W203 Lab 2 Report: Fun in the Sun - Summer '21

Allison Schlissel, March Saper, Matt Whittaker

4 August 2021

1 Introduction

The summer of 2021 is anticipated to be a partial return to normal life after the COVID-19 pandemic, with only certain elements of pre-pandemic life returning in full force. While government-encouraged lockdowns are still the norm in many parts of the world, the United States has seen a surge of demand among those ready to travel after about 1½ years of encouraged distancing. Even though returning to the workplace is still on hold for a sizable proportion of knowledge workers that have the luxury of working remotely, citizens of the United States appear willing to travel - especially domestically.

One anecdotal example of the demand for travel is the current rental car shortage. A recent influx of travelers to Hawaii led some tourists to resort to renting U-haul moving trucks for their vacations - not to move furniture in Hawaii, but instead to circumvent astronomical rental car prices. While Hawaii is more of a "destination" travel location, the demand for travel and the rental car shortage is not limited to only "destination" locations.

Airport screening data from the Transportation Security Administration (TSA) shows that as of June 30th, 2021, airport security officers at the TSA were screening about 2 million people daily. This is still down significantly from 2019 when about 2.5 million passengers passed through airport security each day. However, this is still up considerably compared to June 2020, when only about 500,000 passengers traveled through an airport. This increase in travel observed in 2021 is possibly due to the availability of vaccines, which is likely to increase the willingness to travel in 2021, since the previous summer (2020) saw a near-complete collapse of travel (80% decline in passengers) without the presence of a vaccine. Are people receiving the vaccine willing to travel again, and is this willingness actually resulting in an increase in travel?

Our research question explores a causal relationship between the vaccination rate (as of May 1st, 2021) and travel in June (defined as a "trip" - measured by cell phone location data - over 250 miles). We used the May 1st vaccination cutoff for a first dose to measure June travel data because of the CDC recommendation to wait two weeks after a second dose (if needed) before travel. In order to assess causality, we will control for certain variables in our models that also likely impact travel rates - availability of a car, proximity to an airport, education, and economic status. Our data is at the county level to measure discrepancies within states and determine differences in travel habits between county inhabitants within those states. One example of a discrepancy within states is San Francisco County, CA, where almost 70% are fully vaccinated (as of July 26th). In contrast, Del Norte County, CA, has only a 33% vaccination rate. By measuring travel at the county level, we will be able to identify important differences between counties with higher and lower levels of vaccination rates.

2 Data

To operationalize the data for our modeling process we turned to a variety of sources to gather travel, vaccination, and other covariate data. Information about travel by county came from the U.S. Department of Transportation Bureau of Transportation Statistics Trips by Distance[1] dataset which provides estimated mobility statistics on a county level. These statistics are produced from anonymized mobile device data and a multi-level weighting method that employs both device and trip-level weights, expanding the sample to the underlying population at the county and state levels before travel statistics are computed. Trips are defined as movements that are a longer than 10-minute stay from home (and the home location is imputed on a weekly basis). Trips are grouped by distance (for example, number of trips taken between3-5 miles). For our modeling, we used county-level data for the month of June 2021, and we examined number of trips over 100 miles and number of trips over 250 miles. It is important to note that this dataset is experimental and thus has not met all quality standards necessary for regular production. However, the dataset has been downloaded over 11 thousand times and we believe it is robust enough for our research purposes.

[1] https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

COVID-19 vaccination data comes from the Center for Disease Control and Prevention COVID-19 Vaccinations in the United States, County dataset[2]. This dataset documents overall US vaccine administration at a county level and represents all vaccine partners including retail pharmacies and jurisdictional partner clinics. Data is available by date, and we chose May 1st, 2021 as our cutoff of interest. The dataset contains several benchmarks to describe county vaccination, including total percent vaccinated with their first dose or a completed series. Aggregating data on such a massive and disjointed effort is extremely challenging and we note a few challenges to this dataset. First, not all states report county-level data. Specifically, California does not report the county of residence for a person receiving a vaccine when that county has fewer than 20,000 people. Texas does not report county of residence under any circumstances. Finally, for states that do proport to document all county data, there are no universal systems in place to report county of origin for people vaccinated outside of their home county. In addition, the CDC acknowledges systemic missing datasets due to variations in vaccine distribution policies that may cause some counties to report artificially low numbers. To preserve the integrity of our model, we removed the state of Texas from our dataset. We also removed a further 132 counties in California and other states which did not report vaccination data. The two most populous of these counties were Mesa, Colorado (population 154210) and Pennington, SD (population 113775). The remaining counties all had a population below 72,000 people. These challenges of missing data are inherent to many data modeling questions and given the nature of COVID-19 vaccination efforts, we are working with the best data available.

The remainder of our covariate data came from a variety of sources. To incorporate income per capita in our model we turned to the Bureau of Economic Analysis U.S. Department of Commerce 2019 CAINC1 Personal Income Summary: Personal Income, Population, Per Capita Personal Income[3]. This source presents per capita personal income computed using Census Bureau mid-year population estimates available as of March 2020. Dollar amounts are not adjusted to reflect ongoing inflation. Education achievement levels were sourced from the U.S. Department of Agriculture Economic Achievement Service. This dataset documents various educational achievements such as the number of people who received a high school or college degree by county in 4-year increments. For our research we used data for 2015-2019. Household car ownership levels came from the U.S. Census Bureau American Community Survey 5-Year Data (2009-2019)[4]. This report contains a wider variety of variables, including estimates of numbers of households by county who own 0, 1, 2, 3, or 4 cars. To estimate levels of diversity within a county we used data from the U.S. Census Bureau, Population Division County Population by Characteristics: 2010-2019[5] which estimates racial characteristics of the population by county during non-census years. Finally, to estimate the level of airport activity in a county we turned to the 2017 Federal Aviation Administration Voluntary Airport Low Emissions Program report[6]. We chose this dataset because it listed airports and enplanements by county. Using this we can get a comparison of airport business levels across all counties during normal, non-pandemic times. We removed all regional-only airports from the dataset since we are interested in travel, and these airports provided no commercial passenger service. Since we are primarily interested in relative airport business, the data from 2017 was adequate.

3 Exploratory Data Analysis

3.1 Vaccines

In order to analyze June travel data, we chose to use a cut-off date of May 1st for the vaccine data in order to ensure that there was a month of buffer time for the recently vaccinated to be comfortable with traveling. Additionally, this reflects CDC guidance that the vaccine is not immediately effective. Interestingly, the vaccine % is mostly normally distributed - however there is a spike in the center of the chart that indicates that the impact of any outliers is minimal. As a result, this will be a satisfactory variable to include in our model and will not require transformation.

3.2 Number of trips

To examine travel in June 2021 we chose two trip distances to explore in our EDA: total trips over 100 miles per person in the county, and total trips over 250 miles per person in the county. The average county had 1.4 trips over 100 miels and only 0.22 trips over 100 miles. The spread of the 100 mile trips was quite normally distributed, while the distribution of trips over 250 miles dropped precipitously.

