Caleb Hong, Shusheng Li, Noel Ngui, Zilu Wang
Professors Eaman and Ramesh
Data 102
8 May, 2023

Final Project Written Report

Data Overview

For the Bayesian Hierarchical Modeling question, we solely focused on using Google's Daily Community Mobility Data – specifically that of Alameda County. Google's data provides insights into how communities across the globe have been differentially impacted by the effects of the COVID-19 pandemic. In particular, Google's data used mobility relative to a pre-pandemic baseline (with a normal baseline value calculated for each day of the week as the median mobility during the five-week period from January 3, 2020 to February 6, 2020) in a variety of different categories–retail/recreation, groceries/pharmacies, parks, transit, workplace, and residential buildings–as a proxy for their utilization.

The provided data is a census of all users who have turned their Location History setting on through their Google Accounts. It is important to note, however, that by default Location History is turned off for Google users. As such, it is only those who have enabled Location History that will be included in the data. This may potentially lead to selection bias within our data as users must opt into having their data collected/used (e.g. for real-time traffic updates, find your phone service, etc.). By accepting the terms and conditions and enabling Location History, users are theoretically aware that their location data is being collected and used although this may not prove true in practice. However, measurement error should not be a concern because the information is collected directly from users' mobile devices and Google Accounts.

As mentioned above, the Google data included mobility scores for six different categories with each row containing data pertaining to a single date ranging from February 15, 2020 to October 15, 2022. It is important to note that data was taken from individual users' devices but de-identified and aggregated for differential privacy concerns by adding artificial noises to Google's datasets, which enables Google to generate analysis while keeping users' activity data private and secure. For areas where the number of users on a given day was too low and anonymity could not be guaranteed, mobility data was omitted. In the Alameda County data, no dates had data that were omitted. That being said, two columns (besides the other mobility scores not considered in this analysis) were dropped due to missing and irrelevant data; these columns were column_metro_area and iso_3166_2_code.

For the Causal Inference question, we used Google's Daily Community Mobility Data for all counties in California rather than just Alameda County. On top of this, we also used additional data sources including fully vaccination counts, education, income, population, and race for Californian data all come from census data. We also had data on median age per county from the Federal Reserve Economic Data (FRED). These additional data in the causality question were included to account for confounding variables. Data on education, income, and population were from data.census.gov, which is the official source of census data for the U.S. government. We pulled vaccination data from data.chhs.ca.gov, which is the website for the California Health and Human Services Agency. The Census data actively tries to avoid convenience sampling and selection bias. However, since the data comes from self-reports, measurement error could arise from people incorrectly reporting information about themselves. According to census.gov, the Census Bureau uses data from many different sources, and some

data are collected directly from surveys, while others are collected from different sources. The U.S. Census is a trusted curator, and hence uses differential privacy by injecting statistical noise into the data.

Since each row in our data represents a county within California, our data's granularity is quite large, and so we cannot make assumptions about the data of individual people living in a certain county. We also take the means of statistics across multiple days—and so time wise our data's granularity is quite large. Therefore, we cannot make assumptions about individual days. We would like to have accurate data for all of our variables in the year 2022 but for a lot of our confounders in particular, we had to use data from 2020 and 2021 since data for 2022 is too recent to obtain. Additionally, features that we would like to have are changes in covid policies and also medical history of each county. However, these are very hard to practically measure for each county. We excluded people not living in California and also population below 18 years old for bachelor degrees. On top of that, we also excluded 5 counties across all data to exclude NA values; specifically they were Alpine, Sierra, Mariposa, Modoc, and Trinity. Dropping these few counties may have affected our results slightly.

In addition to dropping counties with missing data, we cleaned the data by formatting column headers, transposing dataframes, and located relevant columns for the data. We also created variables that were in percentages, hence dividing the raw count by the population for that county. We also had to filter data to get data only after January 1, 2022. (Most of the data cleaning was very technical, as seen in the Jupyter notebook). Using percentages allowed us to consolidate data for each county. As a result this means that each person has a different impact depending on what county they are in. People in counties with a smaller population would have a greater impact and vice versa. However using percentages helps the model because counties are now on the same percentage scale rather than a raw number that doesn't give us much interoperability. Using data only after January 1, 2022 allowed us to only focus on certain days, which may affect how generalizable our results can be to other stretches of days.

Research Question 1

Our first research question investigates how has COVID-19 impacted the utilization of retail and recreational activities in Alameda County. Although it is undeniable that COVID-19 is still around us, many have seemingly moved on from the quarantined life in 2021. Here in the United States, the Federal Public Health Emergency is slated to end May 11, 2023, and yet despite the end of the Federal PHE, the COVID-19 virus will not simply disappear. There is no clear, definite end to COVID-19, which can make efforts to study the effects of coronavirus difficult. By quantifying the impacts of COVID-19 on mobility in activities, such as retail and recreation or groceries and pharmacies, policymakers and government officials can utilize these findings to better understand the needs of the people from different regions after the pandemic as well as better support sectors that have been severely impacted by COVID-19 socioeconomically.

