The Effects of Travel on COVID-19 Cases

W203 Lab 2

Chandler Haukap, Courtney Smith, Hassan Saad

2021-08-02

1. Introduction

1.1 History and Context

The COVID-19 pandemic changed, at least temporarily, how and why we travel. Highways that were normally packed with commuters were suddenly deserted as offices closed indefinitely and shelter in place advisories rolled out. International borders closed and interstate movement came with a slew of quarantine regulations.

Despite the threat of the virus and the state-mandated travel restrictions, there were still thousands of local and interstate trips taken. Some people traveled hundreds of miles for leisure trips and others only left their homes to get groceries and other necessities. This travel spread the virus to every corner of the globe within the span of a few months.

With the initial wave behind us, the threat of the Delta variant rising, and a wealth of data waiting to be analyzed, we are presented with a novel opportunity to study how travel is shaping the pandemic. Our proposal aims to combine data from the Bureau of Transportation Statistics, the New York Times, and the Census Bureau to explore the impact of long-range travel and day-to-day mobility on COVID-19 case rates on a county level.

1.2 High Level Description of Data Sources

Our analysis will use population estimates, mobility reports, and COVID-19 case data.

Population Estimates

Population estimates are calculated by the United States Census Bureau[1]. The Census Bureau defines their estimate as:

$$Estimate = BasePopulation + Births - Deaths + Migrations$$

where the base population is from the 2010 census.

[1] https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-counties-total.html

Trips By Distance

Trip totals come from the Bureau of Transportation Statistics[2]. The data was collected by the Maryland Transportation Institute and Center for Advanced Transportation Technology Laboratory at the University of Maryland. Data was generated using anonymized cell phone data and categorized by trip length. A trip, according to the documentation, is:

> *movements that include a stay of longer than 10 minutes at an anonymized location away from home*

A movement with multiple stays of longer than 10 minutes at separate locations is counted as multiple trips. Trips include all modes of transportation. Data are not reported for a county if it has fewer than 50 mobile devices sampled on any given day, but considering the fact that Loving, Texas, with a population of 181, is contained in the dataset, we believe that even small counties are well-represented.

New York Times Cases Data

The number of daily cases comes from the New York Times COVID Dataset [3]. Data is sourced from state and local health agencies. Case counts are the total of both confirmed and probable cases. Confirmed cases indicate positive PCR test results. Probable cases include positive antigen and antibody test results, or evaluation by public health officials based on state or federal health department criteria.

This data is cumulative, meaning for a given day and county it counts all of the people that have been infected up until midnight Eastern time on that day. Cases are counted on the date they are first announced.

County FIPS Codes

The BTS data and the NYT data can be joined on the county FIPS code, but the Census data does not contain FIPS codes. To join the Census data to the other data, we used a mapping from FIPS to county and state found on the United States Department of Agriculture's website[4]

2. Model Building Process

2.1 Explanation of Variables/ Distributions

The BTS trips data is split into 10 categories based on distance traveled: less than 1 mile, [1,3) miles, [3,5) miles, [5,10) miles, [10,25) miles, [25,50) miles, [50,100) miles, [100,250) miles, [250,500) miles, and 500+ miles. We consolidated these categories into short, mid-range, and long trips. We converted the new variables from direct trip count to trip count per 100 people and converted case counts to cases per 100 people.

We looked at case and trip counts per 100 people on a weekly basis instead of daily to account for volatility in reporting and testing practices (for example, testing rates are much lower on Sundays). We sliced the data to look at trip and case counts by county for a single week, Week 22 of 2021, as a baseline before identifying new cases in Week 23 and estimating a relationship between the change in cases and trips taken the week prior.

We also created an indicator variable for densely populated urban counties (population $> 50{,}000$, as per the US Census Bureau definition [5]). Given the different perspectives and stances people have taken towards

[2]https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv
[3]https://github.com/nytimes/covid-19-data
[4]https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697
[5]https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html

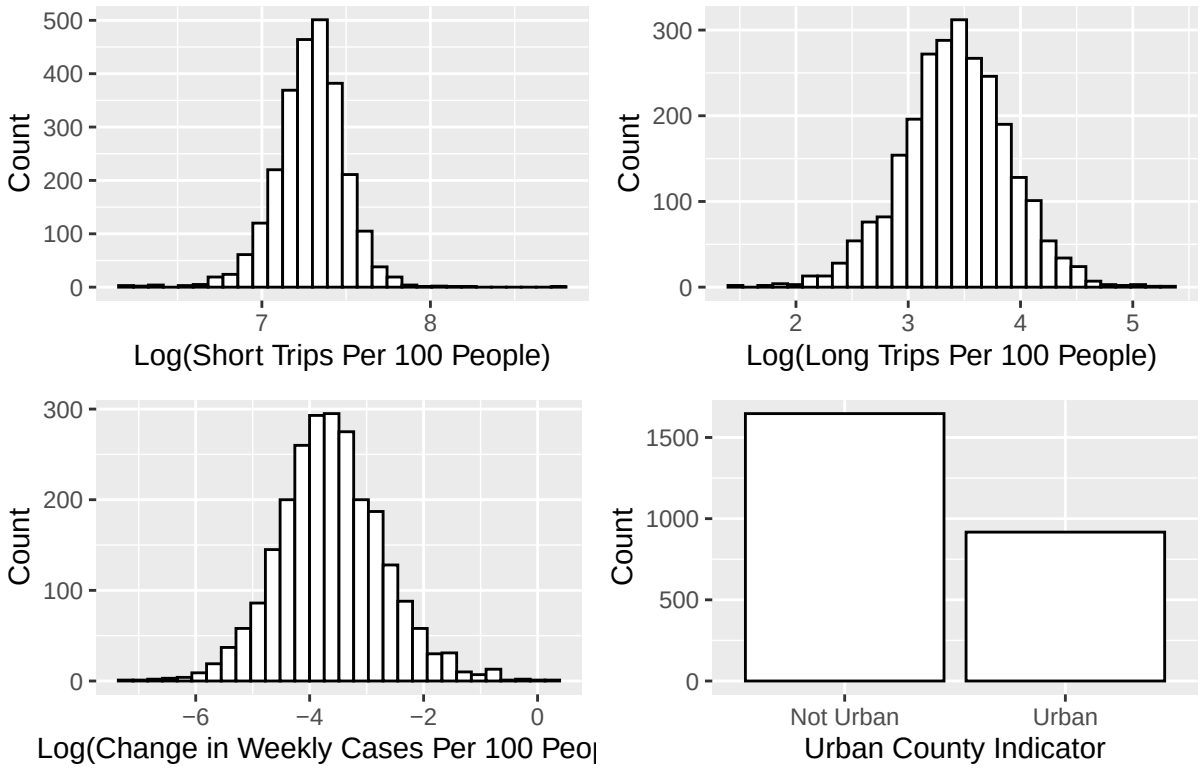
COVID over the last year and a half, it seemed important to account for any underlying groupings that wouldn't be brought out without a control term. This dummy variable then gets reapplied as an interaction variable in our extended model to see if this same grouping has an impact on the way our most important covariate `long_pp` interacts with the cases.