[2]https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh
[3]https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&acrdn=6
[4]https://www.census.gov/data/developers/data-sets/acs-5year.html
[5]https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html
[6]https://www.faa.gov/airports/environmental/vale/media/vale_eligible_airports.xlsx

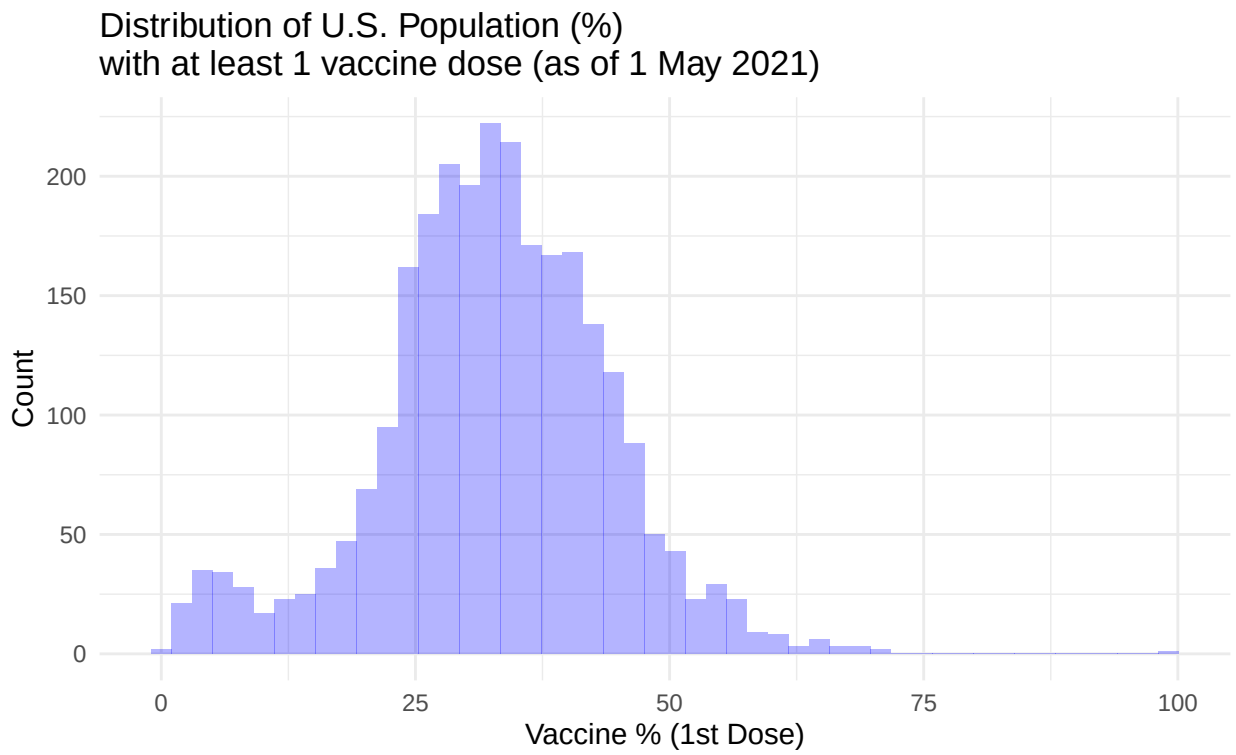


Figure 1: County population % which had received at least 1 dose of the COVID-19 vaccine as of May 1 2021.

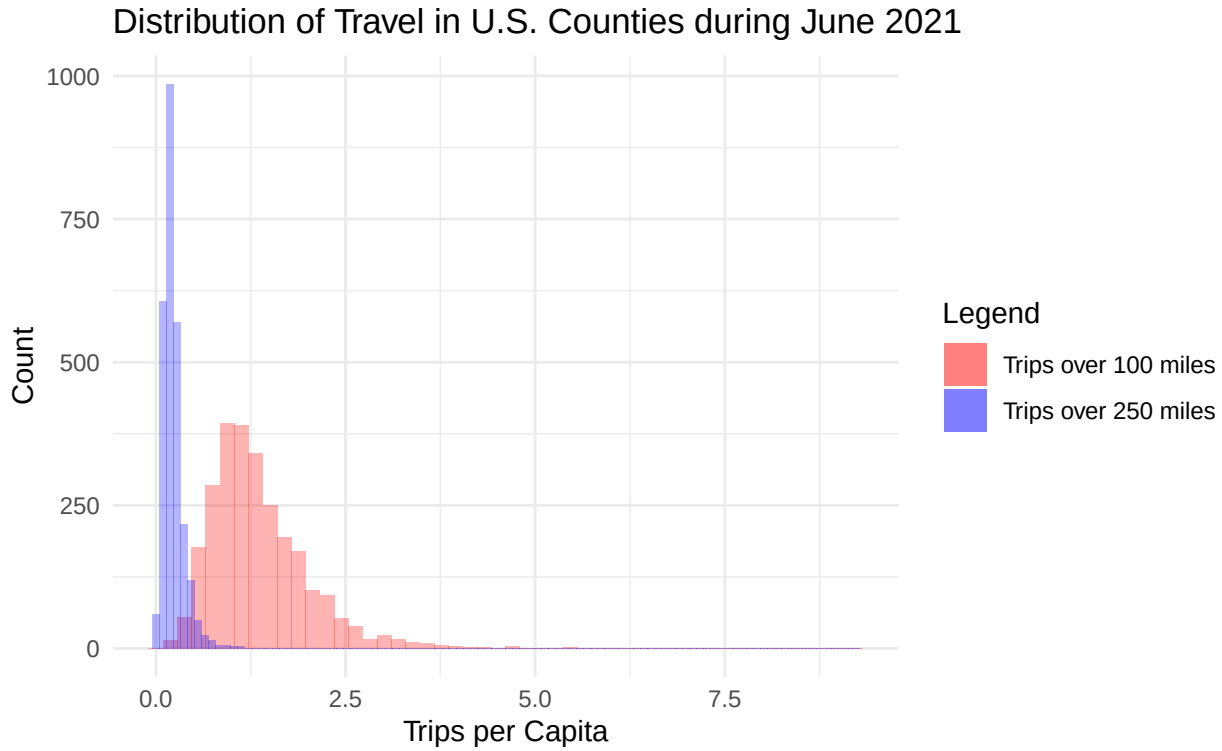


Figure 2: Distribution of trips per capita for trips over 100 miles and trips over 250 miles, by county.

3.3 Deep Dive: Vaccinations and Trips

After excluding Texas, Hawaii, and Alaska (as outlined in the data section of the report) - the vaccination rate by county shows that there are pockets of higher vaccination rates along the east and west coast, along with some northern states (e.g., Minnesota, Michigan) with high vaccination rates. Low vaccination rates are seen in less populated or rural areas in the country.