We will be using Bayesian hierarchical modeling to answer our first research question. This method is a good fit for quantifying the impact of COVID-19 because it estimates the designated parameters of interest of the posterior distribution using observed data while accounting for uncertainty. We will be using PyMC3 to sample for the posterior distribution of the mean of mobility scores.

Despite the compatibility of the method with our research question, there are quite a few limitations to using Bayesian analysis. The model lacks guidance regarding how to select a good

prior; therefore, the process of using Bayesian inferences to formulate a prior can be very subjective and arbitrary. On top of that, the outputted posterior distributions can be highly influenced by the prior beliefs, which are constructed with subjectivity and personal biases. With an increasing number of parameters for estimation, the training time of the Bayesian hierarchical model can also increase drastically.

Research Question 2

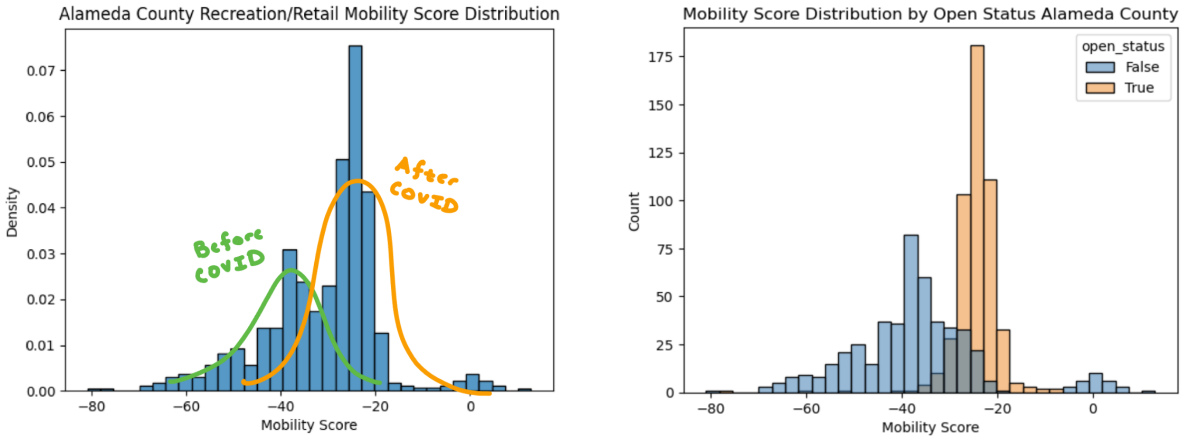
How do high vaccine rates in a county affect average mobility scores? Answering this question can help policymakers and economists decide how to best improve the economy coming out of the pandemic. If there is a causal effect from counties with high vaccination rates and mobility scores, then for those authority figures who only care about how to best improve the economy could push for vaccination policies accordingly in order to increase recreation and retail mobility. If there is no effect, then authorities would not worry about vaccine rates from an economic standpoint but instead a public health standpoint. They will then look for other ways of improving the economy.

To answer this question, we used causal inference because this method treats counties with either a high or low vaccine rate as a treatment variable and their average mobility scores as an outcome variable. Hence we can form a causal DAG and take into account all confounding variables that could affect both the treatment and outcome and use the unconfoundedness assumption to eliminate the confounding effects. Therefore, we can ultimately analyze the true average treatment effect of high vaccinated counties on average mobility scores.

For causal inference, a big limitation would be the unconfoundedness assumption. Since this is an observational study, we can't bring randomness into the setup and instead have to assume unconfoundedness. This requires us to have a good domain knowledge of what confounding variables would affect both treatment and outcome. On top of this another limitation is the lack of data. Even if we have perfect domain knowledge and are able to identify all confounding variables, the data will not exist for every single one of them. Hence, there is no way for us to actually achieve unconfoundedness; we have to instead assume it so we can get an estimate of the average treatment effect. If there is a significant amount of additional confounders that we did not include, this will also cause poor performance since we won't have an accurate estimate of the average treatment effect of high vaccine counties on average mobility scores.

EDA

For our first research question, we created a binary variable called *open_status*, to separate the dates into before and after when California implemented the Reopening Plan on June 15, 2021. The plan aimed to fully reopen California's economy by removing all capacity limits and physical distancing requirements. We then defined a quantitative variable, called *mobility_score*, to be the *retail_and_recreation_percent_change_from_baseline* variable from the 'Google: Daily Community Mobility Data.'



The illustrations above are visualizations from our EDA. The left plot shows the distribution of daily mobility scores in Alameda County from February 15, 2020 to October 15, 2022. The distribution curve roughly suggests the *Mobility Score* for each are pulled from two separate Gaussian distributions: one which we will say is a pre-new-normal distribution and another which represents the new-normal distribution. The plot on the right displays the same data but stratified by the variable "*open_status*". It is clear from the right-hand histplot that the distribution of "new-normal" *Mobility Score* (*open_status* = True) has a median to the right of the pre-"new-normal" (*open_status* = False) distribution.

