within a week or count of vaccinations for a given day. Therefore, I had to compute the rates and the change of the rates myself, which is similar to what was done during our A4 assignment.

With regards to the statistical analysis I used, I used an OLS regression using the statsmodels Python package. I also used the ttest_ind package also included in the statsmodels Python package. I used these two statistical tests because I wanted to see if there was a significant correlation between vaccination and hospitalization, where hospitalization is the variable I want to predict. I hoped to find that as vaccinations increased the hospitalizations would decrease. I believe that the use of OLS regression here is ethical as the output from the model does not contain much information about the people in Middlesex County. One could argue that one could compare the results of this analysis to another county or state in order to make some conclusion about the people in Middlesex County, New Jersey, but I don't I was unable to find a more ethical method for such analysis. To the best of my abilities, I tried to show that the statistical methods I used were valid, i.e. showing that the assumptions of linear regression were satisfied. In addition, one could easily perform the analysis themselves if they believe my methods are unethical since my work will be available for the public to view.

I used a T-Test to compare the mean change in hospitalization rates. I chose one of the largest hospitals, JFK Medical Center, located in Woodbridge Township, to compare to the rest of the county. The remaining hospitals in the data are spread out in Edison and throughout the rest of the county. I believe that the use of the T-Test is ethical and called for given how my problem is set up. I want to compare two averages against each other, which is what the T-Test can be used for especially considering my small sample size. Again, considering that my work is publicly available, I find that even if there were mistakes in my analysis, it could be fixed as I could collaborate with other analysts to correct them.