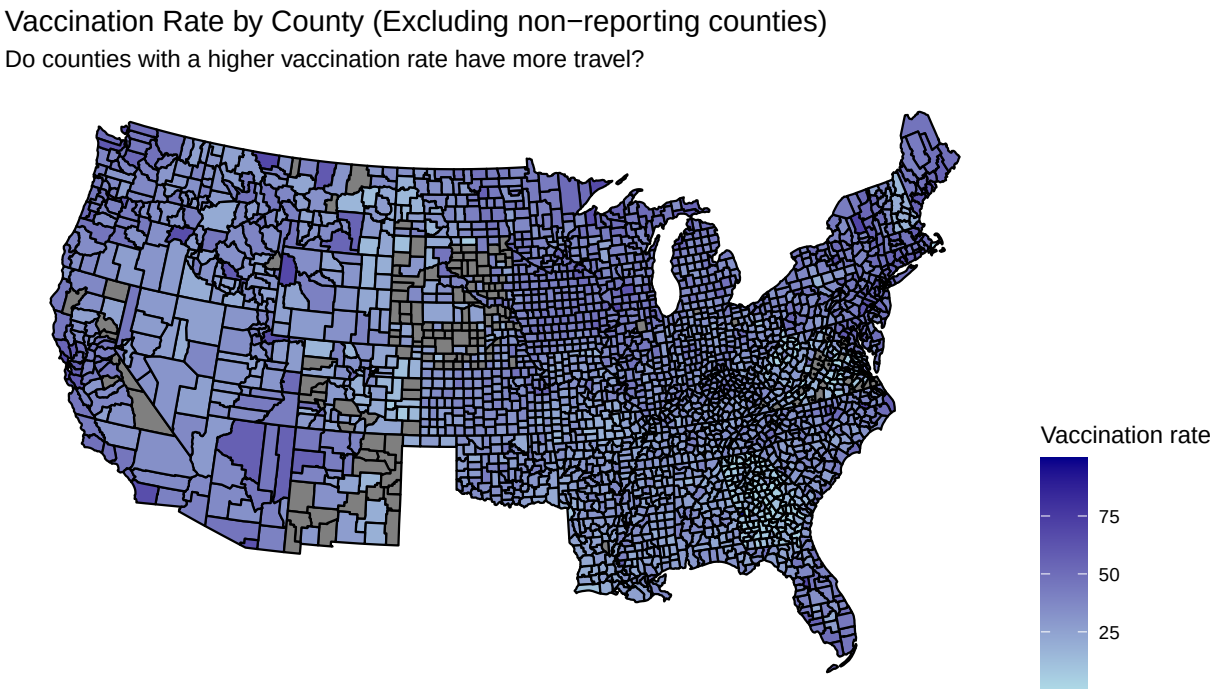


Figure 3: U.S. County-level vaccination rate. Texas excluded due to lack of county-level data.

California is a good example of the "coastal" effect where there are higher vaccination rates closer to the ocean, where there are larger cities and higher population density as compared to the inland areas. In the county view below, note that the gray counties do not report their vaccination data to the CDC because they fall below a 20,000 population threshold.

Using the same map visualization to show the number of trips over 250 miles per capita shows a slightly different effect, with the less populated or rural areas having a higher per capita measurement of trips. This may be due to having to travel much longer distances for basic necessities like food - and not travel for "leisure" which is why we chose a distance over 250 miles to define a vacation.

One anecdotal example of this impact is a higher rate of trips per capita is on Native American reservations, where inhabitants in those counties may have to travel extremely long distances for basic necessities like an affordable grocery store. In the following in-depth view into Arizona, the counties with the highest per capita measurements have significant Native American populations - a highly unique population demographic characteristic. The Tohono O'Odham reside in Pinal and Pima counties (highest per capita measurements), while the Najo, Hopi, and Hualapai Native Americans all primarily reside in other high per capita travel counties. Phoenix, the most populated location located in Maricopa County, is around the median for per capita trips.

```
## [1] ""
```

Another driver of per capita trip distance is the population of the county. Rural counties seem to require more trips over 250 - it is likely that these trips are not for "leisure" but are done out of necessity. Of the top 150 counties in the USA measured by per capita trips over 250 miles, there is not one county with a population over 1,000,000. Even though there are around 50 counties with a population over 1,000,000 in the US, the per capita travel in these counties appears lower compared to more rural areas

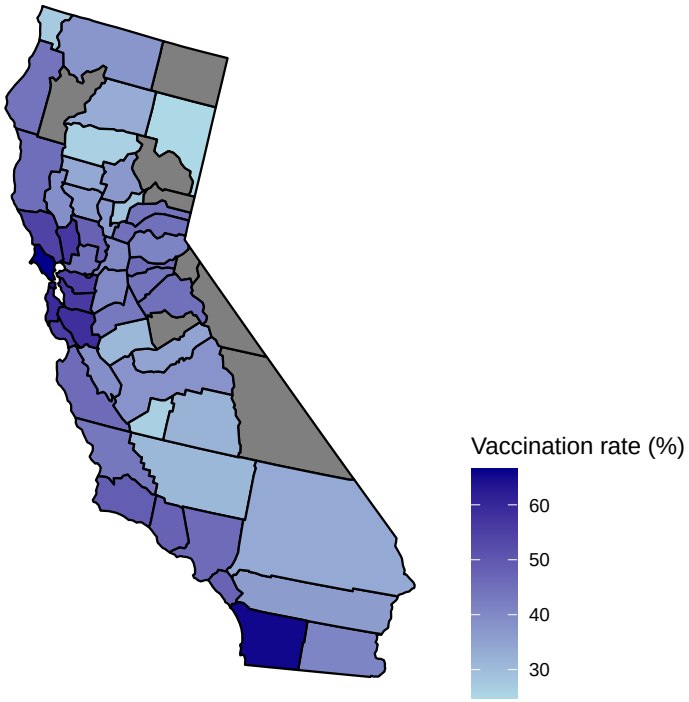


Figure 4: Coastal effect visible in California county-level vaccination rates.