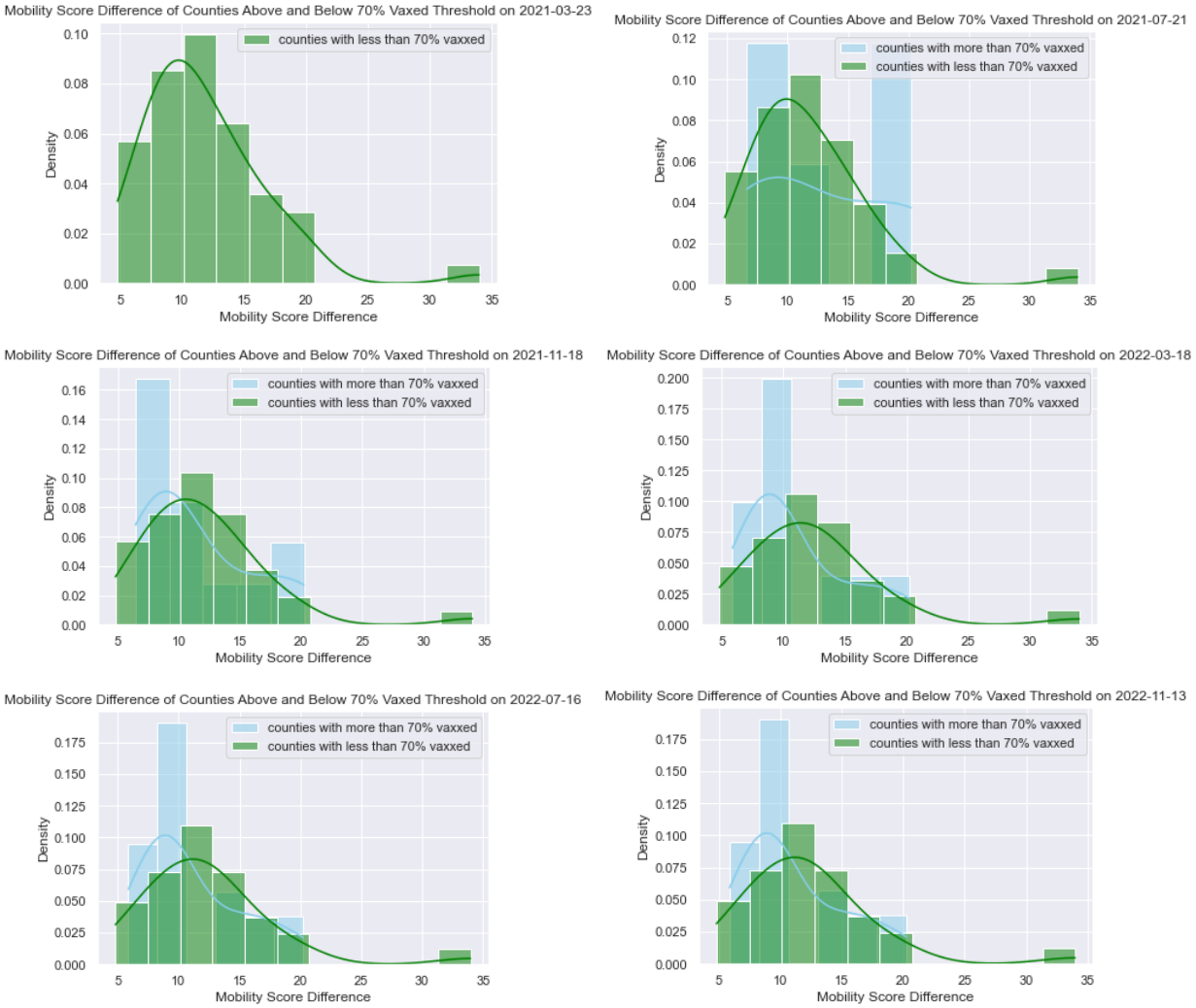
The visualizations are extremely relevant to the question on the effects of COVID-19 on utilization of retail and recreational activities. The results are able to show that the *Mobility Score* from pre-open-status to post-open-status have decreased towards becoming closer to the baseline value, suggesting that there is more retail and recreational activities after the reopening date. This helps provide a potential answer to our question by confirming how the pandemic has impacted retail and recreational activities, guiding us to conduct more data analysis to quantify the effects of COVID-19 using Bayesian inference.