Table 1: Model Variables

Names	Descriptions
long_pp	Long Trips per 100 People Taken Week 22 of 2021 (over 100 miles)
mid_pp	Mid Range Trips per 100 People Taken Week 22 of 2021 (between 5 and 25 miles)
short_pp	Short Trips per 100 People Taken Week 22 of 2021 (under 3 miles)
urban	Binary Variable Representing Whether County is Urban (over 50,000 people)
delta_weekly_cases_pp	New Cases per 100 People in Week 23 of 2021

After deciding to use a control variable representing whether or not a county is urban, we checked that the sample size of each group was large enough. If one group defined by the control variable were composed of very few samples and there existed low variance within those samples, it could skew the prediction for this group. Both have sample sizes of at least several hundred, and the data series has a mean of 0.3576, which gives us confidence in our ability to use this variable effectively in our model. We also explored the shape of the other data series, and we comment on our findings in section 2.3.

Figure 1: Distribution of Variables



2.2 Research Question

The purpose of this study is to compare the effects of longer-range travel versus local mobility on COVID-19 case rates at a county level, controlling for population density.

Our research question is: **What kind of travel (local or long distance) is more strongly correlated with an increase in the proportion of COVID-19 cases in a specific county?**

We will create a series of linear models that regress the number of COVID-19 cases per 100 county residents against characteristics of mobility within each respective county. The models will be descriptive and the coefficients will be representative of the effect each type of travel has on COVID case counts.

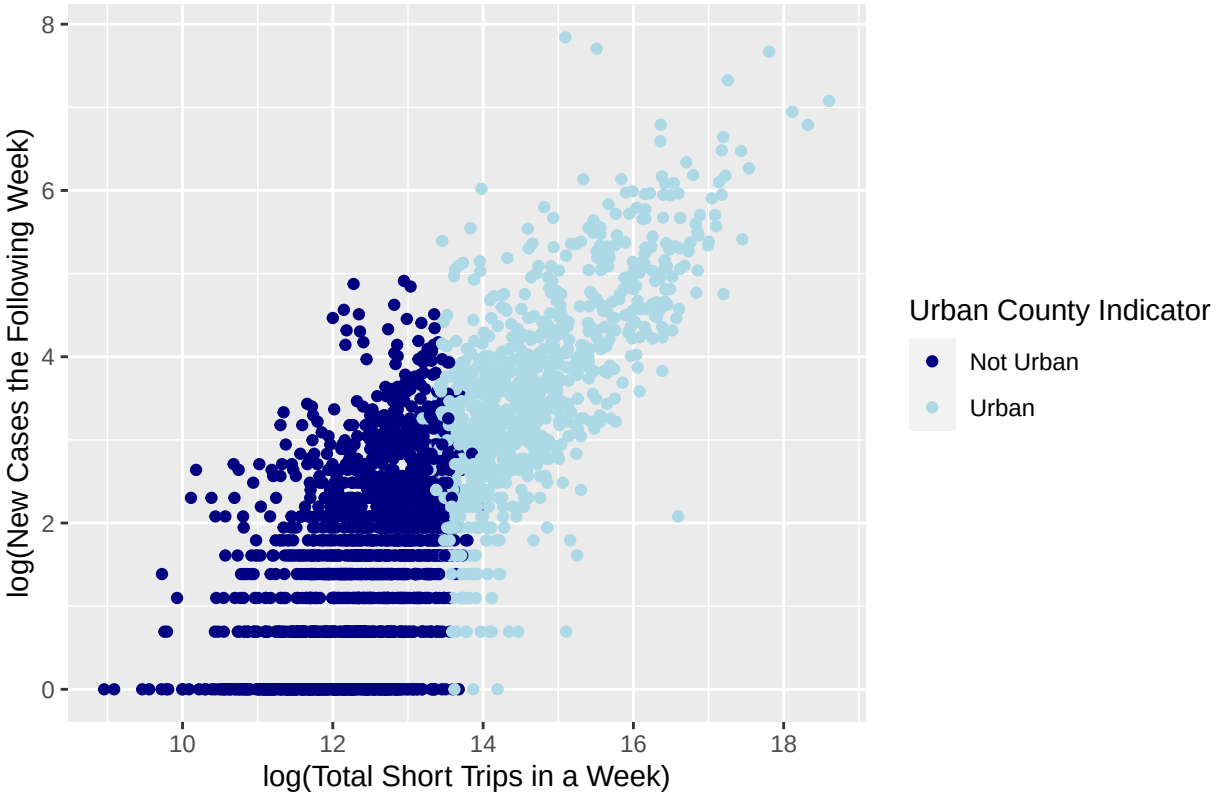
2.3 Model Building

After running our EDA, we agreed on a couple of items that would lead into our model building process:

1) As shown in section 2.1 above, we determined that the trips per person (`short_pp`, `mid_pp`, `long_pp`) had nice normal distributions, as did the dependent variable `delta_weekly_cases_pp` after we carried out a `log()` transformation on them. We therefore decided to implement these transformations in our models. Because we wanted to use `urban` as a control variable depending on a certain population threshold, we don't need to carry out a transformation for this variable.

2) The following plot is one that we created after cleaning our data set, which made us confident that we'd find at least some positive correlation between quantity of trips and case counts. We observed that there is a strong positive correlation between the number of trips taken and new COVID cases. Our goal was to find out if there was a stronger correlation between case counts and long trips (which could signify a vacation) or case counts and short trips (e.g. trips to the grocery store or pharmacy).

