Causal Relationship Between Vaccinations & Mobility

David Miller, Aysun Ozler, John Stilb

8/3/2021

Introduction

The pandemic has driven people back into their homes on a level never seen in modern history. However, vaccine distribution reached scale right around the one year anniversary of initial lockdowns, and the national morale and confidence about the situation seemingly improved by the day. Our research question asks how mobile are the people in each state as of the end of May 2021, and how much of that is driven by the vaccination rate in that state.

Our hypothesis is that mobility can be measured in a simple but elegant way. Using data from the COVID-19 Mobility Report (collected by Google), we hope to define mobility as trips per person. Review of Google's documentation and our knowledge of Google's tracking capabilities in general suggest that the data was collected in a consistent and unbiased manner. Our primary explanatory variable, vaccination rate, is also thoroughly tracked at the state and federal government level. We are reasonably confident that state's are consistent in reporting their vaccination totals.

We acknowledge that there are potentially several other factors that drive mobility in the COVID era. As the pandemic became more political, a state's age and party leanings seemed to have an effect on the general attitude towards the gravity of the pandemic. How hard a state was hit through the duration of the pandemic (defined as the number of cases) may still be playing a role in people's mindset and mobility even if cases are currently low and vaccination rate is high. We will assess the impact of other potential explanatory variables by conducting a series of regression models, rather than just one.

Model Building Process

Measurement Goal

We aim to measure the amount of mobility that people in the United States have during the COVID pandemic. Specifically, our variable of interest is the total number of trips per person in May 2021, and we are investigating how that measure differs across states. We have considered several definitions of what constitutes a trip, as the data we used stratified trips by distance (in miles). We eventually decided to consider all trips of any distance, to eliminate any unknown bias for size of the state, as this could potentially impact the typical length of trip a person takes.

We also collected data from several sources to assess the correlation between input variables and our outcome variable of interest. As will be described in more detail in the next section, this included state (or county) level data on COVID infections and deaths, vaccinations, resident age, and voting history.

Modeling Goal and Causal Theory

The goal of this modeling exercise is to determine the causal relationship of a state's vaccination rate and the mobility of its residents. Our hypothesis is that vaccination rate is the primary determinant of how mobile people are recently, given the intense national focus on the current state of the pandemic. However, we appreciate that there are other factors that could prove to have a sizable impact in describing states' mobility.

In the same way that political preference, age, and cumulative cases may have an impact on mobility; we also expect they may have a confounding impact on the model because of correlation with vaccination rate. We expect median age almost certainly will, given the consistent emphasis on the fact that older people tend to be more adversely impacted by the virus.

We will produce a series of models that include progressively more variables in an attempt to isolate causality, significance, and correlation.

Exploratory Analysis Findings

We set out to build a model that used some variable containing vaccination rate information to explain mobility on a state-by-state basis. Additionally, we were interested in variables related to infection prevalence, political leanings, and age. Our primary focus throughout our exploratory data analysis was to determine which variables we would use to represent this information, their structure, and any transformations we would need to apply to better explain the relationships.

Data Collection

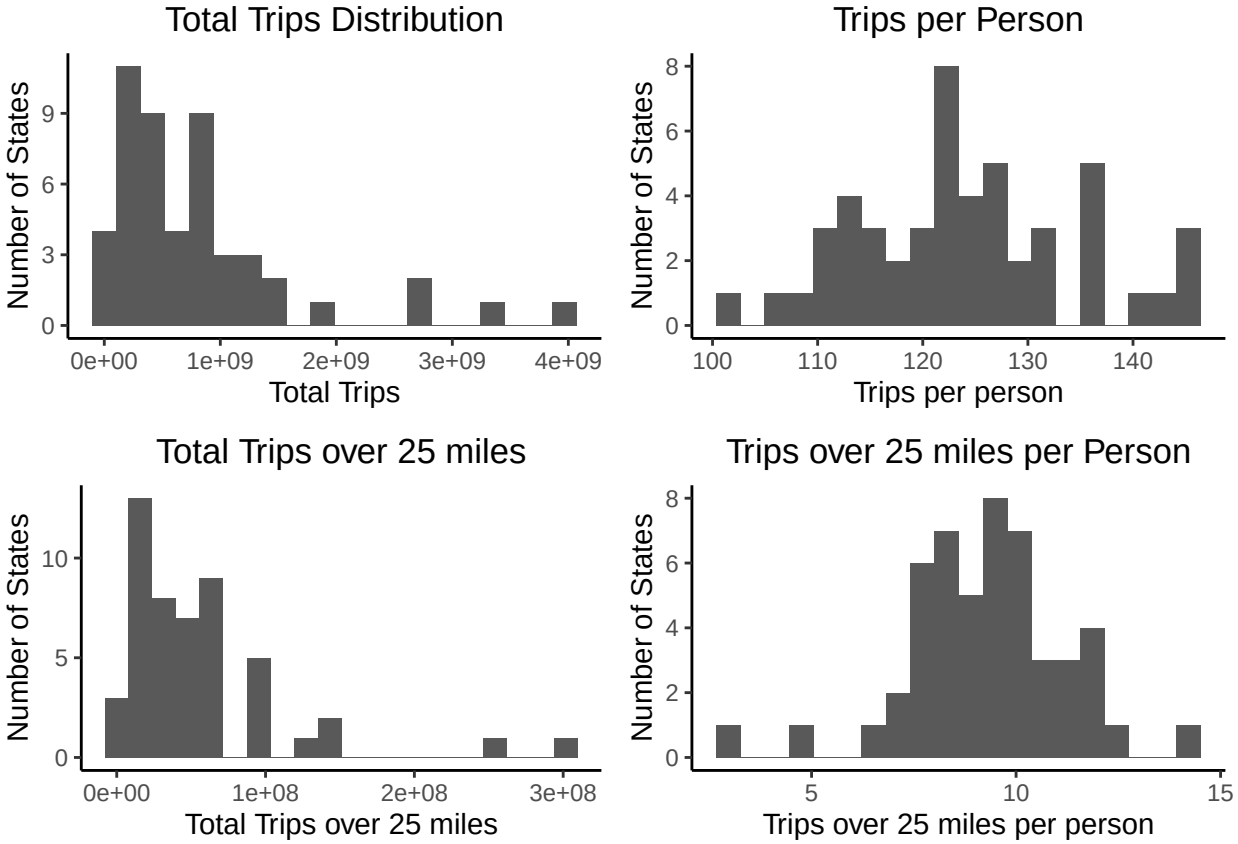
We compiled our data from a series of datasets (listed below) aggregating each variable on our unit of observation: state. We used the most recent available data up until, but not including, June 1, 2021.