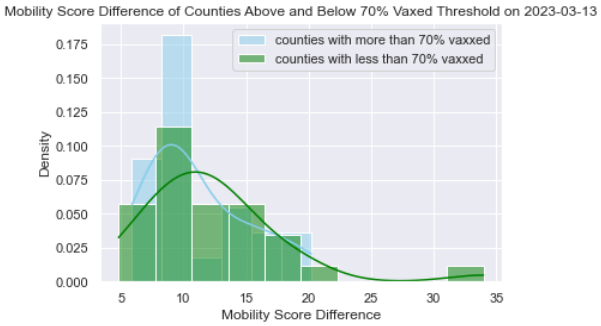
Moreover, we were initially interested in determining when exactly the shift to a "new-normal" occurred by trying to include Z (Bernoulli variable indicating whether a data point belongs to pre-new-normal or new-normal), explained more in details in the method section, as an unknown variable so that we could use the data to help us determine when this shift to a post-COVID new normal occurred. Our effort in trying to achieve this was largely unsuccessful, but this helps provide insights into a possible extension of our current research question and potential future research. The right-hand plot shows overlaid distributions of *Mobility Score* stratified by *open_status* with *Mobility Score* that correspond to "open" dates being those from dates on or after June 15, 2021. The stratified distributions suggest that our hypothesized Gaussian mixture model for *Mobility Score* is an accurate one.

Transitioning to our next research question, we used the 'retail_and_recreation_percent_change_from_baseline' column from the global CSV from https://www.google.com/covid19/mobility/, which shows percent change in length/visits to retail and recreation places from the baseline, which is the median value for the corresponding day of the week, during the 5-week period (Jan 3-Feb 6, 2020). *Mobility Score Difference* we defined then, is the difference in the mean value from that column for days before and after June 15,

2021, which is the date of California's reopening. We use this score to represent the average change in "mobility" for each county before and after California's reopening. *Greater than 70% vaccinated?* (Categorical Variable) is a binary variable, which represents whether or not a county has a population where at least 70% have received one dose of a vaccine.

We'll use *Mobility Score Difference* to quantify the mobility of each county. We would like to quantify the causal relationship between vaccination and mobility, because that's our question. We plotted 70% vaccination status against mobility score for multiple dates from 2021-2023 to pick a date to look at, and also to determine if there was some sort of difference of mobility score between counties with 70% vaccinated.

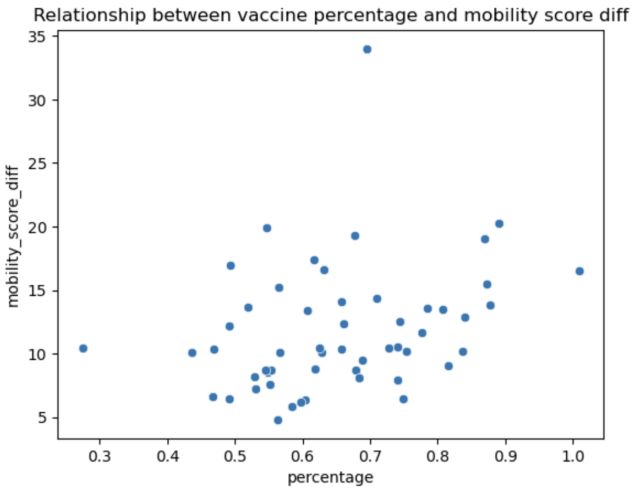


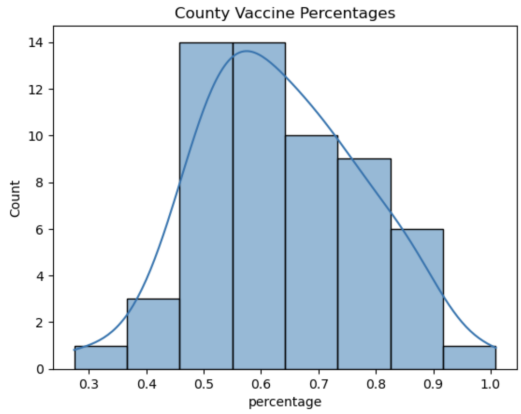


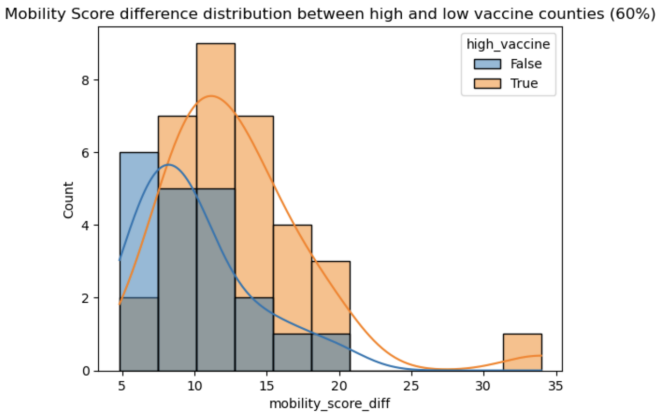
The dates "2022-03-18" through 2023-01-01" seems like a good place to start, since it is a date that is a good amount of time away from when vaccinations were made available, and should therefore represent a more stable estimate for percent vaccinated for each county, and it seems according to the plots that the histograms have somewhat stabilized as well. We plotted a few histograms to view the relationship between percent vaccinated and mobility score difference for each county to confirm if 70% was a good threshold to pick, based on how far the two distributions are from each other.