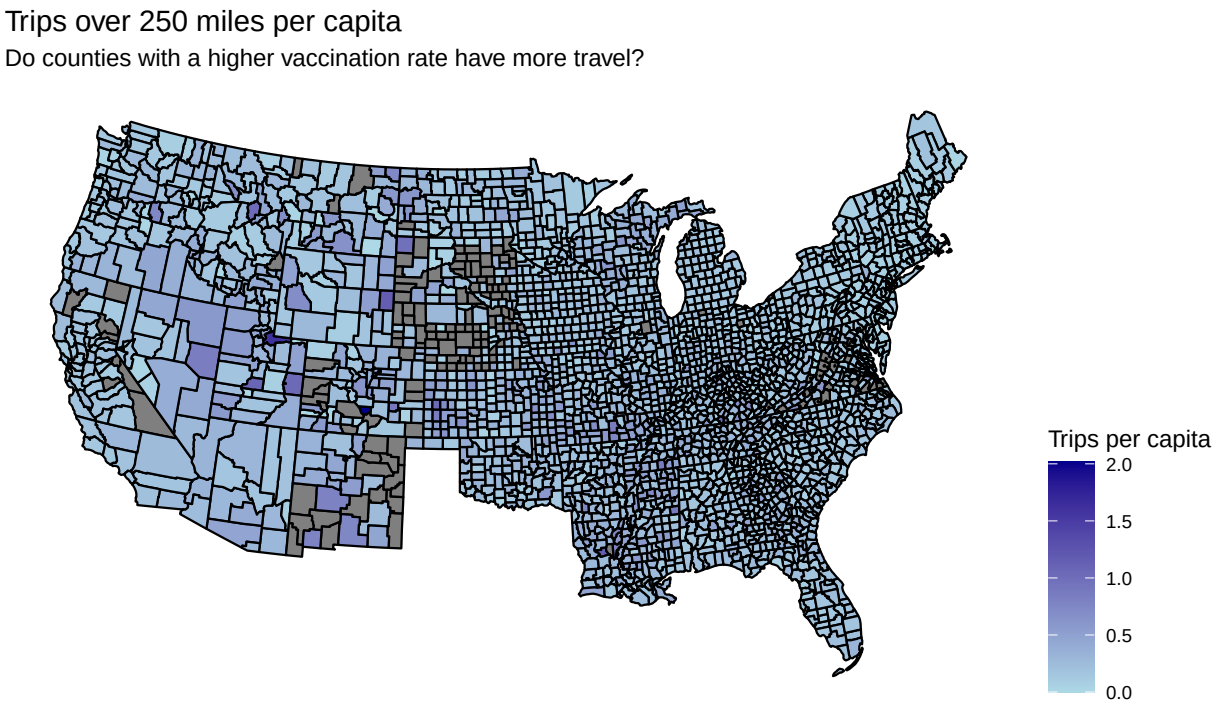


Figure 5: Trips over 250 miles per capita show greater levels of tavel in rural areas.

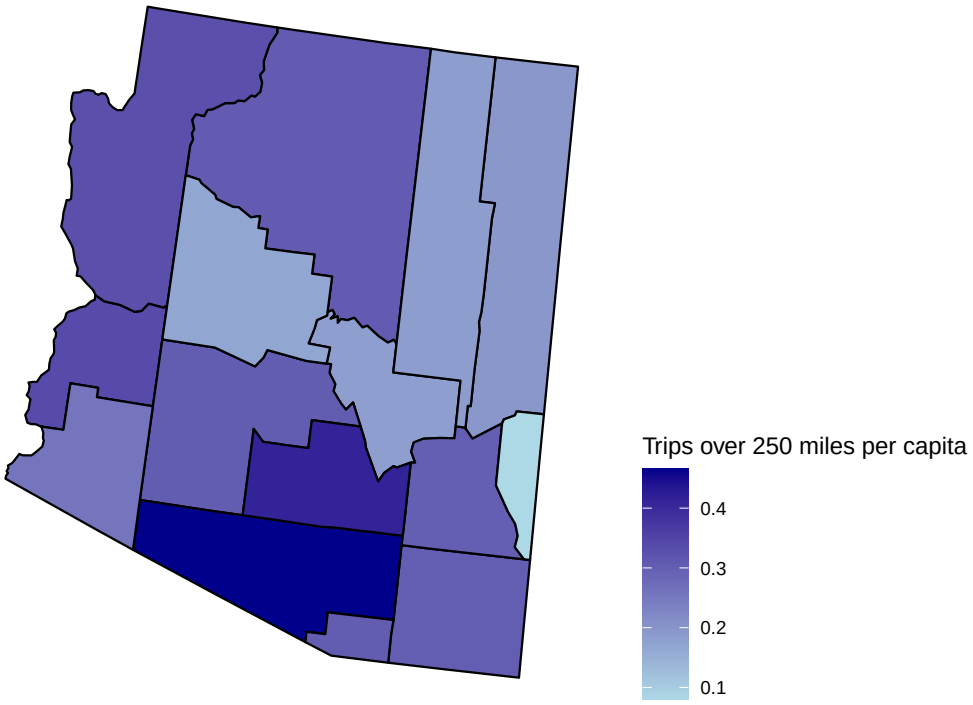


Figure 6: Trips over 250 miles per capita in Arizona highlight rural disparities.

3.4 Car Access

A variable that may impact the number of trips taken over 250 miles is access to a car. For this variable, we used the percentage of households (by county) with access to a car. Without car access, it would be difficult to travel over 250 miles, so this is an important control measure to include as we would like to understand the causal relationship between vaccines the and travel over 250 miles. As shown below, the access to a car is also mostly normally distributed between 85% and 100%, with a slight skew towards higher vehicle access. Consequently, this variable will be useful for the causal model to control for vehicle access, as we would like to remove that predictive coefficient from the vaccine dose.

3.5 Airports

In our filtered dataset, 278 counties had an airport which provided non-regional passenger service. The top 5 busiest airports were located in Fulton, GA, Cook, IL, Queens, NY, Los Angeles, CA, and Denver, CO . These airports are responsible for 31% of air travel in the U.S. The average trips over 100 miles per capita in these counties was 0.75 compared to a national 1.37. The average trips over 250 miles per capita was 0.3 compared to a national 0.22 indicating that for these high-travel cities, the population didn't necessarily travel more. A histogram distribution of airport enplanements per county highlights just how many counties do not have a airport and just how busy the top 5 airports in the U.S. are comparatively.

3.6 Income

The income variable was chosen because various studies and media sources have demonstrated somewhat of an association between vaccine compliance and education level. While income is not normally distributed in the United States, rather, it follows more of a Cauchy distribution, we find that the data reflect this distribution and it is a reasonable variable to include in our analysis.

3.7 Education

The education variable was chosen because various studies and media sources have demonstrated an association between vaccine compliance and education level. Interestingly, education at the county level appears to be normally distributed, and we think this is a reasonable variable to use in our report.