Datasets

- **The American Community Survey**: A dataset from the US Census Bureau containing state-level demographics and other indicators of general interest from the 2019 census, which we used to gather the median age for each state.
- **CDC Data on Vaccinations**: A dataset from the US Centers for Disease Control containing county-level information about vaccination rates, which we used to calculate vaccination rates for each state.
- **Bureau of Transportation Statistics**: A dataset with the number of trips of a given distance taken by residents of a specified geography, which we used to gather the avg. number of trips per person by state.
- **New York Times Covid-19 Data**: A database, compiled from several authoritative sources on the occurrence of COVID-19 at the county level, which we used to calculate infection rates by state
- **2020 General Election Results Data**: A dataset with the number of votes towards a given presidential candidate broken out by state and election year, which we used to calculate the percentage of votes for the Republican Party in the 2020 Presidential Election.

Variable Distributions & Operationalization

Avg. Trips Per Person (Outcome Variable) We examined 4 potential variables for mobility: total trips, avg. trips per person, total trips over 25 miles, and avg. total trips over 25 miles per person. Total trips and total trips over 25 miles both had skewed distributions while avg. trips per person and avg. trips over 25 miles per person followed approximately normal distributions. We chose to include all trips and not just those over 25 miles, because each state has a unique geography and infrastructure. For example, a population-dense, urban city like New York City may have many more trips that are less than 25 miles causing the state of New York's results to look dramatically different from a state such as Wyoming that is largely spread out.



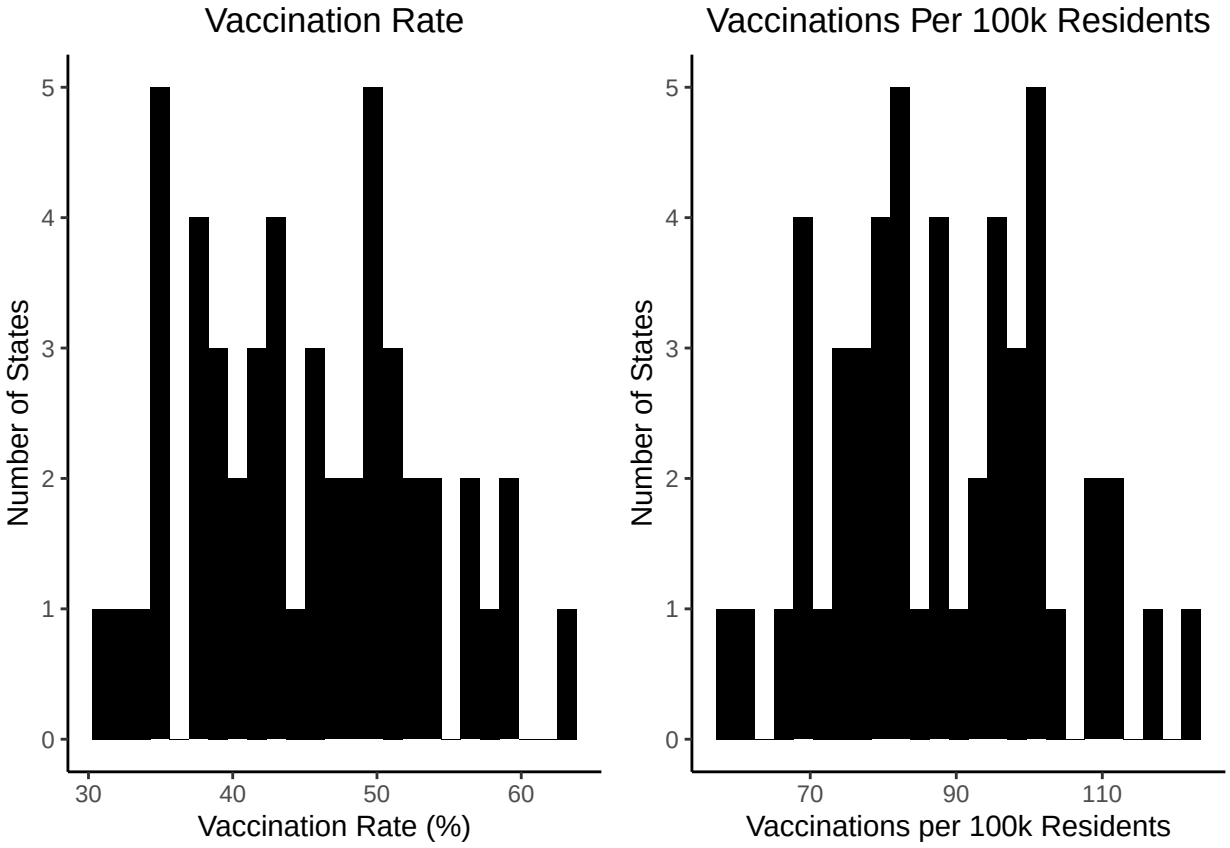
Vaccination Rates (Primary Causal Variable of Interest) We examined two potential variables for vaccinations: total vaccinations per 100k residents and vaccination rate. Total vaccinations treated all vaccinations (those fully vaccinated and those partially) as equal and considered how prevalent vaccinations were for every 100k residents in a state. Because this distribution appeared to be bimodal and it treated all vaccines as equal, we opted to use vaccination rate.

Vaccination rate was calculated as follows:

$$\text{Vaccination Rate} = 100 * \frac{\text{People Fully Vaccinated} + (\text{People Vaccinated - People Fully Vaccinated } * 0.5)}{\text{People}}$$

We weighted those who weren't fully vaccinated as half of vaccination; because while these people still required a shot, evidence suggests they are more immune to the virus compared to those without the booster.