Figure 2: Cases and Trip Counts



We speculated that long trips are representative of a greater opportunity for COVID-19 transmission in comparison to short trips. The underlying travel would likely involve more time away from home and therefore more opportunity to make contact with unknown people. We also hypothesized that a large subset of the shorter trips could consist of walks around the block, trips to the grocery store, or other safe activity that is carried out in a more responsible manner, therefore yielding a relatively smaller correlation with case counts in comparison to long trips.

For all of our models, we chose COVID case counts in each county as our dependent variable. As noted previously, most datasets represent case numbers as a cumulative sum with respect to the date. For our research question, we are more interested in modeling the effects of travel within one week on case counts in the next week. This being said, we had to implement a delay factor in our dataset. According to one of our sources,[6] the median time a person develops symptoms after a COVID exposure is between 4 and 5 days. Therefore, for each sample in the data, we use the cumulative trip information from week 22, and we explore the number of new cases for that same sample (the specific county) at the end of week 23.

Below are three models we developed to attempt to model the correlation between long trips taken during Memorial Day weekend and new COVID cases.

2.3.1 Limited Model

Our limited model regresses new cases on trips over 100 miles. :

$$\log(delta\_weekly\_cases\_pp) = \beta_0 + \beta_1 * \log(long\_pp)$$

This yields the linear model:

[6] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

$$\log(delta\_weekly\_cases\_pp) = -4.5134 + 0.2614 * \log(long\_pp)$$

The interpretation of this model is: for every 1% increase in the number of trips taken per 100 people within a county, the number of new COVID cases per 100 people goes up 0.26% the next week, within the same county.

To make this more clear with an example:

> Suppose County A has a population of 200,000 people, and 40,000 long trips are taken in week 22 (i.e., 20 trips per 100 people). In week 23, there are 50 *new* covid cases in County A. Now suppose that instead of 40,000 trips, 80,000 trips were taken. That's a 100% increase in trips per 100 people. We can expect that the number of new COVID cases in week 23 would be 26.14% higher, meaning there would instead be 63 (truncating from 63.07) *new* cases in week 23 instead of 50.

This seems like a very small effect until considering that one county had upwards of 1.5 million long range trips in week 22 of 2021. Moreover, it's important to note that our model does not represent the estimate that a 1% increase in trips results in a 0.26% increase in total cases.

Table 2: Limited Model Coefficients

Term	Estimate	Rob. Std. Error	T-Stat	P-Value
(Intercept)	-4.5133638	0.1336703	-33.518024	0
log(long_pp)	0.2614233	0.0393083	6.715654	0

The table above shows that we can consider our dependent variable that our `long_pp` variable is statistically significant (given the very small P-value). That is to say that the coefficient for the variable representing 'long trips per 100 people' is not zero. However, we were a little disappointed to see how small our coefficient of determination was when running the `summary()` function on this model. We got an R^2 value of 0.0181. It signifies that, while our dependent variable `long_pp` contributes to the model, the majority of the variation within our dependent variable is not captured by variation within `long_pp`. In other words, `long_pp` is statistically significant as an independent variable but lacks substantially in practical significance.

We thought back to why this was the case, especially after having seen the seemingly obvious correlation in Figure 2. It became clear after some brainstorming that between our exploratory data analysis and creating our models, we had standardized the variables by each county's respective population. We are not 100% sure why this would affect our R^2 value so much- after all, dividing the dependent and independent variable by the same factor within each sample should leave us with the same proportional effect. What we did conclude, however, was that our scatter plot in Figure 2 was representing a different and more trivial relationship than the one we set out to model. The correlation present in the plot represents the simple fact that counties with higher populations experience more trips as well as more new weekly COVID cases (compared to counties with lower populations).

2.3.2 Model Two

Our second model regresses new cases on trips over 100 miles, trips under 3 miles in distance, and on the control variable `urban`, which again indicates whether a county has a population greater than 50,000 people. We also implement a `log()` transformation on `short_pp` for the same reasons noted in our EDA.

We were not sure how the effects would differ between short and long travel (hence our proposed question). However, looking at the scatter plot above, we anticipated that the correlation between `urban` and `delta_weekly_cases_pp` would be positive. The final model is:

$$\log(delta\_weekly\_cases\_pp) = \beta_0 + \beta_1 * \log(long\_pp) + \beta_2 * \log(short\_pp) + \beta_3 * urban$$

And adjusted to reflect the values of each coefficient:

$$\log(delta\_weekly\_cases\_pp) = -0.0224 + 0.1986 * \log(long\_pp) - 0.5798 * \log(short\_pp) - 0.1433 * urban$$

Table 3: Model 2 Coefficients

Term	Estimate	Rob. Std. Error	T-Stat	P-Value
(Intercept)	-0.0223872	0.6827290	-0.0335205	0.9732622
log(short_pp)	-0.5797923	0.0920898	-6.4749179	0.0000000
log(long_pp)	0.1985950	0.0444934	4.7021855	0.0000027
urban	-0.1433351	0.0423240	-3.4137448	0.0006512

The interpretation of the `log(short_pp)` variable is similar to that of `log(long_pp)` in our limited model. However, it has the opposite correlation with respect to `log(delta_weekly_cases_pp)`. For every 1% increase in the number of short trips taken per 100 people within a county, the number of new COVID cases per 100 people goes down 0.58% the next week, within the same county, holding the number of long trips per 100 people constant. We expected that short trips would have a smaller positive correlation with respect to case counts relative to long trips. It is slightly surprising, however, to see that the correlation is negative. We could theorize that the presence of increased short trips is indicative of a population that has a tendency to limit their movement to essential travel only, but we need to be cautious to not overextend our assumptions.

The coefficient for our `urban` control variable was also not immediately intuitive. Again, we reverted back to Figure 2 and noticed that all the samples representing urban counties experienced higher case counts, so we initially anticipated seeing a positive coefficient for this variable. The logic behind this was that urban counties are more densely populated, providing a setting for COVID to be more easily transmitted. Our model suggests that this intuition is also incorrect.

After doing some research, we again realized that this has to do with the fact that we standardized our dependent variable (new cases) by each county's population. Figure 2 is instead representative of the fact that urban counties have higher populations and therefore more overall cases relative to urban ones. According to the CDC website, in addition to having lower vaccination rates, "rural communities often have a higher proportion of residents who lack health insurance, live with comorbidities or disabilities, are aged >65 years."[7] It therefore makes sense for our `urban` control variable to be negative in this context. In urban communities ($urban = 1$), we end up with the equation:

$$\log(delta\_weekly\_cases\_pp) = -0.1657 + 0.1986 * \log(long\_pp) - 0.5798 * \log(short\_pp)$$

In rural communities ($urban = 0$), we get a model with a slightly higher intercept (representing a higher baseline of new cases across weeks 22 and 23) for our dependent variable:

$$\log(delta\_weekly\_cases\_pp) = -0.0224 + 0.1986 * \log(long\_pp) - 0.5798 * \log(short\_pp)$$

The coefficient of this control variable is a large value in the context of the the rest of the model. The percentage change in new cases caused by a 1% increase in trips is 0.199%. Given that rural counties can expect to see 0.1433% more new cases given the same change in the number of trips, this is a practically significant value (again, with respect to the rest of the model). It also shows statistical significance given that the P-value in the table above is well below 0.05.

[7] https://www.cdc.gov/mmwr/volumes/70/wr/mm7020e3.htm

2.3.3 Model Three

In this model, we add the variable `log(mid_pp)` to represent mid-ranged trips on week 22. We also add an interaction term between `log(long_pp)` and `urban` to see how if there is a difference in the coefficient for `log(long_pp)` between urban and rural counties.

$$\begin{aligned}\log(delta\_weekly\_cases\_pp) =& \beta_0 + \beta_1 * \log(long\_pp) + \beta_2 * \log(short\_pp) \\ &+ \beta_3 * \log(mid\_pp) + \beta_4 * urban + \beta_5 * \log(long\_pp) * urban\end{aligned}$$

We end up with the following coefficients:

$$\begin{aligned}\log(delta\_weekly\_cases\_pp) =& 1.1635 + 0.0994 * \log(long\_pp) - 0.5611 * \log(short\_pp) \\ &- 0.1456 * \log(mid\_pp) - 1.1411 * urban + 0.2996 * \log(long\_pp) * urban\end{aligned}$$

For rural communities (`urban` = 0), we get the equation:

$$\log(delta\_weekly\_cases\_pp) = 1.1635 + 0.0994 * \log(long\_pp) - 0.5611 * \log(short\_pp) - 0.1456 * \log(mid\_pp)$$

For urban communities (`urban` = 1), we get the equation:

$$\log(delta\_weekly\_cases\_pp) = 0.0224 + 0.3990 * \log(long\_pp) - 0.5611 * \log(short\_pp) - 0.1456 * \log(mid\_pp)$$

The interaction term in this model acts somewhat similarly to the control variable we added in Model 2. However, it also creates a different coefficient for the `log(long_pp)` term between urban and rural counties. The rate of change created by an increase in long trips for rural counties is a little over 4 times that for urban counties. The interaction term is statistically significant given the P-value noted in the table below (meaning, we reject the null hypothesis that the value of the coefficient is 0). However, we note that this interaction term does not do much to increase the coefficient: $R^2 = 0.0456$. It also increases the P-value for the `log(long_pp)` over our established 0.05 threshold.

Table 4: Extended Model Coefficients

Term	Estimate	Rob. Std. Error	T-Stat	P-Value
(Intercept)	1.1634707	0.7965264	1.434520	0.1515523
log(long_pp)	0.0994184	0.0542833	1.959797	0.0501333
log(short_pp)	-0.5610652	0.0932587	-6.260783	0.0000000
log(mid_pp)	-0.1456052	0.0801328	-2.014053	0.0441137
urban	-1.1411304	0.3028811	-3.716329	0.0002067
log(long_pp):urban	0.2995482	0.0925702	3.241189	0.0012064

3. Regression Tables/ Charts

3.1 Stargazer Regression Table

The table below shows the coefficients implemented in all three of the models discussed above. We made sure to include the robust standard errors in this table.

```
##
## Regressing New Cases on Trips Taken
## ================================================================================================
##                                              log(delta weekly cases pp)
##                        ------------------------------------------------------------------------
##
##                              Limited Model              Model 2              Inclusive Model
##                                   (1)                     (2)                     (3)
## ------------------------------------------------------------------------------------------------
## log(short_pp)                                          -0.580                  -0.561
##                                                       (0.092)                 (0.093)
##                                                      t = -6.296              t = -6.016
##                                                      p = 0.000               p = 0.000
##
## log(mid_pp)                                                                  -0.146
##                                                                             (0.080)
##                                                                             t = -1.817
##                                                                             p = 0.070
##
## log(long_pp)                      0.261                  0.199                  0.099
##                                  (0.039)                (0.044)                (0.054)
##                                 t = 6.651              t = 4.463              t = 1.831
##                                 p = 0.000             p = 0.00001             p = 0.068
##
## urban                                                  -0.143                  -1.141
##                                                       (0.042)                 (0.303)
##                                                      t = -3.387              t = -3.768
##                                                      p = 0.001               p = 0.0002
##
## log(long_pp):urban                                                           0.300
##                                                                             (0.093)
##                                                                             t = 3.236
##                                                                             p = 0.002
##
## Constant                         -4.513                 -0.022                  1.163
##                                  (0.134)                (0.683)                (0.797)
##                                t = -33.765             t = -0.033             t = 1.461
##                                 p = 0.000              p = 0.974              p = 0.145
##
## ------------------------------------------------------------------------------------------------
## Observations                      2,443                  2,443                  2,443
## R2                                0.018                  0.040                  0.046
## Adjusted R2                       0.018                  0.039                  0.044
## Residual Std. Error    0.915 (df = 2441)      0.905 (df = 2439)      0.903 (df = 2437)
## F Statistic         45.100*** (df = 1; 2441) 34.090*** (df = 3; 2439) 23.302*** (df = 5; 2437)
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

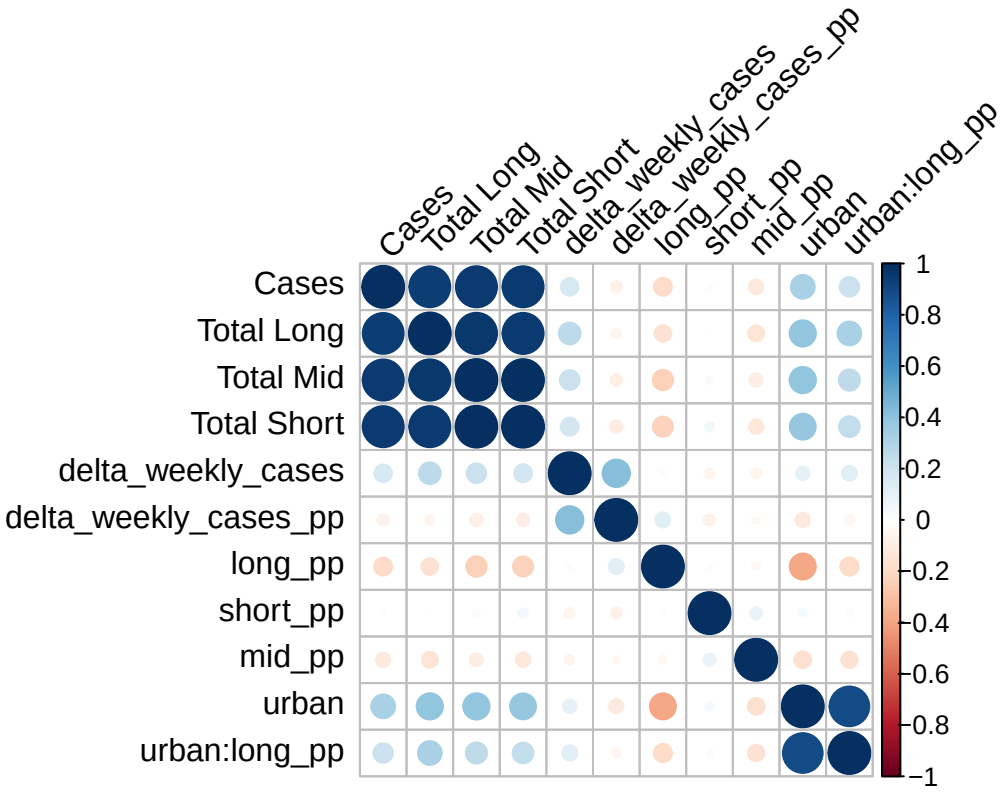
3.2 Covariate Correlation Plot

We wanted to produce a chart that would help us understand the correlation between all the different independent variables after creating our models. We were not surprised when we saw very high correlations between certain variables before standardizing by population (this is the cluster of large blue points in the upper left corner of the chart). Note that we didn't include these variables in our models, but we wanted to include them in the chart to reiterate our previous oversight.

More importantly, the chart below also does a good job of representing the relationships we've discussed within our models. Specifically, we confirm that `long_pp` and `delta_weekly_cases_pp` have a mild positive correlation (a small, light blue dot), and that `urban` and `delta_weekly_cases_pp` have a small negative correlation (the slightly larger, but orange dot).

Lastly, we can use it to make sure we don't have too much collinearity between the terms of interest in our models (that is, between our independent variables). There are no large or densely colored dots between any of the covariates we selected to represent our independent variable.

Figure 3: Correlation Chart



4. Limitation of Model (Assessment of CLM Conditions)

4.1 IID

There are a few dependencies in our data worth mentioning.

4.1.1 Time

The number of cases for a given county is heavily dependent on the number of cases in that county on the previous day. To overcome this dependency we limited our research to a *single week*, week 22 of 2021. This coincides with the dates of May 31st to June 6th of 2021. Where previously we measured effects between one day and the next, grouping the data this way was sufficient to overcome the time dependency since our samples instead consist of differences across counties.

4.1.2 Proximity

We are looking at travel information. For long trips, this implies leaving one county and entering another. Therefore, cases in one county might be dependent on cases in another county. We decided not to account for dependence both for practical reasons and theoretical reasons.

In theory, each county's travel counts per capita were very similar. Therefore, any given county has roughly the same chance of giving a traveler COVID as they do getting a case of COVID from a traveler.

In practice, we do not know the destination of the recorded trips. This is for the privacy of cell phone carriers, but unfortunately, it prevents us from accounting for inter-county travel.

Given the fact that in theory each county should be impacted by inter-county travel equally, we do not expect this dependency to impact the credibility of this study.

4.1.3 County Selection

All of the counties we are studying are in the United States. This dependency is addressed in our research question. We do not claim that our results are applicable outside of the United States.

4.1.4 Case Count

We do not know how many undocumented or asymptomatic cases of COVID exist. There is a possibility that cases are under-reported in some counties, and overreported in others. However, the NYT cases repository is an academically accepted source of COVID-19 case data, and we trust that it is the most reliable (although not perfect) source of information.

4.2 No Perfect Collinearity

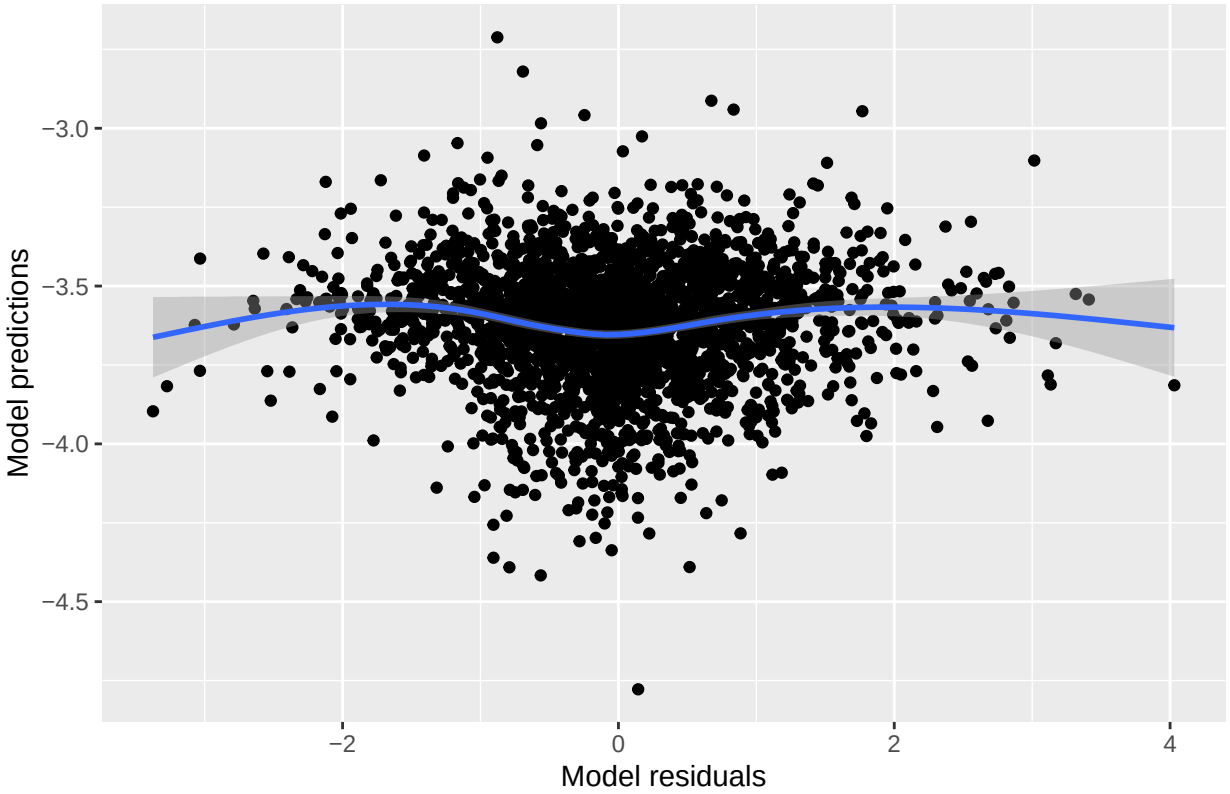
None of the correlations between variables suggest that there is perfect or near-perfect collinearity between variables. The highest correlation value between any two variables is -0.38425538 between `long_pp` and `urban`. In addition, we can see no intuitive reason to suspect that these variables are collinear.

4.3 Linear Conditional Expectation

Given that our model contains multiple variables, it can be hard to determine if we are actually describing a linear relationship between the dependent variable and the independent variables. However, by plotting the residuals of the model vs the predictions of the model, we can reach a conclusion on the linear relationship of the overall model to the dependent variable.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

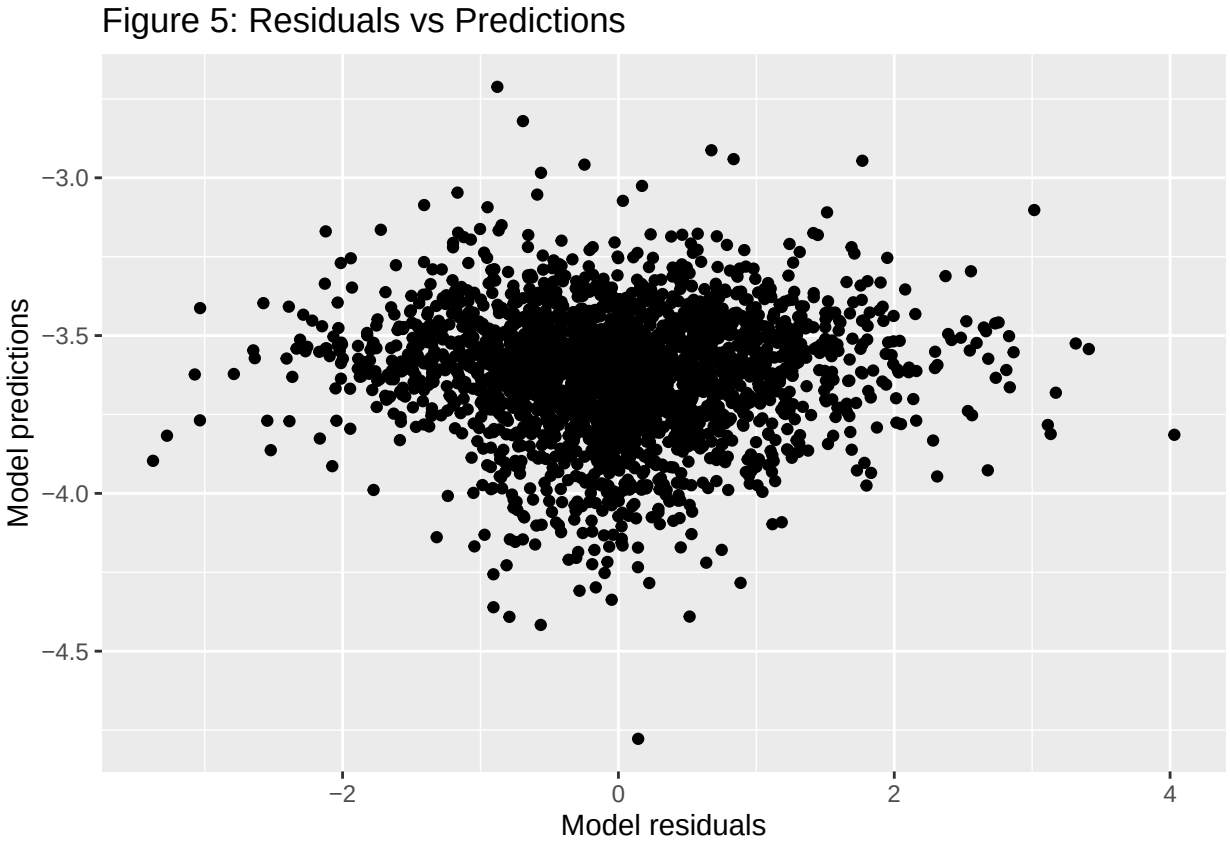
Figure 4: Residuals vs Predictions



Plotting the residuals and the predictions we can see some unexpected symmetry on either side of 0 on the x (residual) axis. Why the means of the predictions seem to increase on either side of 0, then dip at the extremities is unclear. This might suggest that case rates do not perfectly follow a logarithmic curve. Further research into this topic could yield interesting results.

However, for this study, the symmetry in the figure above is not indicative of a failure in our modeling methods, rather, they suggest that our model could benefit from additional predictors that will become evident as this pandemic is studied in greater detail.

4.4 Homoskedasticity

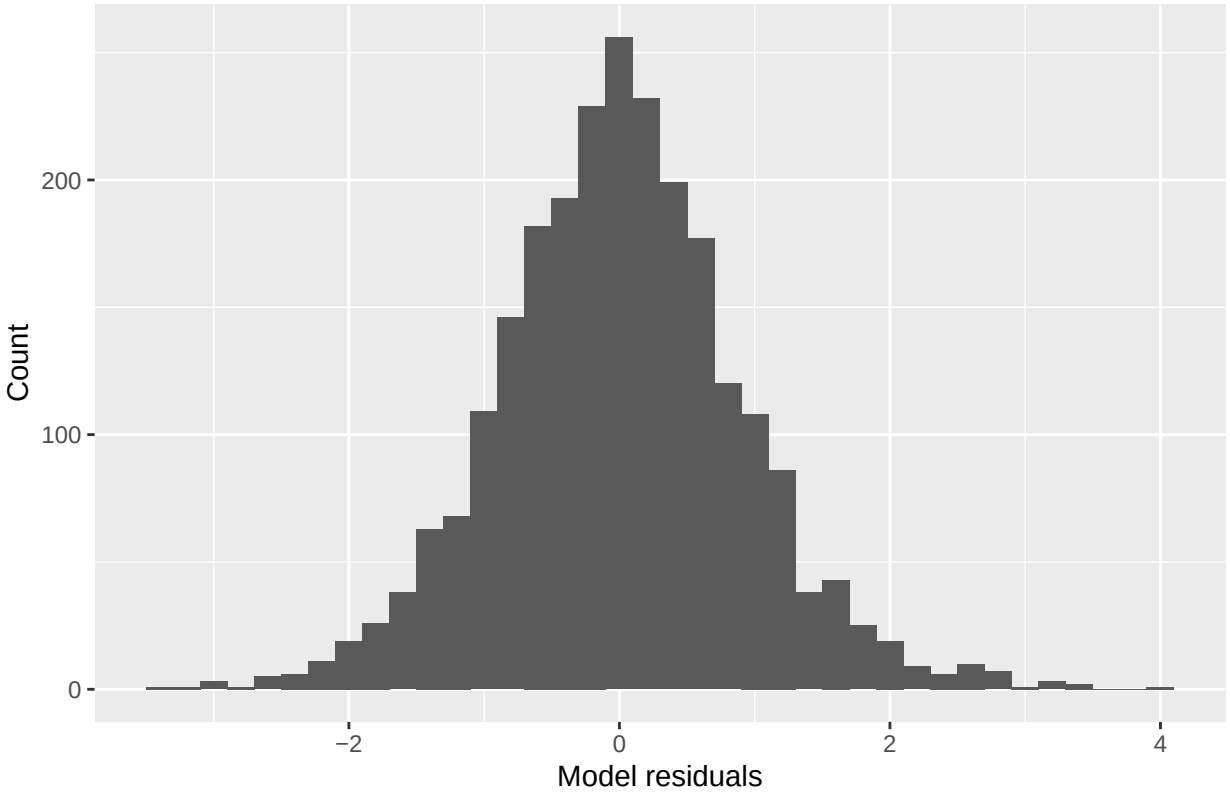


The figure above does not suggest any reason to believe that our data is not homoskedastic. There is a greater *range* in data near the center, but there are also far, far more data points near the center. This will therefore counteract any non-uniformities in variance.

4.5 Normal Distribution of Residuals

A histogram of the residuals of our model suggests that they fit a normal distribution well.

Figure 6: Distribution of model residuals



5. Discussion of Omitted Variables

Two omitted variables are worth mentioning.

5.1 Destination

For privacy reasons, the destination of the trips data is anonymized. We therefore cannot account for the obvious dependency in the proximity of counties. However, we are not concerned that this omitted variable could invalidate the findings of this study.

5.1.1 Variable of Interest and Causal Tree Graph

Ideally, we would have another variable, `visitations` that records the number of trips originating from outside of a given county and ending in that county on a given day. This would allow us to account not only for the origin of travelers but also their destinations. `Visitations` would not affect our current trips data. Rather, it would give us a better understanding of the effects of all travel in the county.

From a causal perspective we do not believe that visitors would impact case counts, but none of the dependent variables used in this study.

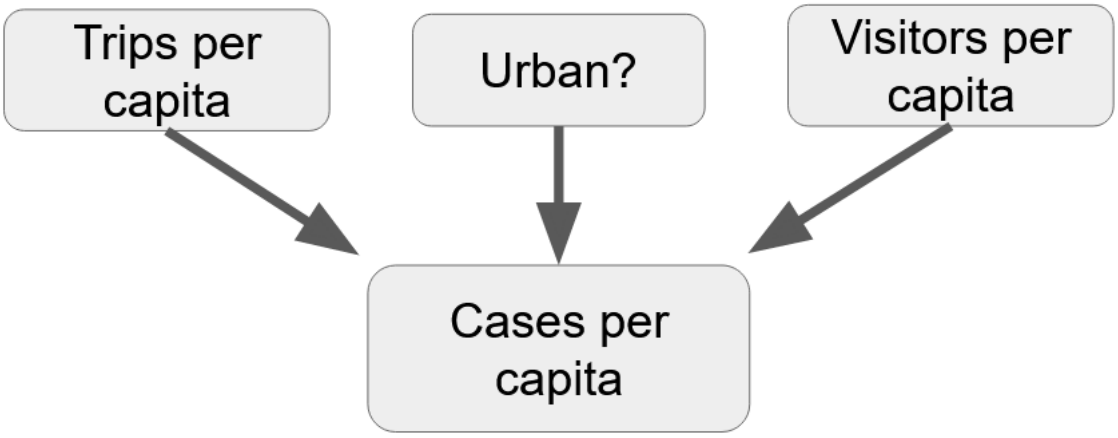


Figure 7: Destination dependency graph

5.1.2 Discussion of Direction of Bias

Because trips originating in a county and trips originating outside of a county are not dependent on each other, we do not expect that this lack variable would significantly impact any of the variables in our study. Rather, it would explain a portion of the variance that our model currently cannot account for. Therefore, while this variable would be useful in the bigger picture, it does not impact the conclusions drawn in this study.

5.2 Vaccinations

Preliminary research suggests that vaccination rates are not constant across counties. We chose not to include vaccination data for two reasons:

1) The efficacy of the vaccine, while being studied rigorously, is still not entirely known. Each vaccine brand has a different efficacy, and each county received differing amounts of each brand. On top of that, new strains of COVID-19 show varying resistance to the vaccine and varying transmissibility in infected, vaccinated people. Including vaccinations in our study might have added variance that we cannot explain given our limited knowledge of the vaccine.
2) The CDC data on vaccine rates is incomplete, and privacy laws across states affect the reporting of vaccine administrations. For example, the CDC has no data on vaccinations in Texas. Excluding the entire state of Texas from this study because of a lack of information on a control variable added a dependency to sample selection that we believe would not have been justifiable, given the variations in vaccine efficacy discussed in 1).

5.2.1 Variable of Interest and Causal Tree Graph

If we were to add vaccination rates to our study we would normalize the data to vaccinations per capita. Vaccinations would affect not only case counts but also trips.

Populations with high vaccination rates are less likely to sustain a COVID outbreak because vaccinated people are less likely to transmit the disease while going about their daily activities. Therefore, cases are expected to be lower where the per capita vaccination rate is high.

Vaccinated individuals also feel a sense of security upon taking the vaccine, because they are less likely to get sick and if they do get sick they are less likely to be hospitalized. This sense of security intuitively leads us to believe that vaccinated individuals will be more likely to travel.

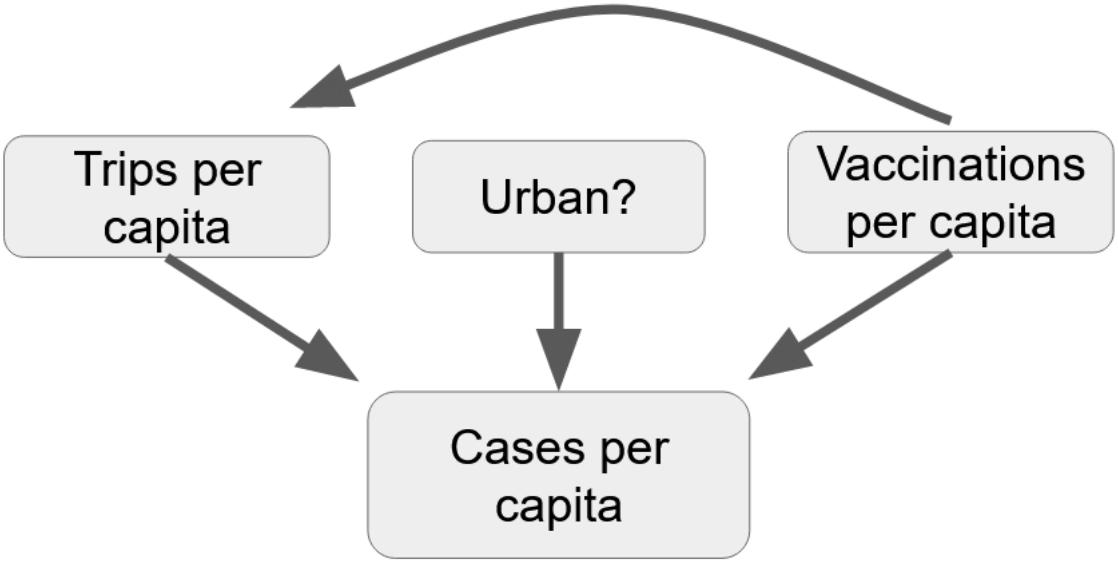


Figure 8: Vaccine dependency graph

5.2.2 Discussion of Direction of Bias

Vaccinations are expected to have a negative impact on case counts, but a positive impact on trip counts. Therefore, we cannot conclude whether this will bias our results toward zero or away from zero.

If the effect of vaccines on case count is much stronger than the effect of vaccines on trips, we can expect vaccinations per capita to bias out trips coefficients toward zero. Trip counts will be much less significant in modeling future cases if the vaccine becomes the dominant predictor of case county.

On the other hand, if the effect of vaccines on travel is much stronger than the effect of vaccines on case counts, we expect that vaccines per capita would bias our trips coefficients away from zero. The actual weight of trips on COVID cases would be much higher if vaccinated individuals feel emboldened to travel more but still spread the disease.

The fact of the matter is that we simply do not have enough information on the vaccines yet to conclude how it would affect the significance of trips counts in predicting cases. From the perspective of our study, our results would be most valid if the two opposing effects, increased trips and decreased cases, balanced each other out and the significance of our models did not change.

6. Conclusion

Overall, model 2 does the best job of describing the relationships of our underlying data set. While it does not significantly increase the R^2 value over the limited model, it helps us address the question of what kind of travel contributed more to changes in new COVID cases between weeks 22 and 23 of 2021. The limited model allowed us to establish a baseline and observe if there was a strong correlation between long trips and case counts. Then, by including both the short and long trips as variables in our model, we could compare the coefficients and interpret their relative correlation to case counts. More importantly, in model 2, we implement more independent variables while not sacrificing much statistical significance within each one.

This being said, it's important to recognize the lack of practical significance within our models. We again draw attention to the fact that standardizing our variables by population seemed to drastically decrease the correlation between the variables of interest- specifically, the number of trips taken by county residents

versus the new COVID cases one week later. On one hand, if we had not standardized these variables, the models of interest would represent a trivial pattern (i.e., more highly populated counties inherently have more trips and more COVID cases, while less populated ones have fewer trips and fewer COVID cases).

Standardizing the variables by population, therefore, seemed like the right approach; however, the R^2 values suggest that our selected independent variables do not explain very much of the variation within our dependent one. It's difficult to imagine that more vacations and increased amounts of travel do *not* have a large effect on COVID cases. In other words, we feel that there *should* be a larger positive correlation between trip counts and case counts. We believe our low R^2 values are representative of how we operationalized our data set rather than being indicative of the idea that there is no correlation between trips and COVID cases.

Specifically, we do not know what the underlying trends are within the travel data. Recall that we compartmentalized travel into three specific groups- short, mid, and long-range trips. Taking the long trips as an example, we cannot guarantee what behavior characterized this kind of travel. The practical significance of our models relies on the fact that the `long_pp` variable represents trips in which people are increasing interaction with people who are not in their household. This is a big assumption that wasn't explicitly stated in our data source.