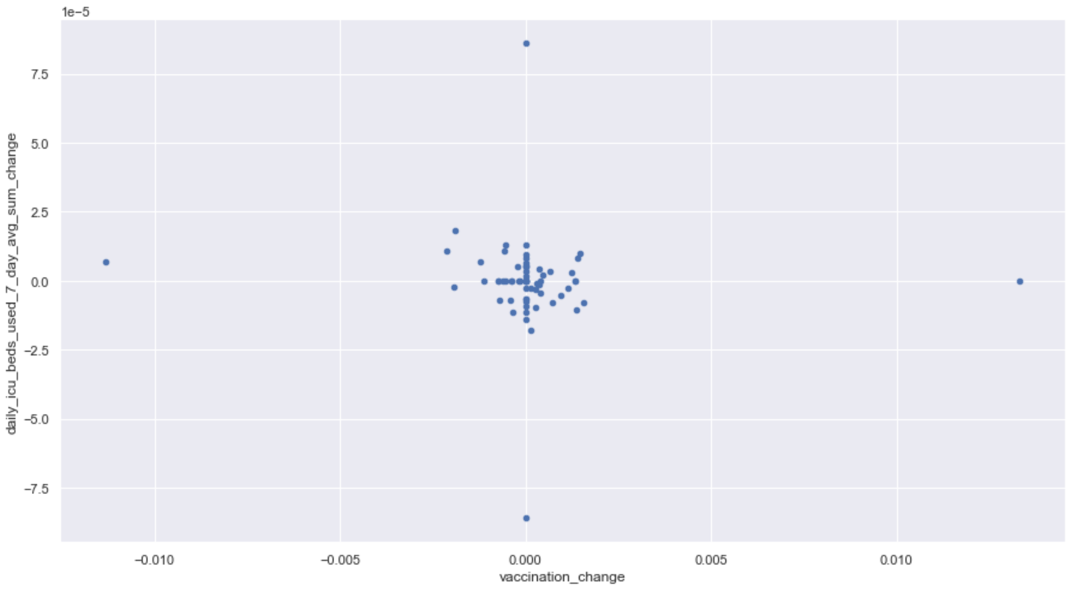
D. Findings

After applying my statistical analysis to my data, I found some particularly interesting findings. After joining my datasets I was only left with 65 weeks to be represented in my data. So, for all tests n=45. For my regression analysis, I had a p-value of 0.609, so I fail to reject the null hypothesis that all coefficients are equal to each other. However, the coefficient of the model was slightly negative, which was what I was expecting. I expected that as the number of vaccinations increased, the hospitalizations would decrease. Of course, correlation is not causation, so it is hard to say whether this relationship has real validity. There could be a number of other factors at play. It could be that many non-COVID patients could be replacing open beds from previously recovered or deceased COVID patients. Also, it could be that the introduction of vaccination in Middlesex County caused people to go out more and possibly spread the virus to more people, thus increasing the hospitalization rate, but in any case, there was no statistical evidence to say that hospitalizations were a good predictor for vaccinations. In addition, there is still a large portion of the population that are not vaccinated in Middlesex County, so those people are still at risk to be hospitalized. The complete results of the linear regression can be found below.



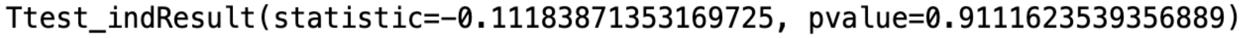
```
                                         OLS Regression Results                                          
=========================================================================================================
Dep. Variable:     daily_icu_beds_used_7_day_avg_sum_change   R-squared (uncentered):                   0.004
Model:                                                  OLS   Adj. R-squared (uncentered):             -0.011
Method:                                       Least Squares   F-statistic:                             0.2648
Date:                                      Tue, 14 Dec 2021   Prob (F-statistic):                       0.609
Time:                                              14:16:01   Log-Likelihood:                          623.36
No. Observations:                                        65   AIC:                                     -1245.
Df Residuals:                                            64   BIC:                                     -1243.
Df Model:                                                 1                                                  
Covariance Type:                                  nonrobust                                                  
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
vaccination_change    -0.0005      0.001     -0.515      0.609      -0.002       0.001
==============================================================================
Omnibus:                       33.755   Durbin-Watson:                   3.075
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1040.541
Skew:                          -0.007   Prob(JB):                    1.12e-226
Kurtosis:                      22.601   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

However, if one looks at the relationship between hospitalizations and vaccinations, this result would probably not be a surprise. The scatter plot between these two variables is shown below, indicating that there is not a strong linear relationship.



Since there is no clear linear regression, I would expect the OLS regression to output what it did.

For my hypothesis about hospitalization rates comparing JFK Medical Center to the rest of Middlesex County, I found a large p-value, which again indicates that I fail to reject the null hypothesis. This results I did expect as I felt that the hospitals are pretty evenly distributed throughout the county. While JFK Medical Center is a large medical center, that isn't to say that the other hospitals in the county are not large. I would imagine people would probably go to the closest hospital to them. The output of the T-Test can be found in the image below.