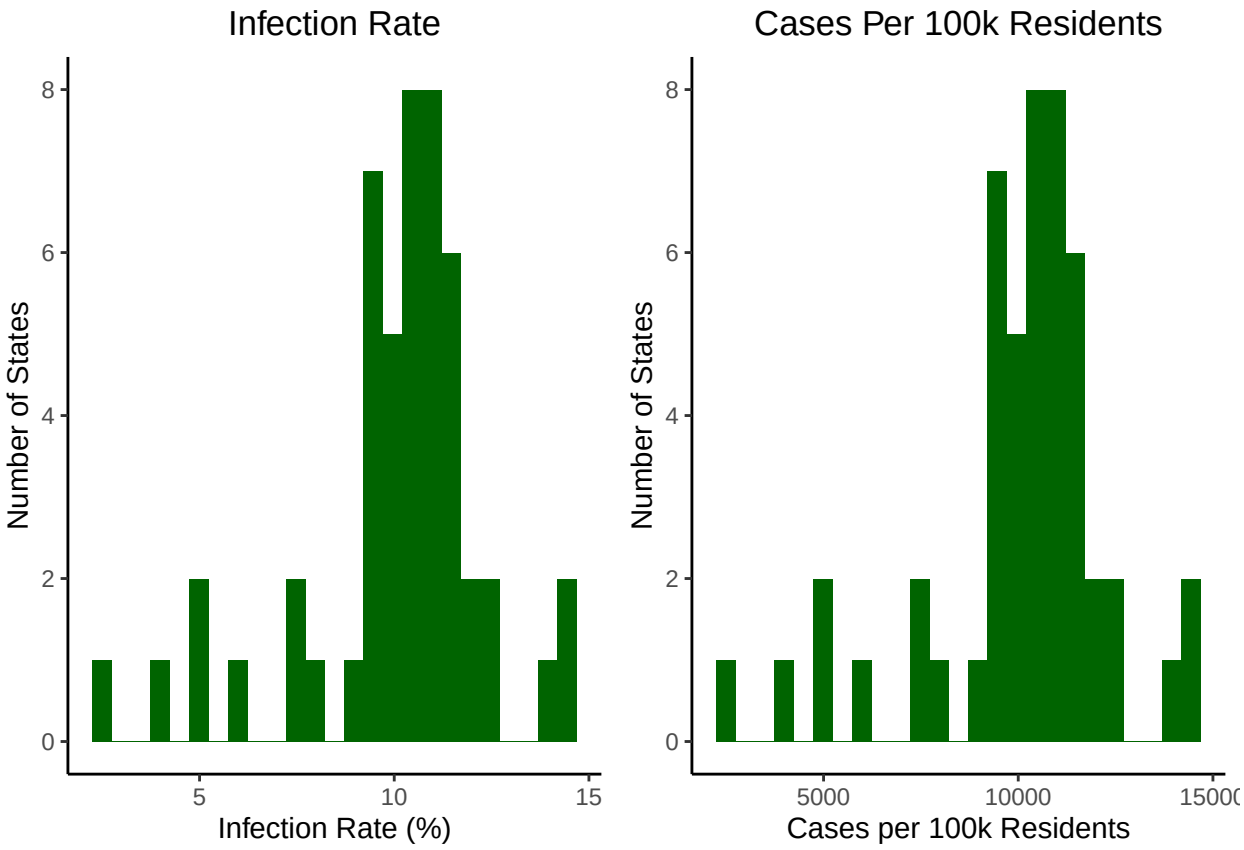
Vaccination rate also appeared bimodal, but we believed differentiating between those partially and fully vaccinated, as well as, dividing by the entire population provided a much more accurate view of vaccine prevalence.



Infection Rate We examined three potential variables for infection prevalence: cases per 100k residents, all time infection rate, infection rate with 7-day rolling avg. Cases per 100k residents had a large skew in its distribution, which we believed would dramatically skew our results.

Generally, the 7 day rolling average number of cases as of the end May in each state were as low as the average had been throughout the entire pandemic. This caused significant skew in the distribution of the average, and downstream caused model results to be difficult to interpret. Because of this, we reverted to using the all time infection rate, defined as total cases / total residents. This variable still represented the current snapshot scenario for each state, but created a more normal distribution while also aligning with our initial understanding of which states experienced the highest infection rates.

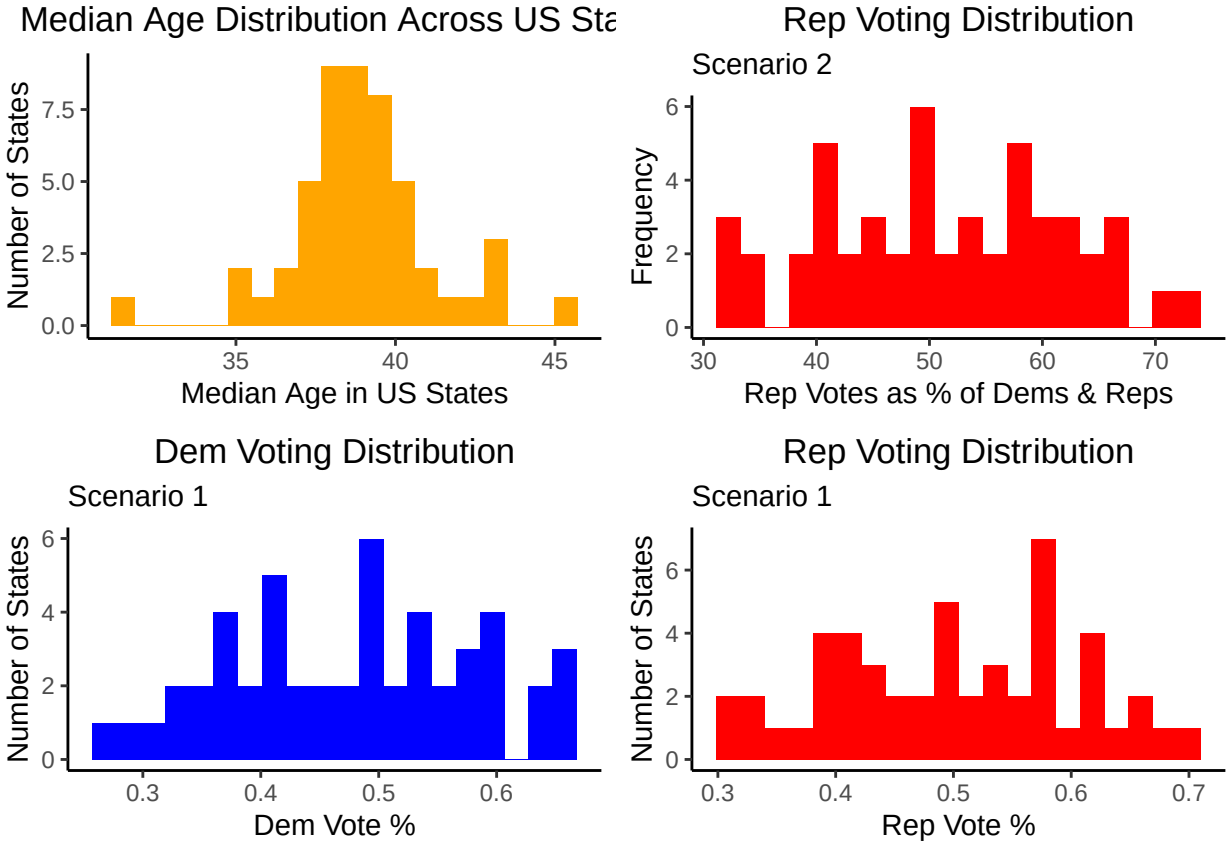
$$\text{Infection Rate} = 100 * \frac{\text{Cumulative Cases}}{\text{People}}$$



Median Age We only examined one potential variable for age, because of our dataset limitations: median age. Media age had an approximately normal distribution with unconcerning outliers.