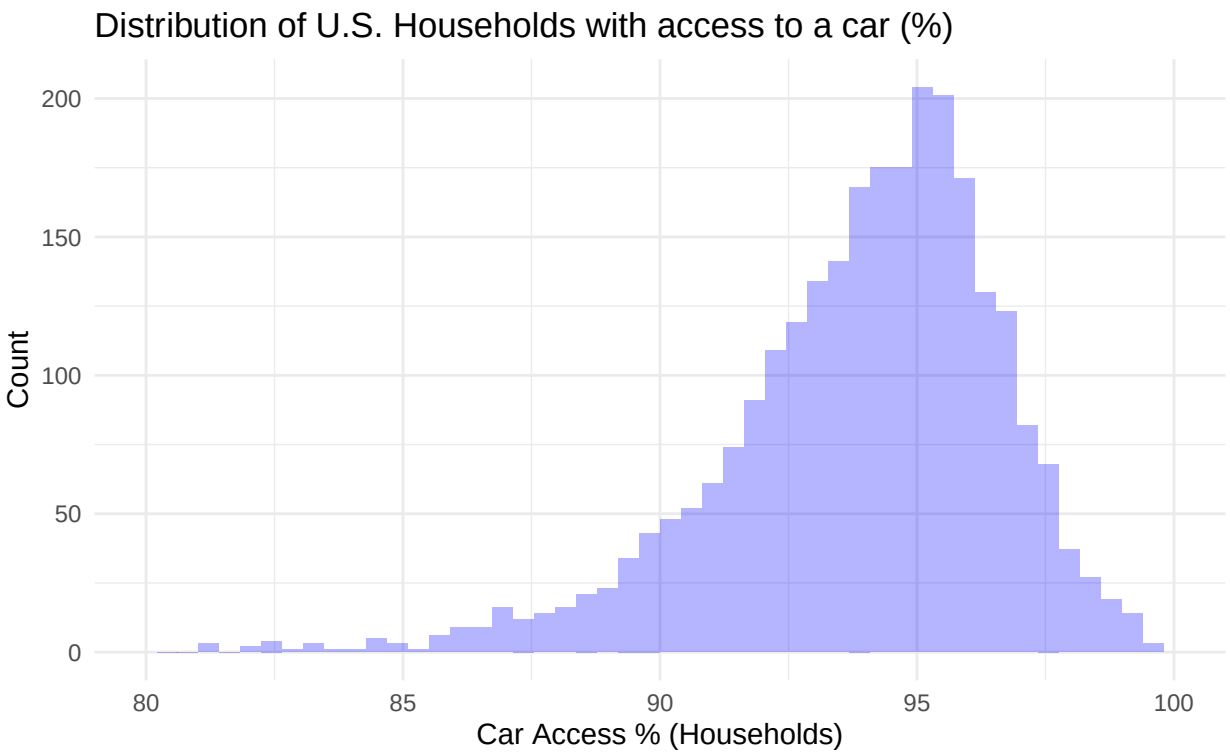


Figure 7: Percent of households by county with access to at least 1 car.

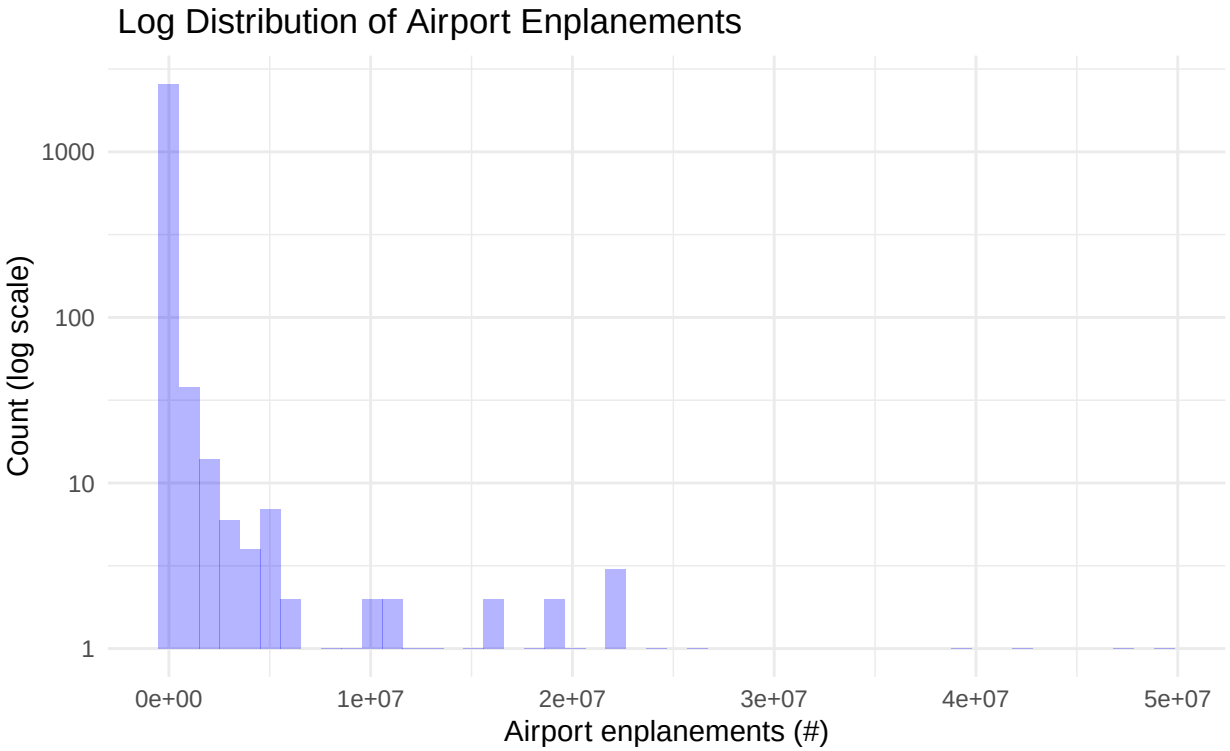


Figure 8: Log scale distribution of airport enlanements by county in 2017, used to approximate airport busyness level.

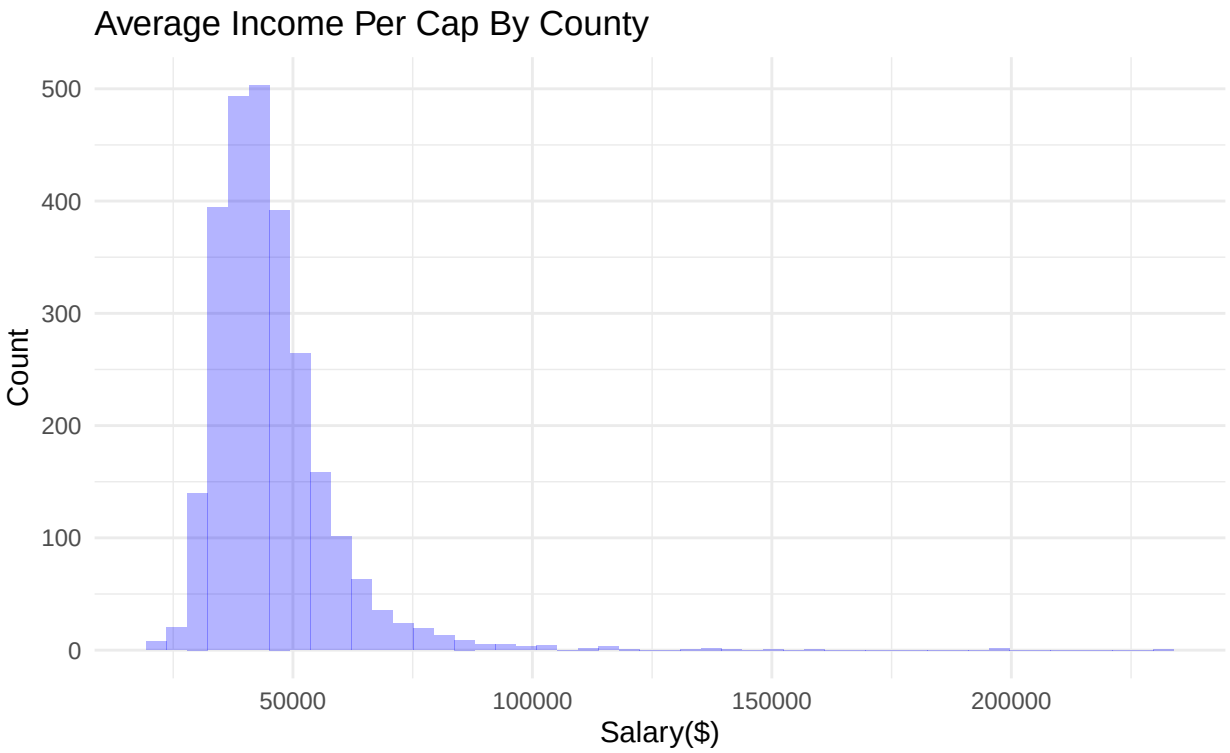


Figure 9: Salary distribution per capita by county follows a Cauchy distribution.

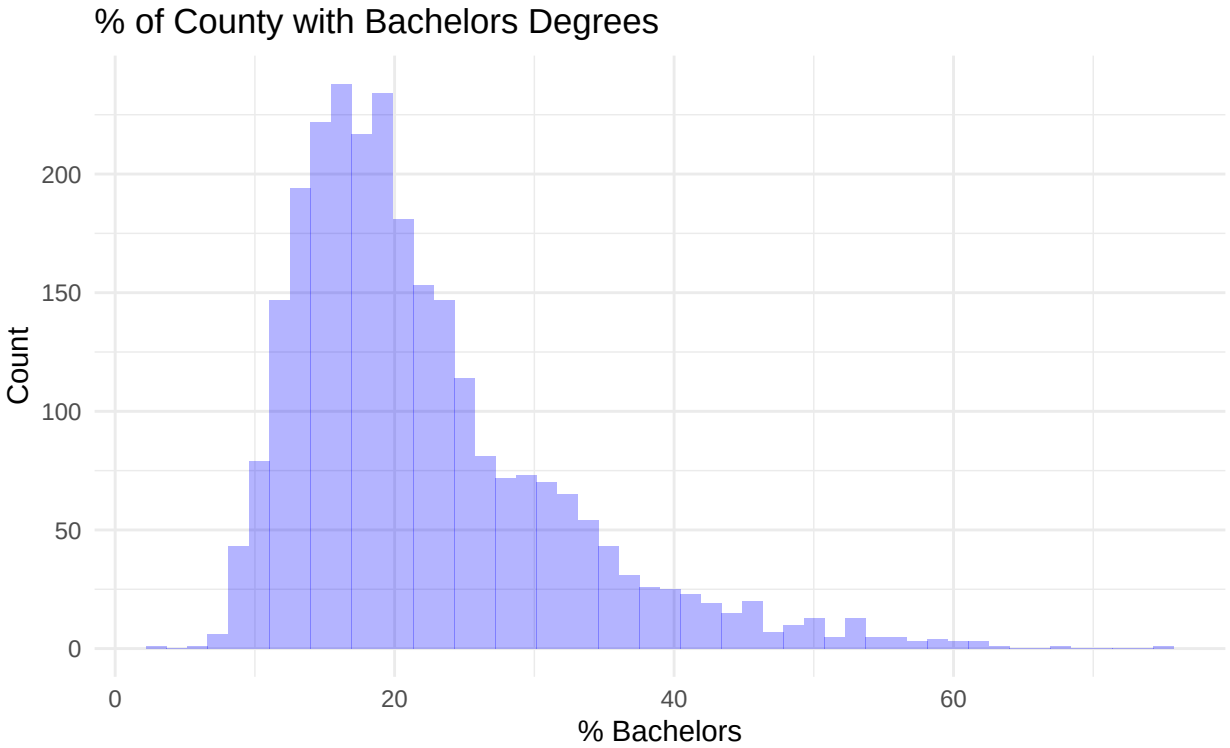


Figure 10: Bachelor's degree achievement levels by county distribution.

4 Model

After performing EDA in preparation for the modeling goal of building a causal regression model for the impact of Vaccination Rate (1st Dose) on Travel (Trips over 250 miles) - we ultimately selected the following variables (all at the county level) to include in the models. The variable selection process was iterative with many different combinations tested, with discussion on the qualitative relationships prior to the quantitative analysis. Note that further details on the data selection and associated methodology to prepare for model building can be found in the "Data" section.

Variables and Covariates

- `TripsOver250PerCap`: Number of trips per capita over 250 miles
- `Dose1Pct`: First coronavirus vaccine dose (%)
- `CarPct`: Households with car access (%)
- `log(AirportEnplanements+1)`: Log transform of airport enplanements (#, used as a proxy for airport access)
- `IncomePerCap`: Income per capita
- `BachelorsPct`: Bachelors degree (% that have obtained)
- `WhitePct`, `BlackPct`, `HispanicPct`, `IndigenousPct`, `AsianPct`: Race and demographic control variables (% White, % Black, % Hispanic, % Indigenous, % Asian)

The plot below shows a scatter plot comparison between each variable (excluding Race control variables) to help visualize the correlation relationships between the variables. The strongest associations were expected (e.g., bachelors degree vs. income per capita) but we expected to see a stronger relationship between Dose 1 percentage and the trips over 250 miles, which are visualized on the top-left panels.

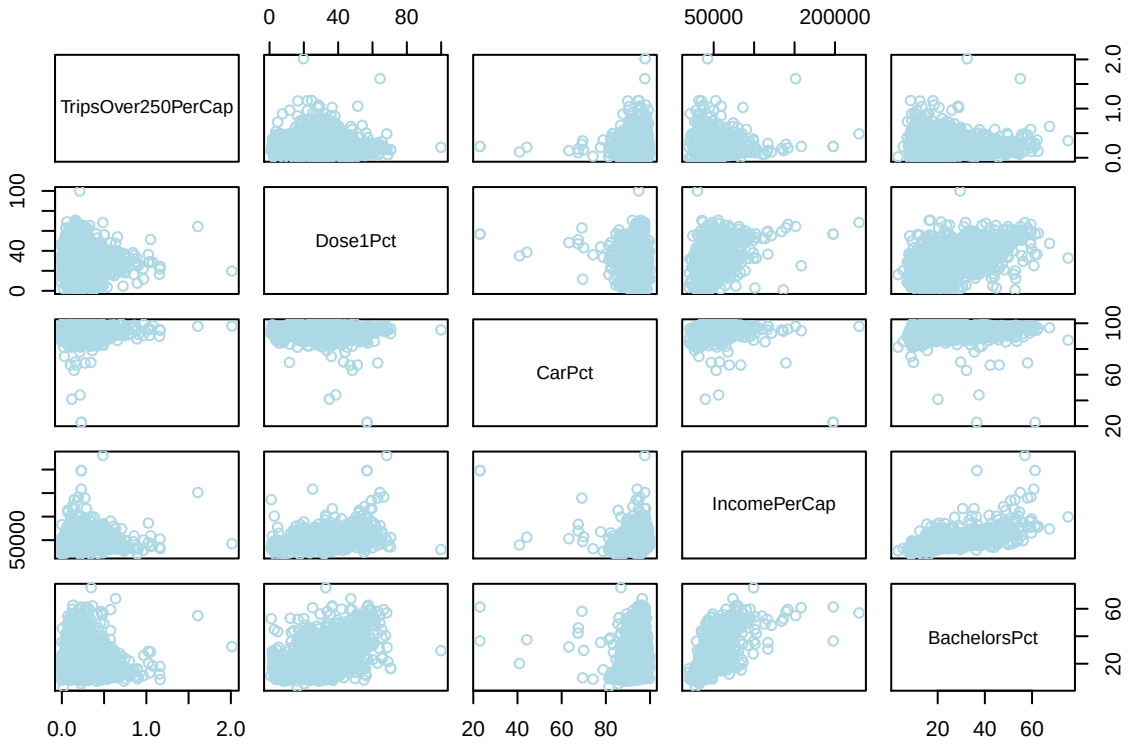


Figure 11: Variable comparison among key covariates

In order to select the best combination of these variables to include in our model, we used an F-test to compare the model outputs and a t-test to identify a subset of the predictors that we believe to be the best at explaining the response. While the results are detailed in the following section, the only significant variables other than the Dose 1 percentage in model 3 (model with all variables included) were the percentage of households with cars, access to airports, and % Hispanic. For our model with the most

important covariants, we kept the households with cars and access to airport variables, choosing to not control for race as only one race was significant.

To separate impact and analyze a wide range of variables, we used 3 different models with separate subsets of variables to conduct the analysis (as shown in the following regression analysis section)

4.1 Model Regression Analysis

The model with the highest R^2 was our Model 3, which was higher than the first two models, but still fairly low. The R^2 for this model is 0.044, and the Adjusted R^2 is 0.040. With observing a low R^2, we would hesitate to draw any conclusion from the impact of vaccine doses on travel over 250 miles, even with controlling for different factors.