It seems across the board that counties that have a larger proportion of vaccinated people have a lower mobility score difference. Based on our definition of mobility score difference, this could be because counties that have lower vaccination rates stayed inside more before the reopening date, or counties with lower vaccination rates went outside more after the reopening date, or a combination of both, or as the result of a potential confounding factor. There does seem to be some correlation here between vaccination status and mobility score difference, which encourages us to do more analysis to figure out if there is any causality between vaccination and mobility score difference.

Next, to study the effects of a county's vaccine status on mobility score rate, we graphed a scatter plot of percentages with mobility_score_diff (pre-open vs post_open) to see the relationship between the two. Then a histogram of the vaccine percentages (fully vaccinated / population) each county has on April 19th, 2023 to determine what is the best threshold to use for high vs low vaccine rate. Deciding to go with a threshold of 60% vaccination rate for high vaccine counties and below 60% to be low vaccine counties, this gives a somewhat equal amount of data on the control and the treatment group.







This finding of somewhat similar distribution between both groups without controlling for confounders seems like vaccine status does not have too big of an impact on the change in mobility score between when the county was open mandate and non-open mandate. This leads us to wonder whether other confounding variables have a larger effect on mobility score difference.

Bayesian Hierarchical Modeling: Quantifying the Effects of COVID-19 in Alameda County

To address this question, we will be using Bayesian Hierarchical Modeling. As specified in the EDA portion, we hypothesize that the mobility score data is made of two groups: one that comes from a pre-new-normal distribution and another that comes from a new-normal distribution.

Random Variables

Let Z_i be the group that the i[th] data point belongs to (with $Z_i = 0$ referring to a pre-new-normal datapoint and $Z_i = 1$ referring to a new-normal datapoint)

- $Z_i \sim \text{Bernoulli}(\pi)$

Let X_i be the retail/recreation mobility score for the i[th] datapoint

Let μ_0 be the mean pre-new-normal retail/recreation mobility score

- $\mu_0 \sim \text{Normal}(\mu_p, \sigma_p^2)$ for fixed μ_p, σ_p

Let μ_1 be the mean new-normal retail/recreation mobility score

- $\mu_1 \sim \text{Normal}(\mu_p, \sigma_p^2)$ for fixed μ_p, σ_p

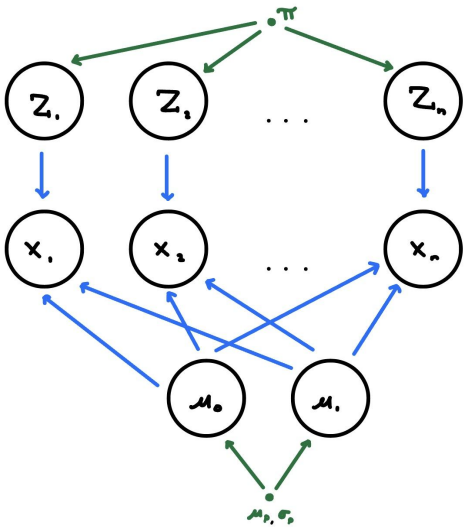
Hyperparameters

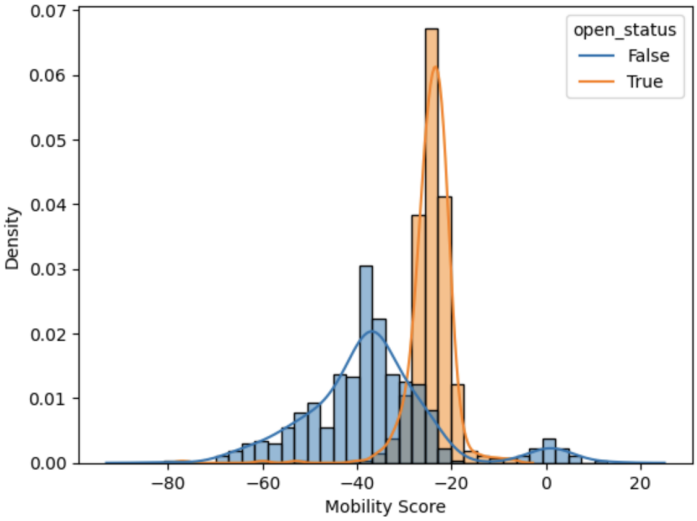
π: Bernoulli hyperparameter for Z_i

μ_p, σ_p: Normal hyperparameter for distribution of both pre-new-normal and new-normal mean retail/recreation mobility scores

Graphical Model

As stated above, the unknown variables we are trying to infer are μ_0 and μ_1 (the mean of the pre-new-normal mobility score distribution and the mean of the new-normal mobility score distribution, respectively).

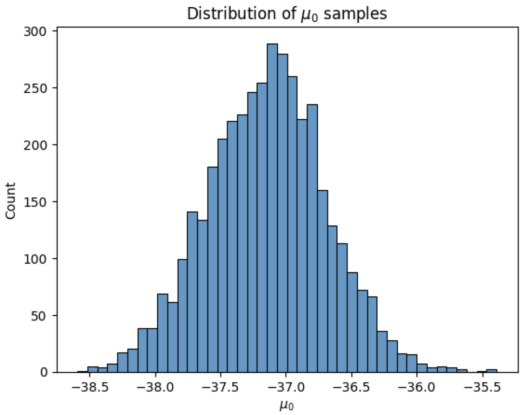


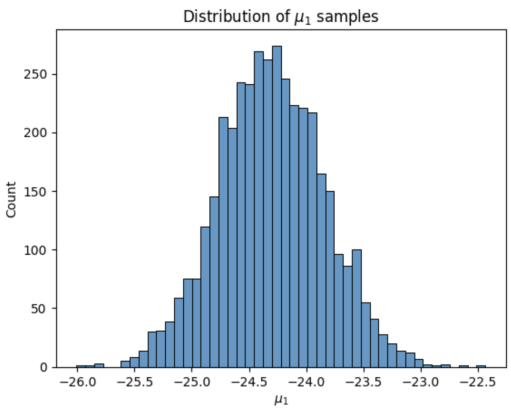


There is a lack of domain knowledge about the data; therefore, there will be limitations to how good our chosen prior can be. We plotted the distribution of pre-new-normal and new-normal mobility scores as overlaid histograms; our observations tell us that both follow a roughly normal distribution. For the pre-new-normal distribution, we calculated the mean of pre-open status mobility scores in Alameda County to be approximately -37 and set that as the value of μ_0 . We then chose $\sigma_0 = 14$ because the distribution of pre-new-normal mobility scores (the blue histogram) has a standard deviation of roughly 14. For the new-normal distribution, we followed the same calculations and set $\mu_1 = -24$. We then chose a much smaller value $\sigma_1 = 5$ because the new-normal distribution is a lot more concentrated compared to pre-open distribution. The mean and standard deviation of the new-normal distribution (the orange histogram) was roughly -24 and 5, respectively, which motivated our choice of prior.

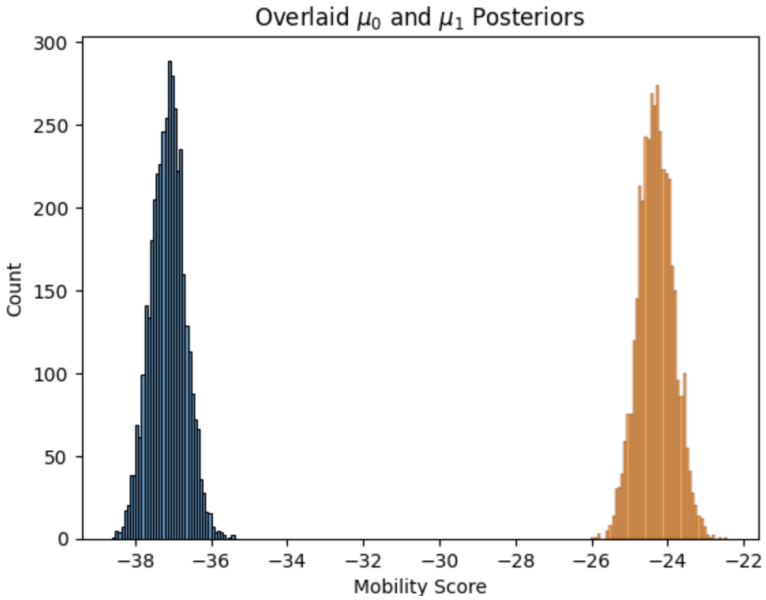
Results

We trained separate PyMC3 model for pre-open and post-open status to find the posterior distribution of μ_0 and μ_1 respectively. The samples seem to center around the calculated value of μ_0 and μ_1 mentioned previously.





Lastly, we plotted the and μ_1 samples on the same plot, shown in the visualization below. The two distributions are very separated from each other, therefore showing that mobility scores for post-open dates generally have a much smaller percent change from baseline compared to those of pre-open dates. In conclusion, we quantified the effects of COVID-19 on utilization of retail/recreational activities to be 12.843 as the difference in means between the sampled μ_1 and μ_0 distributions.



Discussion

Originally, when using one PyMC3 model and treating Z as an additional unknown to help determine when the transition to a new-normal occurred alongside quantifying the effects of COVID-19 on recreation/retail activities there was trouble converging. There would frequently be too few accepted samples in the new-normal category which resulted in a wide posterior for pre-new-normal and a very skinny new-normal posterior distribution. We speculate that the inability to fully flesh out the two posteriors is the result of the two distributions (pre-new-normal and new-normal) being too similar and the sampler thus struggles to generate samples from the true posteriors.

We initially tried training only one PyMC3 model for the sampling of both μ_0 and μ_1; however, we ran into errors, such as 'divergences after tuning', 'acceptance probability does not match the target', and 'estimated number of effective samples is smaller than 200 for some parameters'. We tried increasing 'target_accept', reparameterization, and increasing the number of tuning steps, but the results were still undesirable. As a result, we then moved onto training separate PyMC3 model for μ_0 and μ_1 because the previous model had trouble differentiating between the two parameters.

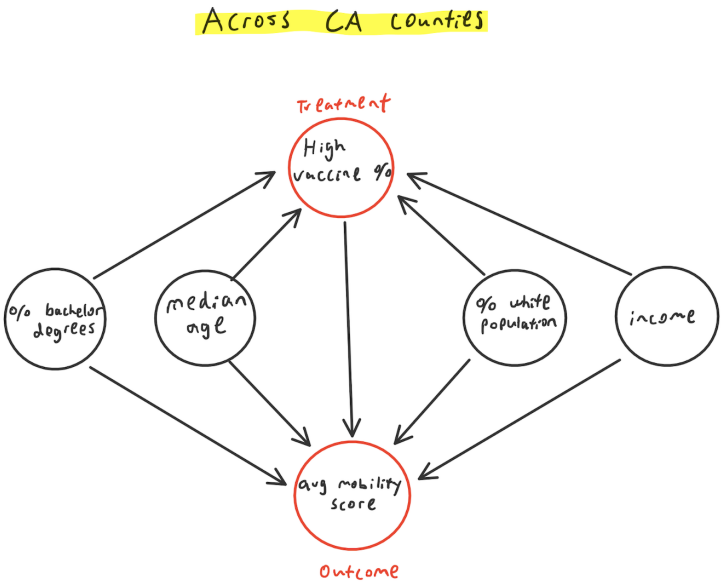
Although the data provided was sufficient for answering the research question at hand, the inclusion of more data (e.g. the expansion of the research question's scope to include all nine Bay Area counties or all of California) would likely enable us to look into investigating when the transition to a new-normal occurred in California.

Causal Inference: Relationship Between County-Level Vaccination and Mobility Scores

The treatment of interest was the binary variable *high_vac* (determined by whether the county had > 55% fully-vaccinated rate on 1/1/2022) and the outcome of interest was *avg_mobility* defined as the average mobility score for each county from 1/1/2022 to 10/15/22. In this case, four primary confounders were identified:

- *bachelor_perc*: the county-wide percentage of people with a Bachelor's degree in 2020
- *median_age*: the county-wide median age in 2021
- *white_perc*: the county-wide proportion of self-identified White individuals in 2020
- *income*: the county-wide median household income in 2020

No collider variables were identified. The causal DAG is shown below:



Obviously due to limitations of our domain knowledge and data, we cannot say that we have accounted for all confounding variables. We have included multiple variables that we believe have a confounding effect on both treatment and outcome. Therefore we tried our best to come up with an unbiased estimate of the average treatment effect. Regardless, we are using Outcome Regression to adjust for confounders. Assuming unconfoundedness given the confounding variables and assuming that the linear model correctly describes the relationship between the variables. The OLS regression predicts *avg_mobility* based on the treatment variable *high_vac* and confounders *median_age*, *white_perc*, *bachelor_perc*, and *income*.

Results

We hypothesized the vaccination percentage of a county to have a positive effect on the average mobility score of a county since more vaccination means that people would be more comfortable to go outside. By naively fitting an OLS regression on *avg_mobility* from *high_vac*, we got the average treatment effect of *high_vac* to be -14.2. This naive interpretation means that a county with a vaccination percentage higher than 55% would actually decrease the average mobility score by 14.2, which is opposite from our hypothesis. In terms of statistical significance, the naive model of the ATE for *high_vac* had a 95% confidence interval between

-17.33 to -11.07 and a p-value of 0.00 which means that this variable is deemed statistically significant.

Next, we fitted a new OLS model including all the confounding variables. Assuming unconfoundedness given the confounding variables and also assuming that all relationships across variables are correctly described by the OLS linear model; in particular, assuming that the effect of each confounder is the same in both the treatment and control group. The results were very shocking because the ATE of *high_vac* decreased to -2.24 with a 95% confidence interval between -7.13 and 2.64, and a p-value of 0.36; meaning that this variable is not statistically significant. The coefficient of the confounders were the following: *median_age* = -0.06 (not statistically significant: p-value=0.79 and 95% CI from -0.56 to 0.43), *white_perc* = 24.72 (statistically significant: p-value=0.04 and 95% CI from 1.33 to 48.11), *bachelor_perc* = -38.18 (not statistically significant: p-value=0.06 and 95% CI from -78.59 to 2.23), and *income* = -0.002 (not statistically significant: p-value=0.11 and 95% from -0.0 to 0.00004). This concludes that there isn't strong causality between high vaccination percentage and average mobility score across California counties after taking into account all the confounders. This totally goes against our initial hypothesis that there is a positive causal relationship between high vaccination counties and their average mobility score.

Discussion

The OLS outcome regression method we used assumes that all variables interact linearly and consistently across both the treatment and control group. This could be a limitation if the variables don't have a naive linear relationship or if each confounder affects the treatment group differently from the control group. Another limitation with our method of assuming unconfoundedness is that this could never be the case since no one will have enough domain knowledge and also the data available to allow them to achieve perfect unconfoundedness. Additional data on even more confounding variables such as changes in covid policy, medical history, or even weather could all both affect the vaccination percentage of a county and their average mobility score. This would move us closer to the unconfoundedness assumption.

Overall, the model shows that there is no evidence of a causal relationship between the treatment and outcome. However, the fit of the model has a log-likelihood of -171.96 over 53 numbers of observations which does not seem to be the best. However, this is better than the naive model with a log-likelihood of -190.92 over 53 numbers of observations. This gave us some confidence that the confounding variables helped but we can never be certain on what the true ATE of the treatment variable is since there will always be more domain knowledge/confounders that could be incorporated.

Conclusion

Our findings from the first research question suggest that COVID-19 has drastically increased the percent change in mobility scores from baseline. We believe that this result is generalizable to other regions within California and other states as well because of policies regarding stay-home, masking, and vaccination requirements that emerged as a result of the rise of coronavirus. Our results further suggest that after the announcement of California's reopening plan, percent change in mobility scores from baseline has decreased, implying that there has been more activity within Alameda County in the retail and recreation sector. Based on these results, policymakers can plan accordingly to fund and support sectors that have had a hard time

transitioning back to normal after the reopening plan. Continuing to enforce vaccination requirements can further safely encourage mobility in public facilities, such as grocery markets and transit, as well as encourage the activity of entertainment businesses, such as restaurants and shopping centers. Although we did not merge any additional data sources for the first question, there still remain limitations in the Google Mobility Reports. The main limitation is that it does not include those who have not enabled location services. It would seem as if the vast majority of users would be included in the dataset as many critical functions of modern smartphones require users to opt into using location services (i.e. using Google Maps, find my phone, location sharing, etc.), but it is not explicitly mentioned in Google's Mobility Report how many users choose not to opt into using location services and so this is not accounted for in the data nor our analysis.

In terms of potential future studies, it would be incredibly interesting to compare this analysis across different counties/regions in California and even across state lines throughout the country. It is no doubt that COVID restrictions very quickly became politicized and it is critical that some semblance of ground truth be established. Although there is not a one-size-fits-all solution, objectively some states had better COVID-19 outcomes than others. Only by quantifying and comparing outcomes on a state-by-state basis can the best approaches to pandemic prevention be determined.

For the second research question, we found that there was no significant causal relationship between high vaccination and mobility score in a California county. Our results are confined to time and space—a few months after January 2022, and also the California region. Although Californians do not necessarily represent the rest of the world, by virtue of having a large and (somewhat) diverse population, there is a probable chance that there are similar results in places across the world similar to California. Since there is no evidence of a causal relationship between high vaccination and mobility score, if policymakers initially saw getting people vaccinated as the way to open the economy back up, they may need to look for other means of encouraging public mobility. In terms of data collection, we combined data to include confounding variables, which were age, race, bachelor degrees, and income. We needed that data to get an accurate estimate of the average treatment effect. However, we used data across different years, and so the data that we used does not exactly match consistently with reality, and may have affected our results. A limitation in our use of data is that we excluded data from Alpine, Sierra, Mariposa, Modoc, and Trinity, since there were NA values in some columns for these counties. This excluded 5 out of the 58 counties within California which is a good chunk of data that would have helped bring more consistency in our findings. In future studies, we could pull more data as more counties reach a higher vaccine percentage and use the larger data on mobility scores as well to get a more accurate estimate of the causal effect of high vaccinated countries on average mobility scores. Additionally, since we only focused on how the vaccines affected mobility scores in the COVID-19 pandemic, future works could certainly compare and contrast the causal effects of other vaccines for a different pandemic.

Finally, through this experience, we learned that a real-world project requires a lot more work than an in-class assignment that is already structured out for you. A lot of effort is spent in this project to find and clean the data even though we have the correct understanding of what needs to be done. Additionally, we have to constantly communicate with each other in order to be on the same page and understand the technical details within each section. We needed to have a good understanding of what we are trying to ultimately do because we are the ones that need to build towards that goal. Lastly, we learned that what we originally hypothesize is not always

correct. For in-class assignments, we generally have a good understanding of what our results should be (generally learning the methods to match that result). However, here we were surprisingly completely off from our original hypothesis that high vaccinated counties will cause higher mobility scores.

Additional Datasets

Covid:
https://data.chhs.ca.gov/dataset/e283ee5a-cf18-4f20-a92c-ee94a2866ccd/resource/130d7ba2-b6eb-438d-a412-741bde207e1c/

Population:
https://data.census.gov/table?q=population&g=040XX00US06%240500000&tid=DECENNIALPL2020.P1

Race:
https://data.census.gov/table?q=population&g=040XX00US06%240500000&tid=DECENNIALPL2020.P1

Education:
https://data.census.gov/table?q=bachelor+degrees&g=040XX00US06$0500000&tid=ACSST5Y2020.S1501

Income:
https://data.census.gov/table?q=income&g=040XX00US06,06$0500000&tid=ACSST5Y2020.S1901

Age:
https://fred.stlouisfed.org/release/tables?rid=430&eid=326943&od=#