Rep. Voting Percentage We examined two potential situations for political leaning based on our dataset: statewide results for the 2020 presidential election. In the first scenario, we would use two variables: the percentage of the state who voted for the republican candidate and the percentage of the state who voted for the democratic candidate. Using both variables would allow us to include information about third party voting percentages (100 - the sum of Republican and Democratic voting percentages.) We ultimately decided against this situation, because the variables were nearly perfectly correlated and we would use up degrees of freedom by adding another variable while gaining very little information.

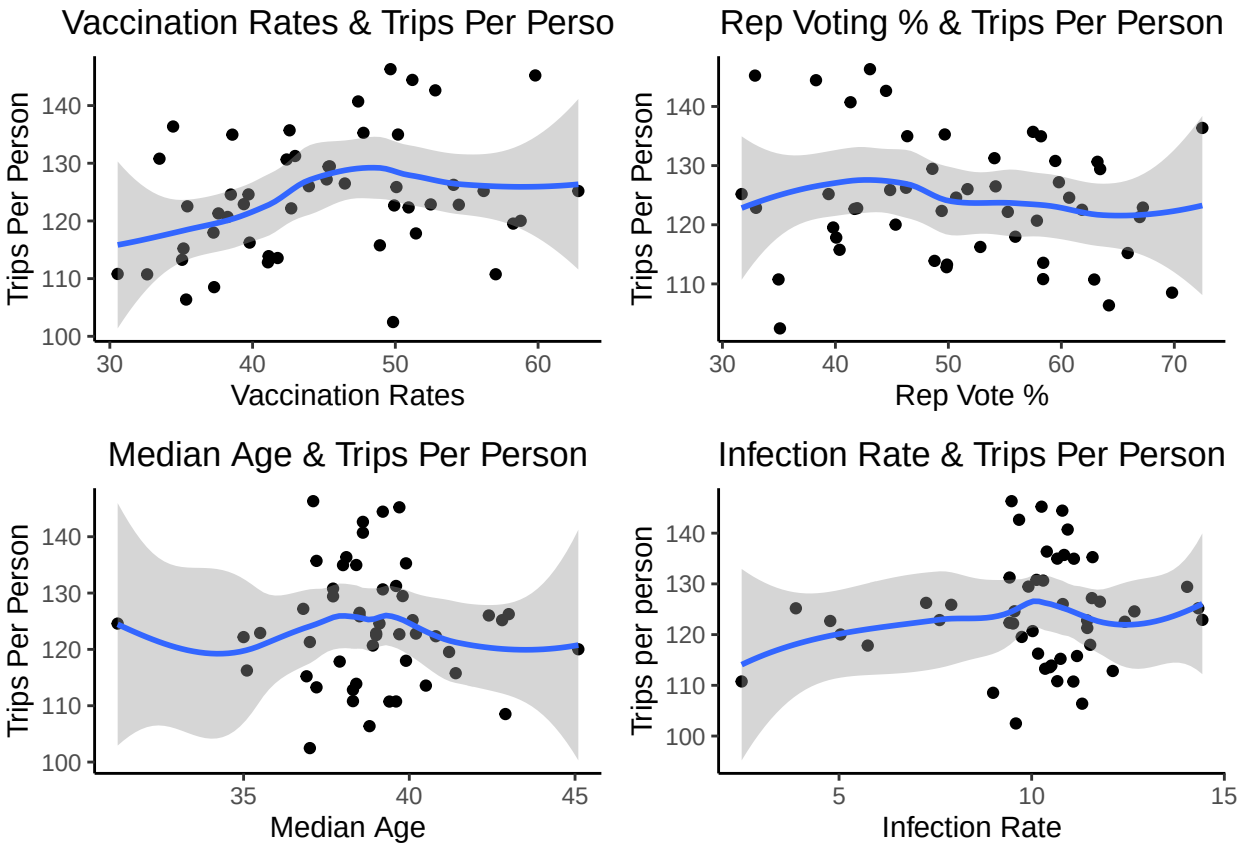
Instead we chose the second scenario, which only considered republican and democratic votes. In this scenario, we used one variable: the percentage of republican votes out of the two major parties. This variable contained information about both major parties and minimized the number of variables we were using. The distribution of this variable was approximately normal with 10 bins, but appeared trimodal with 20. There was one major outlier (DC), which didn't concern us, because we wouldn't be including DC in our date, because many of our other variables didn't contain information for DC.



Variable Relationships & Transformations

In the second phase of our exploratory analysis, we examined the relationships our variables had with each other—most importantly, the relationship with our outcome variable. From the scatterplots we created, we discovered that the variable republican voting percentage had moderate to strong linear relationships with all of our variables except median age. Because of this relationship, we prioritized republican voting percentage in our model specifications hoping it would help reduce noise in our model.

Based on the findings from this section of our exploratory analysis, we elected to not transform any variables. The variables in their current form don't have any dramatic skew that needs to be corrected. Nor do they appear to gain any explanatory information about our outcome variable through transformation.



Model Specifications

Due to our causal question that we're interested in, that is how the vaccination rate impacts trips, we decided to start with the bare minimum model where our output variable is trip per person and input variable is vaccination rate. Then, we decided to build on top of this limited model as follows:

1. Limited Model (Model 1): Output variable is trips per person and input variable is vaccination rate. We looked at the causal impact of vaccination rate on trips per person. This variable also had a stronger relationship with the output variable compared to other variables which was inline with our causal theory.

$$Trips\ per\ Person = \beta_0 + \beta_1 * Vaccination\ Rate$$

Figure 1: Model 1 Formula

2. Model 2: For this model, we added one more variable. Based on our causal theory, we think political views, median age, and infection rate have the potential to impact both the vaccination rate (main causal input variable) and trips per person (output variable). Out of these variables, political views (republican voter percentage) seem to have a stronger relationship with the output variable compared to the other variables. Therefore, this was our next choice of variable addition to the model.

During our EDA (i.e. from our scatterplots), we have also seen republican voter percentage has meaningful linear relationships with vaccination rate and infection rate. This definitely brings up multicollinearity concerns. We decided to follow this modeling specification as this was following our causal theory, but we would later check Variance Inflation Factor (VIF) to assess the degree of multicollinearity to be able to talk about how confident we would be about the outcome of the model.

$$Trips\ per\ Person = \beta_0 + \beta_1 * Vaccination\ Rate + \beta_2 * Republican\ Voter\ Percentage$$

Figure 2: Model 2 Formula

3. Model 3: In this next model specification, we considered adding median age since this variable is also in our causal theory. Both median age and infection rate had similarly seemingly low linear relationships based on scatter plots during our EDA process. We expect both of these variables to impact both vaccination rate and trips per person, so we decided to add these variables into our modeling gradually starting with median age. Median age has a moderate linear relationship with vaccination rate based on our EDA, therefore, this might bring up some multicollinearity issues. However, we still wanted to go with this specification since this aligns well with our causal theory. Then, we would later check VIF to assess the degree of multicollinearity as mentioned in model 2 specification as well.

$$\begin{aligned} Trips\ per\ Person = \quad & \beta_0 + \beta_1 * Vaccination\ Rate + \beta_2 * Republican\ Voter\ Percentage \\ & + \beta_3 * Median\ Age \end{aligned}$$

Figure 3: Model 3 Formula

4. Full Model (Model 4): Last but not least, we're including all measurable/available variables in our causal theory in the full model. This is where our input variables are vaccination rate, republican voter percentage, median age, and infection rate and output variable is trips per person.

$$\begin{aligned} Trips\ per\ Person = \quad & \beta_0 + \beta_1 * Vaccination\ Rate + \beta_2 * Republican\ Voter\ Percentage \\ & + \beta_3 * Median\ Age + \beta_4 * Infection\ Rate \end{aligned}$$

Figure 4: Model 4 Formula

We have mentioned in our EDA section, why we have decided not to do any transformations. This is something we would revisit as part of our modeling and CLM assessments process. Another thing to note in our modeling choices is that we haven't decided to go with building in any interaction variables as we haven't had a strong causal theory in believing that vaccination rates would have different effects on trips per person for different infection rates (as an example). Realistically, it might be possible that the impact of vaccination rate on trips per person may depend on median age or republican party. But, it means a different slope for vaccination rate for every single value of infection rate, median age, or republican voter percentage since all of our variables are continuous variables. And it would make the interpretation a lot harder for anyone who wanted to consume our model and the usefulness of the model output may diminish due to that. Therefore, we leaned towards not doing that in our model building process.

Model Output & Interpretation

As seen in Table 1, our first model shows a statistically insignificant result between vaccination rate and trips per person. For every one point increase in the vaccination rate (e.g. from 40% to 41%), we see 0.37 increase in trips per person. Considering that the minimum and maximum values of trip per person in May 2021 is 103 and 146 for 50 states may not look practically significant either. However, it is expected not to see a meaningful increase in trips per person for only 1 point increase in vaccination rate. For a 10 point increase in vaccination rate (e.g. from 40% to 50%), an expected increase in trips per person would be 3.7 trips increase which is a meaningful increase. Though, this practical significance discussion is not meaningful as this result is not statistically significant.

In our second model, controlling for republican voter percentage made the relationship between our main causal input variable (vaccination rate) and trips per person statistically significant. We think that this is due to the fact that the relationship between republican voter percentage and vaccination rate is moderate to strong and republican voter percentage and trip per person might be also meaningful. We think that this might have helped absorb the noise and helped us measure the impact of vaccination rate on trips per person more strongly. With this additional information into the model, the impact of vaccination rate on trips per person is statistically significant. Every 10 point increase in vaccination rate results in 6.7 increase in trips per person which is a sizable increase in trips per person from a practical point of view as well given the range of values trips per person variable takes (i.e. 103 to 146). Republican voter percentage's impact on trips per person is not statistically significant.

In our third model, we added median age to the mix which has moderate to strong relationship with vaccination rate and some relationship with trips per person. Due to the similar reason we touched above, we think this addition absorbed some more noise and made the relationship between vaccination rate and trips per person more significant. For every 10 point increase in vaccination rate, we expect to see 8.6 trips per person increase. In this model, neither republican voter percentage nor median age has a strong explanatory effect on trips per person.

In our last model, we added infection rate to the previous model because we believed infection rate would impact both the vaccination rate and trips per person per our causal theory. With the addition of this control variable, we have seen minor changes in the impact of vaccination rate on trips per person variable and that relationship still stays statistically significant. For every 10 point increase in vaccination rate would result in an 8.8 increase in trips per person which is a meaningful increase from a practical point of view as well. All other control variables had statistically insignificant impact on trips per person.

Overall, when we put all of these models together, we think the sweet spot of explanation between vaccination rate and trips per person is the last model because Adjusted R-squared is the highest in this model. However, Adjusted R-squared for the last model being 0.133 shows that there are a lot of potential improvements that could be made to the model to increase our explanatory power of the percentage of variability in the trips per person variable if we were able to measure other variables. We will discuss these in the omitted variables section.

```
##
## Table 1: The relationship between trips per person and vaccination rate
## ================================================================================================
##                                                Dependent variable:
##                                    ------------------------------------------------------------
##                                                     trips_per_person
##                                          (1)             (2)             (3)             (4)
## ------------------------------------------------------------------------------------------------
## vaccination_rate                        0.367          0.668*          0.864**         0.879**
##                                        (0.192)         (0.299)         (0.265)         (0.278)
##
## rep_vote_percentage_overall                             0.273           0.319           0.226
##                                                        (0.265)         (0.251)         (0.278)
##
## median_age                                                             -1.148          -0.829
##                                                                        (0.654)         (0.638)
##
## infection_rate                                                                          1.116
##                                                                                        (0.773)
##
## Constant                              107.563***       80.031**        113.500**        93.965*
##                                         (8.445)         (25.748)        (37.354)        (36.521)
##
## ------------------------------------------------------------------------------------------------
## Observations                             50              50              50              50
## R2                                       0.082           0.106           0.158           0.203
## Adjusted R2                              0.063           0.068           0.103           0.133
## Residual Std. Error             9.993 (df = 48) 9.963 (df = 47) 9.778 (df = 46) 9.613 (df = 45)
## ================================================================================================
## Note:                                                          *p<0.05; **p<0.01; ***p<0.001
```

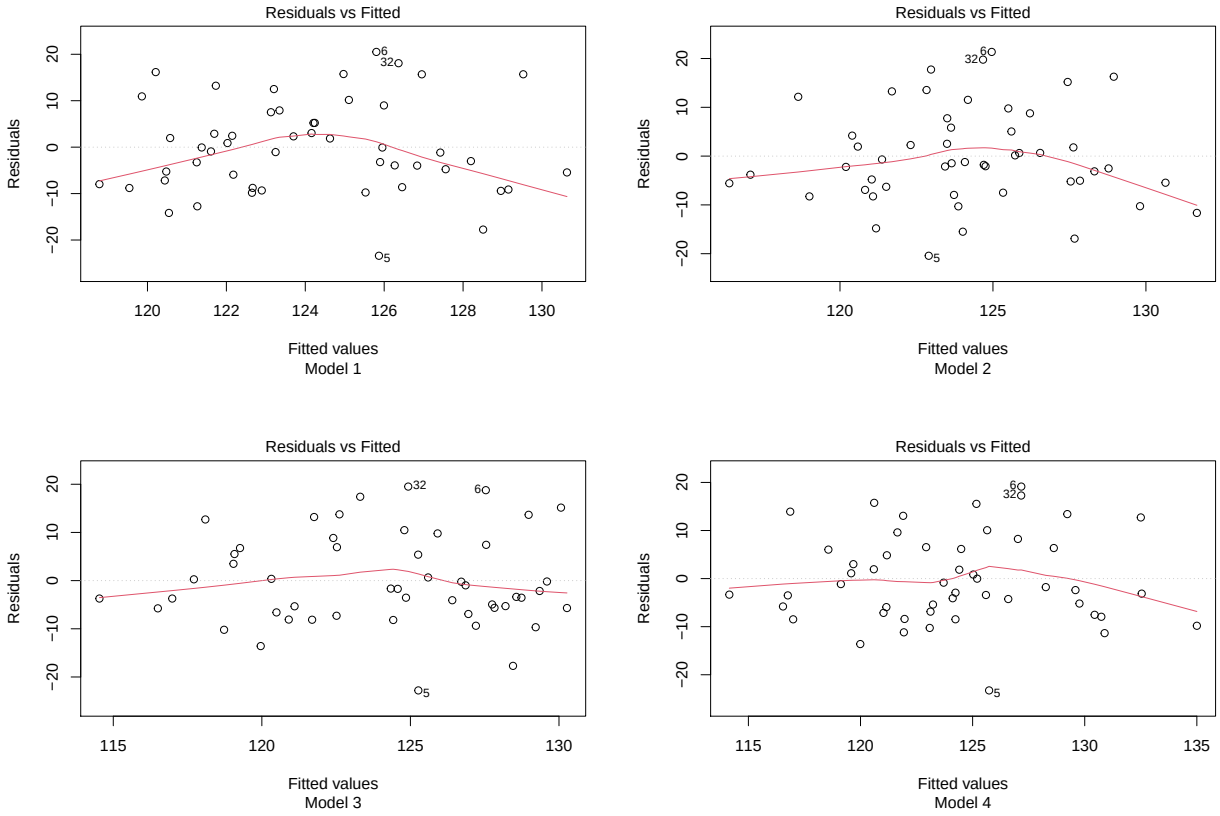
Model Limitations

Up until this point, we haven't fully touched on the limitations of the explanatory model we chose to explain the impact of vaccination rate on trips per person. Since our model is at state-level, we have a pretty small sample size that we need to assess the validity of our model through Classical Linear Model (CLM) assumptions. Without any CLM assessments, we chose the last model to explain the impact of vaccination rate on trips per person as this has the highest Adjusted R2 with the most statistically significant relationship between vaccination rate (main causal variable of interest) and trips per person even though other variables were insignificant. Controlling for other variables helped us to more significantly explain the relationship between vaccination rate and trips per person.

In order for us to feel comfortable with the estimated parameters of the model (i.e. magnitude of the relationship between vaccination rate and trips per person is unbiased), we will need to assess the following assumptions:

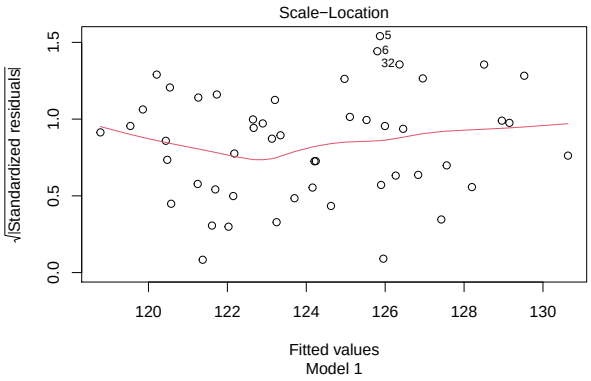
- **IID Sampling:** As mentioned before, our data is at state level. One state's observation related to trips and vaccination rates as well as infection rates may be able to help us understand the other state's situation from the same aspects. There might be some dependency between states. Therefore, IID sampling may not be addressed. However, we want to explain the impact vaccination rates have on people's mobility and some geographical definition is required to assess this. This assumption may not be fully addressed however we define the geography as some level of dependency may persist.
- **No Perfect Collinearity:** Since none of the variables were dropped by our modeling procedure in any of the models, this assumption is addressed.

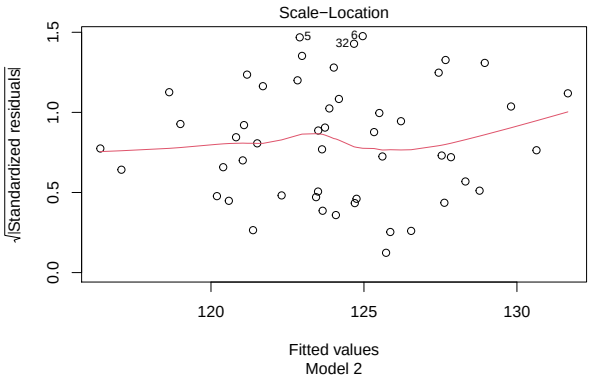
- **Linear Conditional Expectation:** Looking at the residuals vs fitted graphs below, model 3 and 4 seem to reasonably address this assumption. We see some slight deviation from zero in residuals in model 4 on the right hand side of the graph. However, since it is mainly driven by one data point (outlier), we are not concerned with this.

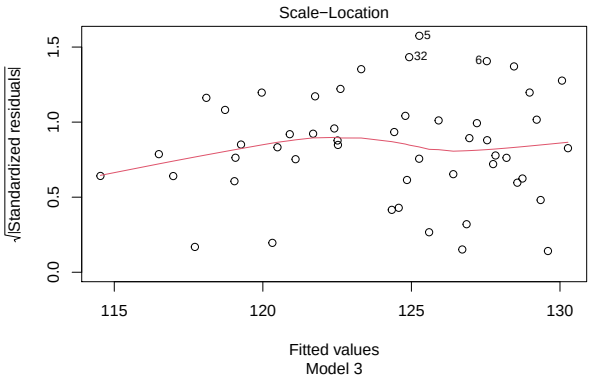


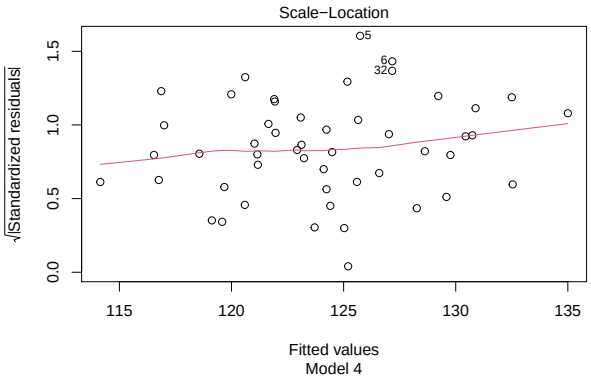
In order for us to feel comfortable with standard errors and hypothesis tests of the estimates, we will need to assess the following assumptions:

- **Homoskedastic Errors:** Based on both the residual vs fitted graph and scale location graphs below, models 3 and 4 look like reasonably they have low heteroskedasticity. Based on the Breusch-pagan test results below, we do not reject the homokedastivity null hypothesis for any of the models. Therefore, we think this assumption is reasonably addressed for models 3 and 4.









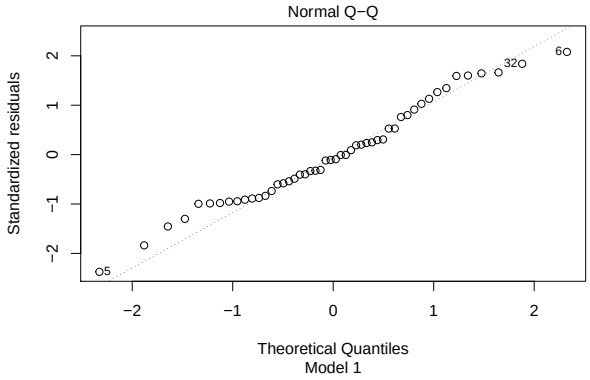

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 1.4892, df = 1, p-value = 0.2223
```

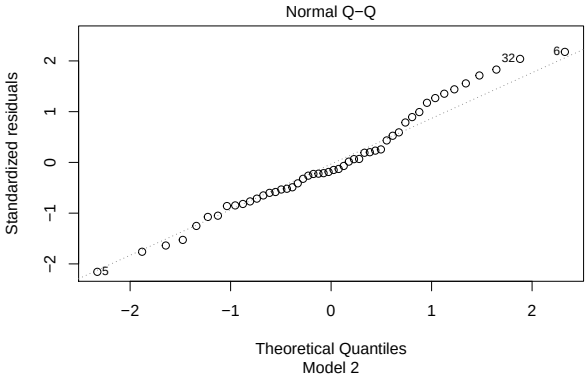
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 3.9078, df = 2, p-value = 0.1417
```

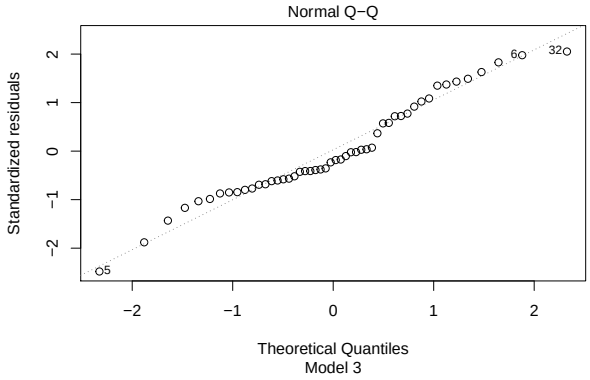
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 7.224, df = 3, p-value = 0.06509
```

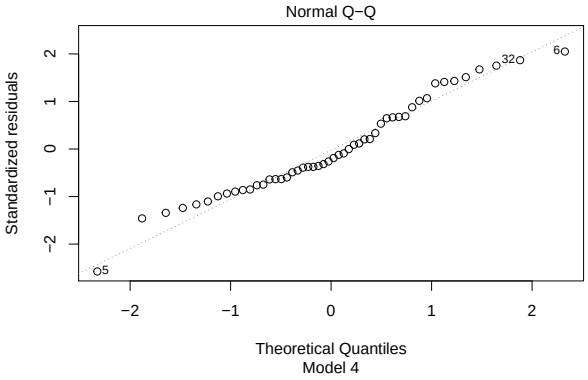
```
##
##  studentized Breusch-Pagan test
##
## data:  model4
## BP = 6.1707, df = 4, p-value = 0.1868
```

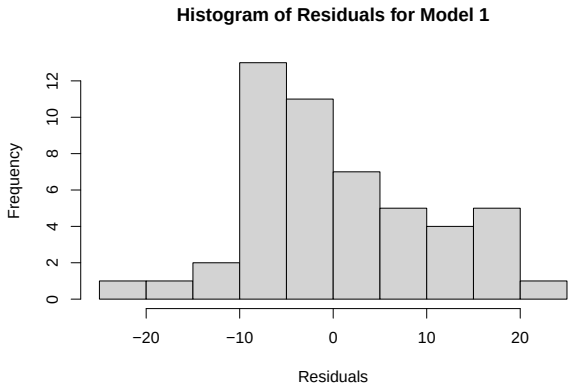
- **Normally Distributed Errors:** Based on the Q-Q plots below, distribution of errors for models 1 and 2 seem to be slightly deviating from the theoretical normal distribution. For models 3 and model4, we don't have significant concerns as the Q-Q plots seem similar to theoretical normal distribution since the majority of the data points are on the diagonal line. We also looked at the histogram of the residuals below which shows close to normal distribution for all models. Therefore, this assumption is reasonably addressed as well for our model of choice (i.e. model 4).

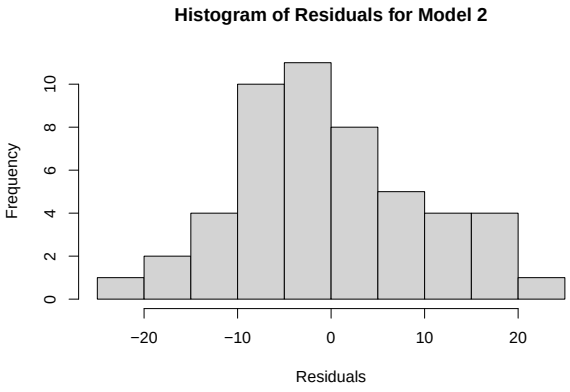


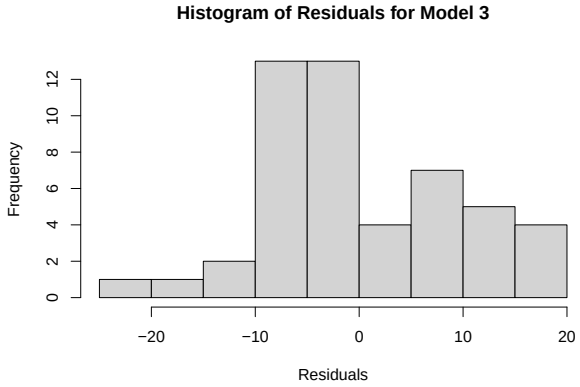


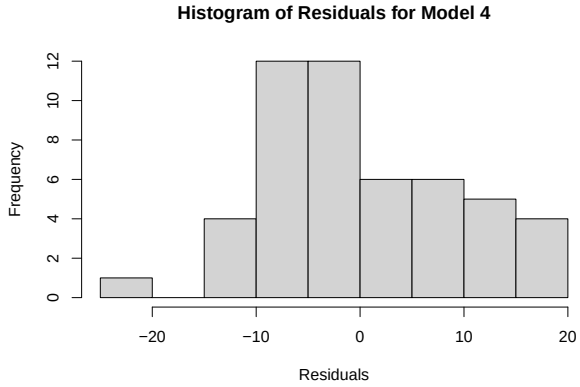












Overall, we feel comfortable with unbiasedness of our estimate of vaccination's impact on trips as well as the hypothesis test assessment on significance of the relationship between our main variable (vaccination rate)'s impact on trips. One last thing we want to assess is the multicollinearity in order to feel comfortable with the actual assessment and the estimated relationship between vaccination rate and trips per person. As we assessed below with VIF for all the models except the first one (since there is no notion of multicollinearity with a single input variable model), all the VIF values are are less than 4, we're okay including all these variables in our models (especially the last model we have chosen as a result of our analysis).

```
## [1] "Variance Inflation Factors for Model 2"
```

```
##            vaccination_rate rep_vote_percentage_overall
##                    3.247062                    3.247062
```

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

```
## [1] "Variance Inflation Factors for Model 3"
```

```
##            vaccination_rate rep_vote_percentage_overall
##                    3.701834                    3.289850
##                  median_age
##                    1.265473
```

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

```
## [1] "Variance Inflation Factors for Model 4"
```

```
##            vaccination_rate rep_vote_percentage_overall
##                    3.705088                    3.489567
##                  median_age              infection_rate
##                    1.374497                    1.464137
```

Discussion of Omitted Variables

We identified three omitted variables whose omission introduced additional bias into our model: the percentage of people who work remotely, the percentage of healthy citizens, and the percentage of people who believe in general scientific consensus. We omitted these variables, because we were unable to procure the appropriate data.

Percentage of People Working Remotely The percentage of people working remotely in a state has a negative relationship with the variable—number of trips per person—because people who work remotely don't commute to and from work daily. Additionally this variable has a negative relationship with the vaccination rate, because there is less incentive to get vaccinated if you aren't interacting with coworkers in person. Together, these relationships create an overall positive bias in our model that is directionally away from zero. This estimated moderate-to-strong effect suggests that our results may appear more significant than they are.

While we were unable to find a stand-in for this variable, many of our controlling variables, such as infection rate and republican voting percentage, likely contain information about the percentage of people working remotely.

Percentage of Healthy Citizens The percentage of healthy citizens in a state has a positive relationship with the average number of trips per person, because healthy people likely travel more than unhealthy people, especially during a pandemic. Additionally, we expect the percentage of healthy citizens to have a negative relationship with vaccination rate, because healthy citizens have less need to get a vaccination; many weren't even eligible until recent months. Together these relationships create an overall negative bias directionally towards zero. We expect this weak-to-moderate effect to make our results appear less significant than they are. We believe we captured some of this information with median age, as age likely heavily correlates with health. However the median inherently neglects key age information; the mean would have provided a better stand-in for this variable.

Percentage of People Who Believe in Scientific Consensus People who believe in scientific consensus likely avoid unnecessary travel, because of the recommendations from scientific organizations, suggesting a negative relationship with the number of trips per person. Additionally, these individuals are more likely going to believe in the importance of vaccines suggesting a positive relationship with vaccination rate. Together these relationships create an overall negative bias directionally towards zero. We expect this moderate-to-strong effect to make our results appear less significant than they are. Some of this information is likely contained in our political leaning variable, republican voting percentage, but how much is hard to say.

Omitted Variable Bias Conclusion

Overall, we have two omitted variables with a bias towards zero and one away from zero. We expect much of this to cancel out and our overall omitted variable bias to be directionally towards zero with a weak-to-moderate effect. Therefore our bias doesn't give us strong reason to question the significance of our results and suggests that there is likely more significance than we discovered.

Conclusion

Our findings support our causal theory that vaccinations drive people to take trips. In our limited model, we didn't achieve significant results, but when we included other controlling variables, we obtained a high level of significance. Because of our unit of observation we were only able to feed the regression function fifty observations. However, since we were able to meet four out of five of our CLM assumptions, we can feel confident that our results will hold in subsequent analyses.