In the table below, we include statistical outputs for all three models. Model 1 contains only the primary covariate; Model 2 contains a subset of covariates, and Model 3 contains all covariates in this dataset.

```
##
## Results
## ================================================================================================
##                                                Dependent variable:
##                              ---------------------------------------------------------------
##                                                TripsOver250PerCap
##                                    (1)                  (2)                    (3)
## ------------------------------------------------------------------------------------------------
## Dose1Pct                      -0.002*** (0.0002)   -0.002*** (0.0003)   -0.002*** (0.0003)
## CarPct                                              0.003*** (0.001)     0.004*** (0.001)
## IncomePerCap                                       0.00000** (0.00000)  0.00000*** (0.00000)
## log(AirportEnplanements + 1)                                              0.001 (0.001)
## BachelorsPct                                                             -0.001 (0.0005)
## WhitePct                                                                  0.002 (0.003)
## BlackPct                                                                  0.003 (0.003)
## IndigenousPct                                                             0.003 (0.003)
## HispanicPct                                                              0.002*** (0.0003)
## AsianPct                                                                 -0.001 (0.003)
## Constant                       0.288*** (0.009)      -0.012 (0.069)       -0.364 (0.277)
## ------------------------------------------------------------------------------------------------
## Observations                         2,668                2,668                  2,668
## R2                                   0.019                0.027                  0.044
## Adjusted R2                          0.019                0.026                  0.040
## Residual Std. Error            0.145 (df = 2666)    0.144 (df = 2664)      0.143 (df = 2657)
## F Statistic                52.033*** (df = 1; 2666) 24.715*** (df = 3; 2664) 12.122*** (df = 10; 2657)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

The 4 coefficients that are statistically significant are: Dose1Pct, CarPct, IncomePerCap, and HispanicPct. One interesting note is that although HispanicPct is statistically significant, none of the other ethnic groups in this analysis are statistically significant. Given this context, we are hesitant to trust HispanicPct as practically significant. Further we note that IncomePerCap is less than 0.00001 and of limited practical significance. It is likely marked as statistically significant due to the large sample size. Overall, all coefficients were small, indicating that the relationship between them and travel is likely weak.

5 CLM Assumptions

While the size of our dataset (over 2,600 observations) is large that we need not use the Classical Linear model approach, we have completed an analysis of the CLM assumptions on model #2 as an exercise. The most interesting/relevant assumptions are highlighted below.

5.1 Homoskedastic Errors

We first evaluate the residual values of our model to determine whether the distribution of the errors is homoskedastic. A visual inspection should show a constant band of even thickness from left to right around the fitted line. For our model, this is not the case. We an area of increasing/decreasing errors which indicate that further work should be done before truly trusting the model.

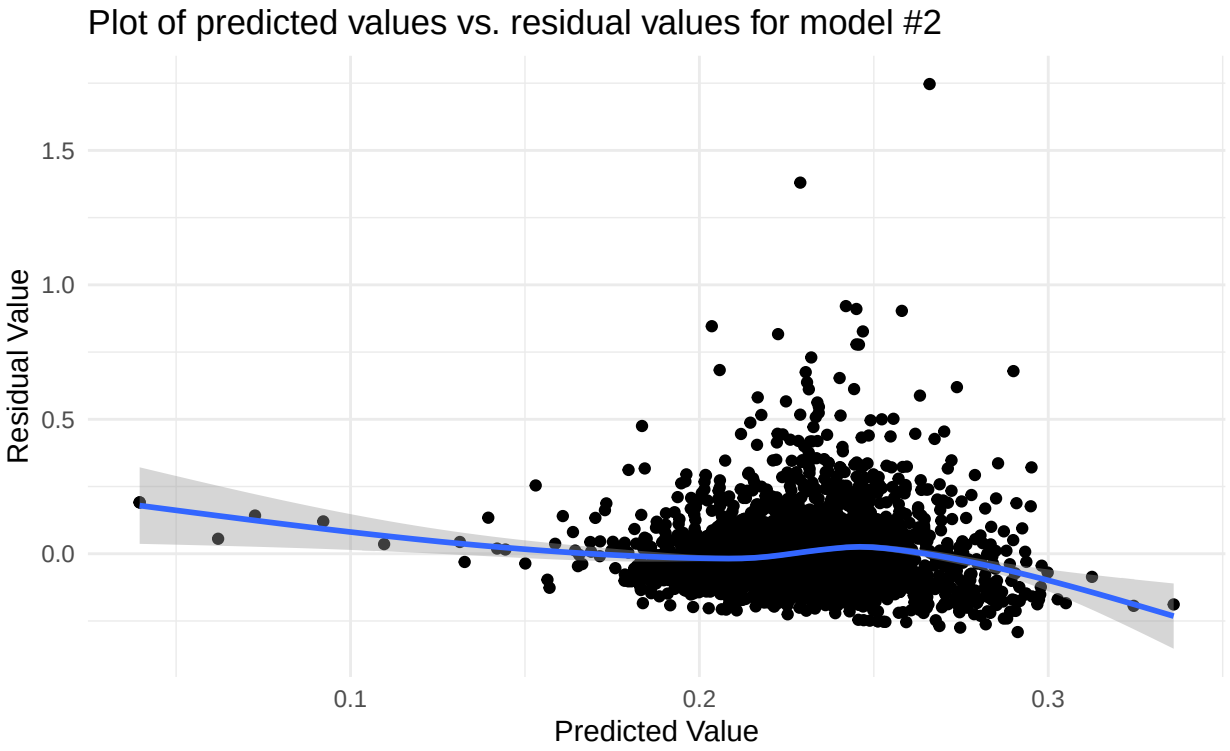


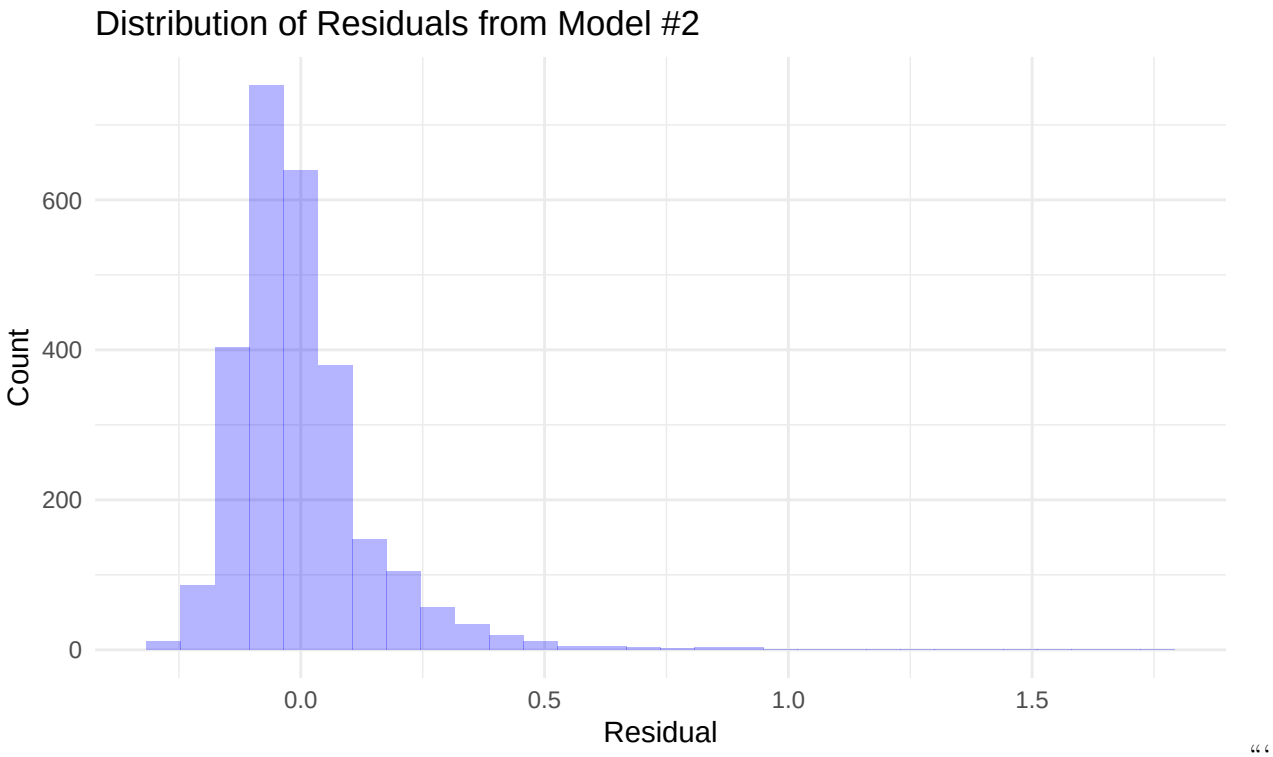
Figure 12: Scale-Location plot to evaluate model for linear conditional expectation andhomoskedastic errors.

5.2 Linear Conditional Expectation

It is difficult to asses whether our model has a linear conditional expectation based on the plot of predicted vs residual values for model 2. The trend line certainly appears quite linear, however the tight distribution of data points. This might suggest the need for a further transformation of one of our variables to create more opportunity for our model to create linear predictions.

5.3 Normally Distributed Errors

To evaluate whether model 2 has normally distributed errors we examine the histogram of residuals for noticable deviation from normality. Fortunately, the residual distribution is quite normal, with only a slight heavy tail on the right hand side. This indicates that our model is likely unbiased and we do not see any results that concern us. The histogram of residuals and the qqplot shows some some deviation from normality, specifically a right skew and perhaps an unusual concentration on the right tail.



“‘

6 Omitted Variable Bias

To establish a causal relationship between vaccination rate and travel, we analyzed two potentially important omitted variables (i.e., variables not included in the model that conceptually impact both vaccination rate and travel) to help rationalize any bias before making a causal conclusion:

1. Trips by county prior to the pandemic - We ultimately chose to not include this variable in the model because of the vast differences in travel habits in 2019 vs. 2021. Certain age groups for example may be more likely to travel in summer of 2021 (e.g., college students with less deadly virus risk) than in 2019 when travel would have been more “normal”. However, this omitted variable would still impact the number of trips taken (given potential historical baseline effects) as well as the vaccination rate, because we assume people who want to travel in 2021 are conceptually more likely to want the vaccine or are even mandated to get the vaccine. Since the omitted variable will likely be positively correlated with trips (and the positive coefficient modeled) as well as vaccination rate (another positive coefficient), the omitted variable will be a positive coefficient. As a result, the direction of bias is away from zero.

2. Population density - Population density may impact both number of trips taken over 250 miles and the vaccination rate. For example, in New York County (i.e., Manhattan) less than 50% of the households had access to a car. This could decrease the number of trips that are taken over 250+ miles because of access difficulty (even if mobility within the city is high due to public transportation / subways). Population density could also increase the vaccination rate because the transmission risk is elevated in densely populated areas. Since the omitted variable will likely be negatively correlated with trips (and the positive coefficient modeled) and positively correlated with vaccination rate (a positive coefficient), the omitted variable will be a negative coefficient. As a result, the direction of bias is towards zero. For these omitted variables, the direction of bias of the two variables will offset the overall impact on the model. This improves the ability to make a causal claim between the vaccination rate and number of trips taken in our analysis.

Potential limitations of our model include the omitted variables listed above, as well as non-modeled behavioral data where securing underlying data would be challenging. For example, knowing anti-vaccination behavioral sentiment would have helped control our model for those counties where there is significant peer pressure within the county to distrust vaccinations. Additionally, further analysis could be performed on available data to control for additional non-modeled variables like political parties or counties who voted for Donald Trump, who has been a vocal public health skeptic (although he received the vaccine in January).

7 Conclusion

The team looked to research any causal relationship between the vaccination rate (as of May 1st, 2021) and the number of trips over 250 miles in June. We are unable to draw a meaningful causal relationship from our research, due to a low R^2 across all of our models. When determining the impact on travel over 250 miles, a different set of variables should be explored to understand any causal relationship. As a result, further analysis would need to be performed on other potential explanatory and control variables in order to draw a causal conclusion. We recommend starting with identified limitations of our model, which did not include political party affiliation or travel habits prior to the pandemic. The summer of 2021 is likely to see a spike in travel, however travel habits do not seem to be any different between counties with high vaccination rates and low vaccination rates.