```
Ttest_indResult(statistic=-0.11183871353169725, pvalue=0.9111623539356889)
```

E. Discussion/Implications

What makes my findings particularly interesting is the lack of statistical significance in having hospitalizations as a predictor variable for vaccinations. This has implications for nurses in Middlesex County that vaccinations may not be helping reduce the amount of work for health care workers who help patients in ICU beds. I would like to think that there is some real negative correlation between vaccinations and correlations, but perhaps the introduction of vaccines caused more people to return to a normal way of living and spread the virus more. It's hard to say. For future research, one could conduct this analysis on more data. Over time, this dataset will continue to grow and one may be able to find more conclusive evidence about whether there is some correlation between vaccination and hospitalization, but I do think that this is a case where correlation is not causation. There are so many different factors that contribute to virus spread and hospitalizations.

With regard to the comparison between JFK Medical Center and the rest of Middlesex County, this could mean that all the hospitals in Middlesex County are equally impacted by the virus. Or, at least all of them could have similar hospitalization rates. The hospitalization data is focused on times during COVID, but it could include other non-COVID related ICU hospitalizations. For future research, one could compare each of the hospitals to one another. I would be curious to see how hospitals or medical centers in more rural parts of Middlesex compare to the large cities. I wonder if people in rural areas who are facing serious COVID complications would be more inclined to seek health care in a larger city, that could have facilities with better care, than those in more rural parts of the county. In addition, it would also be interesting to compare these results to another county or perhaps the nationwide average. This type of analysis would help give context and understanding for how hospitalizations really compare.

F. Limitations

There were a few limitations in this project, specifically the data and the statistical method assumptions. The data itself was limited to COVID times only. That means from around Fall of 2020 to current days. What would have probably been more informative is to observe the hospitalization rates before and after COVID to get a better understanding about how "busy" hospitals are. This would have given me a better understanding of exactly how COVID was impacting medical workers. Without this data, we can only observe the impacts on hospitalization based on the introduction of vaccination, which is what I was interested in, but it is still hard to understand the exact impact on the healthcare system in Middlesex due to this limitation. This was due to my selection of a dataset, which was useful, but did not include any data before COVID. It is likely that this data exists and that this analysis can be expanded to include a sort of context to my hypothesis.

Another limitation was about how this data failed to satisfy a few of the assumptions of linear regression. If you look at the scatter plot between change in vaccination and hospitalization rates presented above, there is no clear linear relationship between the two variables. In fact, it is hard to find any precise relationship between the two variables. Even if the data is time shifted, there does not seem to be a way to make a relationship between the two variables. Thus, the results of the linear regression showed that this probably was not the best method to use.

Furthermore, it is very unclear to me the licenses for the data and how the data can be used. The government websites from which I retrieved the data did not include any information about the licenses for the data. After even deeper research regarding the licenses, I was unable to find additional information that indicates how the data can be used. As I have seen such data used by large publications, I would imagine that it can be used for my purposes, but it is extremely unclear to me about the ethics of using this data for analysis such as mine.

G. Conclusion

In conclusion, this has been a more in depth analysis of COVID in Middlesex County, New Jersey. After observing the infection rate of COVID in this county, I wanted to investigate vaccinations and hospitalizations further. I wondered if there was a statistically significant correlation between vaccinations and hospitalizations, which I was unable to come to a conclusion to due to the high p-value of my OLS regression test. In addition, I investigated whether the hospitalization rate between JFK Medical Center in Middlesex County and the rest of the county itself were significantly different, which I also was unable to come to a conclusion due to the high p-value of my T-Test. While I was surprised about the results of my OLS regression test, I did expect a high p-value for the comparison of hospitalization rates, as I believed that they should be about the same. What one could take away from this is that the hospitals across the county *may* be equally impacted by COVID even though that is not likely to be the case. I say this because it is likely that the hospitals vary in ICU capacity. Further research could be conducted about hospitalization before and during COVID as well as further statistical testing about the change of hospitalization rate through all the hospitals within Middlesex County.

H. References

[1] Washington State Department of Health. (2021). *COVID-19 Cases, Hospitalizations, and Deaths by Vaccination Status* [PDF].

[2] Craig, D. (2021). COVID hospitalizations on the rise across the US, especially in these states. Retrieved 14 December 2021, from https://www.news10.com/top-stories/covid-19-hospitalizations-on-the-rise-across-the-us-especially-in-these-states/

I. Data Sources

- COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Licensed under CC BY 4.0, obtained from https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
- U.S. State and Territorial Public Mask Mandates From April 10, 2020 through August 15, 2021 by County by Day, unable to find license, obtained from https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i

- Mask-Wearing Survey Data, The New York Times, license found here, obtained from https://github.com/nytimes/covid-19-data/tree/master/mask-use
- COVID-19 Reported Patient Impact and Hospital Capacity by Facility, unable to find license, obtained from https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/anag-cw7u
- COVID-19 Vaccinations in the United States,County, unable to find license, obtained from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh