Causal Relationship Study: COVID-19 Caseload and June 2021 Vacation Travel in Texas

Jonas Degnan, Autumn Rains, Lucy Wu

8/5/2021

Contents

1 Introduction

Summer vacations in the United States are synonymous with sunny, leisurely days spent with loved ones. As Ella Fitzgerald so aptly sings, "Summertime, and the living is easy." However, travel in the recent peak-vacation months has certainly deviated from this idealistic frame of mind with the onset of the SARS-CoV-2 (COVID-19) pandemic in early 2020. The global travel industry has experienced sharp declines in Revenue. Specifically in the United States, "travel spending totaled a mere $679 billion in 2020, an unprecedented 42% annual decline (nearly $500 billion) from 2019," according to analysis from the U.S. Travel Association[1]. Our hypothesis as to why this decline has occurred is due directly to the volume of COVID-19 number of cases. As virus infection rates oscillate regionally in the United States, our team of data scientists will look to answer the following primary research question using linear regression statistical techniques:

Do COVID-19 cases impact vacation travel trips for Texas for June 2021?

Our primary hypothesis is that there is a causal relationship between the number of COVID-19 cases and the number of vacation trips; when COVID-19 cases increase, the number of vacation travel trips decrease. To investigate, several demographic, mobility, and COVID-19 data sources will be analyzed. Based on the data, our team will also investigate the following secondary question:

Does the median income, the population density, and/or the age by county type (rural or urban) impact vacation travel for Texas residents in June 2021?

Our team also theorizes that individuals at higher risk of COVID-19 (e.g. older populations) infection would travel less than younger populations. For the analysis, we will create three models to understand if causal relationships exist between the outcome variable, the number of long distance leisure trips, and the control variables for each Texas county: income, population density, age, and COVID-19 cases . We will use an indicator variable, rural counties, to understand performance of our models for both rural or urban counties subsets. By creating causal models with these variables, we intend to answer our research questions to provide insight about how vacation travel has been impacted so that interested stakeholders can gain an understanding of how to improve the revenue declines in the travel industry in the near future.

2 Data Description

Operationalization of our research question required data from four sources: the New York Times (NYT)' *Coronavirus (COVID-19) Data in the United States* dataset[2], the United States Census Bureau's 2010 decennial survey[3], the Texas Association of Counties' *County Information Program*[4] database, and the United States Bureau of Transportation Statistics' *Trips by Distance*[5] dataset.

The NYT has maintained a robust public dataset of county-level, time-series COVID-19 data on GitHub and holds records dating back to the beginning of COVID-19. The dataset aggregates state and municipal government health department reports of cumulative cases and deaths for each US county daily from COVID-19. NYT data does not match deaths or cases for patients to their home state (e.g., vacationers), nor does it identify if a case occurred in one county, but a subsequent death occurred in another. Therefore, there is a likely mismatch between this data and those reported by individual states and counties.

[1] https://www.ustravel.org/toolkit/covid-19-travel-industry-research

[2] The New York Times. (2021, 08 04). Coronavirus (COVID-19) Data in the United States. https://github.com/nytimes/COVID-19-data. Retrieved 07 21, 2021, from https://raw.githubusercontent.com/nytimes/COVID-19-data/master/us-counties.csv

[3] United States Census Bureau. (2011, 11 22). Decennial Census, 2010. Explore Census Data. Retrieved 07 21, 2021, from https://data.census.gov/

[4] Texas Association of Counties. (2021, 08 04). QueriesCIP. County Information Program. Retrieved 08 02, 2021, from https://imis.county.org/iMIS/CountyInformationProgram/

[5] Department of Transportation. (2021, 08 02). Trips by Distance. Bureau of Transportation Statistics. Retrieved 07 19, 2021, from https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

For our investigation, we selected on cases, aggregated by county, in the month of June 2021 to mitigate time-series effects on modeling. The death count measurements were not selected because of the high correlation with the case count measurements: deaths counts are a subset of case counts.

The Texas Association of Counties proved to be a rich source of demographic and economic data for our target population. This database aggregates national and state data sources (e.g., Census Bureau, Texas Workforce Commission, United States Bureau of Labor Statistics, United States Bureau of Economic Analysis, and the Texas Demographic Center). The lineage of the TAC data was challenging to individually trace. All of the data originated from primary sources at the Federal or State level, but there was no readily available documentation to map columnar data back to its primary source. We selected four measurements to support our causal theory: county density, median household income, and populations aged over 65 years and under 17 years. The economic measurement of median income should have a significant role in an individual's ability to take vacations. The county population served as a control for the economic variables and density was hypothesized to encourage longer trips to more open spaces.

The decennial survey conducted by the Census results in an incredibly rich national database of economic, geographic, demographic, and social measures. While the vast database has many of the measurements useful to support our causal theory, as the TAC database queries held a majority of these, the decennial survey did proffer a means to generate an indicator variable identifying rural or urban counties.

Our final final data source was produced by the US Department of Transportation's Bureau of Transportation Statistics (USBTS). These data are experimental *estimates* of mobility gathered from aggregated mobile device data. USBTS states that these data are experimental, and but significant effort was made to ensure proper population weighting, and data multi-sourcing (to reduce variance in spacial and temporal measurements common from a single source) were implemented before any of the statistics were calculated. Trips are defined as a longer than 10 minute stop at a location that is not home. Trips are not distiguish by mode of transporation (i.e., air, train, bus, car, or bike). For our analysis, we were interested in only trips longer than 100 miles in Texas in June of 2021 as this best operationalized the concept of vacation travel[6].

USBTS conducted a survey[7] that characaterized long distance travel as trips over 50 miles from home. The distriubtion of trips in that survey included leisure, business, and commuter travel. To operationalize out outcome variable more concretely, we focused the dataset on trips over 100 miles to reduced noise from non-leisure travel[8].

Source	Description	Variable Name
Trips by Distance Bureau of Transportation Statistic U.S. Dept. of Transportation	Number of trips taken (>100 miles)	long_trip
Coronavirus (COVID-19) Data in the United States The New York Times	Total cases	cases
2010 Decennial Census U.S. Census Bureau	Urban/rural county indicator	is_rural
County Information Program Texas Association of Counties	Count median income County population density Population <17 Population >65	median_income density pct_lt17 pct_gt65

Figure 1: Operationalization Summary for Texas in June 2021

[6]https://www.bts.gov/statistical-products/surveys/national-household-travel-survey-long-distance-travel-quick-facts
[7]https://www.bts.gov/statistical-products/surveys/national-household-travel-survey-long-distance-travel-quick-facts
[8]https://www.brookings.edu/wp-content/uploads/2016/07/Srvy_JobsProximity.pdf

2.1 Exploratory Data Analysis

Table 1 below displays some information about the mean, median, minimum and maximum values for each variable. Figure 3 below depicts the distribution and correlation of each variable in our dataset by county type (where '0' indicates an urban county). From Figure 3, it can be observed that some of the variables are not distributed normally, are skewed, and thus require transformations. More specifically, these variables are: Total Population, Total Cases, Population Density. The outcome variable, `long_trips`, will also require the same transformation. Upon further analysis, logarithmic transformations were chosen to normalize the distribution and minimize skewness.

Below are histograms of variables that did not follow normal distribution and required transformation. We applied a logarithmic transform to each of these variables, which can be seen in the collective Figure 2.

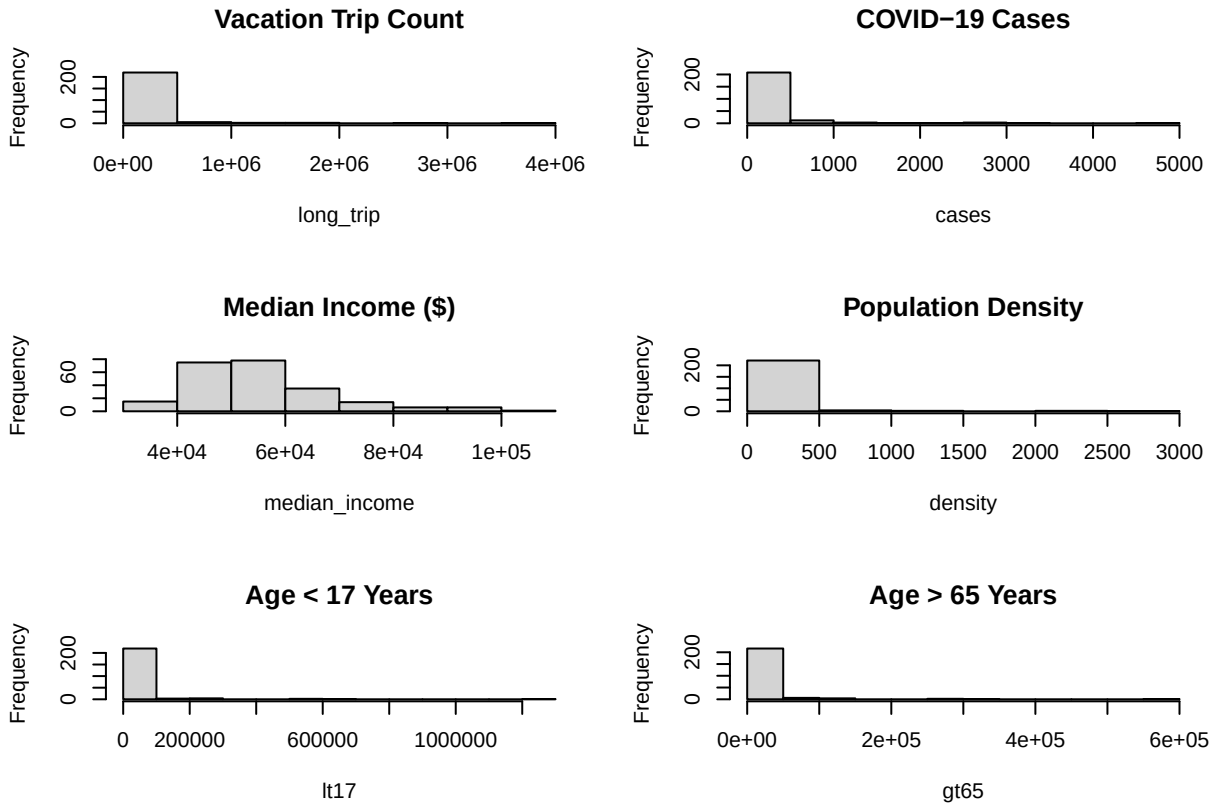

Figure 2: Distribution of raw variable data

Upon logarithmic transformation, below is a condensed view of the statistics of each variable:

2.2 Accounting Tables

To demonstrate the operationalization of the data, Table 2 below display the original value counts from the dataset as well as how many values were removed as a part of filtering the dataset to remove all states except Texas. For the analysis, 230 observations were analyzed.

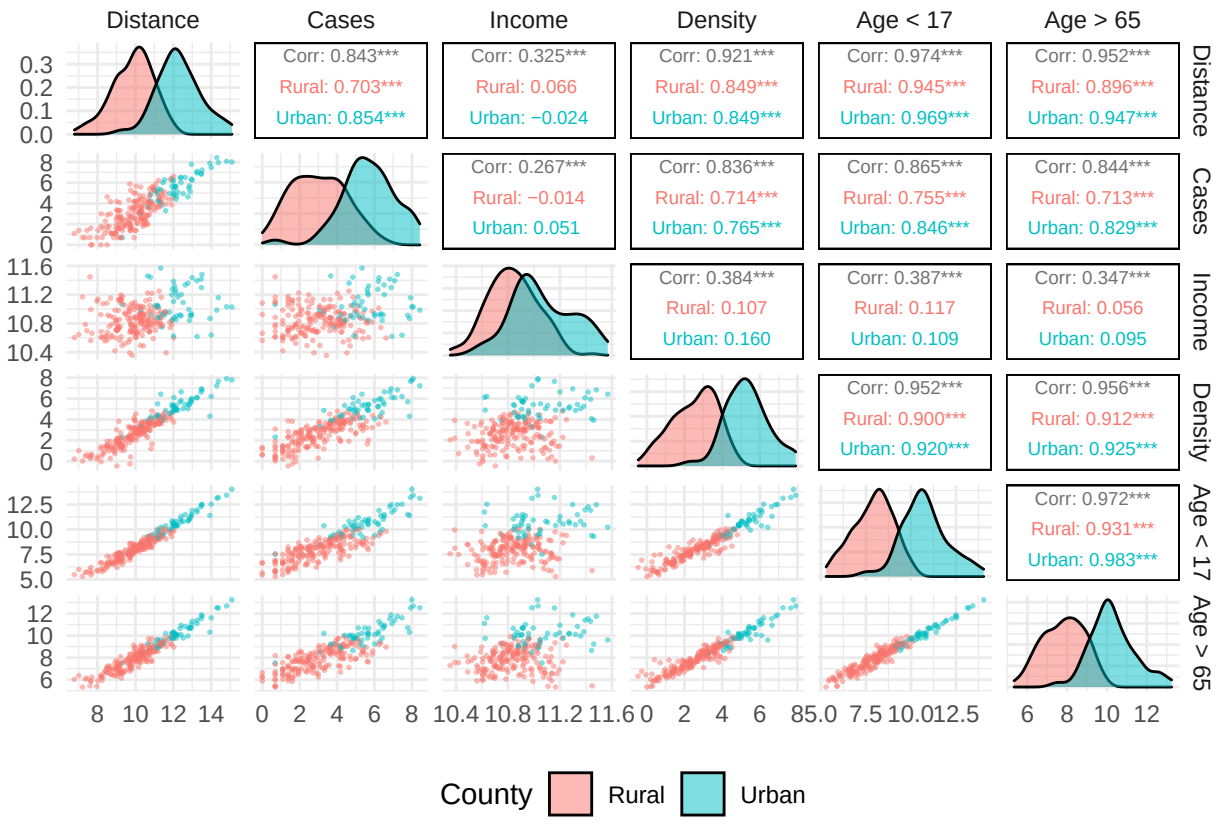

Figure 3: Variable Characterstics

Table 1: Variable Summary Statistics

	Rural	Urban	Total
	(N=178)	(N=52)	(N=230)
Vacation Trip Count			
Mean (SD)	31200 (31500)	437000 (666000)	123000 (359000)
Median [Min, Max]	21400 [878, 166000]	187000 [11100, 3570000]	29800 [878, 3570000]
Population Density			
Mean (SD)	22.4 (23.0)	399 (575)	107 (315)
Median [Min, Max]	15.1 [0.630, 122]	201 [10.0, 2720]	23.0 [0.630, 2720]
COVID-19 Cases			
Mean (SD)	55.4 (100)	695 (981)	200 (542)
Median [Min, Max]	18.5 [0, 752]	284 [1.00, 4580]	35.5 [0, 4580]
Median Income ($)			
Mean (SD)	51800 (10100)	66000 (16100)	55000 (13100)
Median [Min, Max]	50600 [31400, 93800]	61700 [40900, 106000]	53100 [31400, 106000]
Age < 17 Years			
Mean (SD)	4580 (4830)	116000 (212000)	29700 (110000)
Median [Min, Max]	3050 [198, 23900]	43400 [1840, 1250000]	4220 [198, 1250000]
Age > 65 Years			
Mean (SD)	4020 (3940)	59900 (98900)	16700 (52300)
Median [Min, Max]	2610 [207, 20400]	25300 [1850, 567000]	4010 [207, 567000]

Table 2: Accounting Table

Step	Number of Samples	Samples Removed	Reason
1	230	0	Original Dataset
2	230	0	Logarithmic Transformations

2.3 Additional Modifications to the Data:

Upon final review of the dataset, there are no additional changes required to prepare the dataset for data exploration and subsequent model creation. Outliers or incorrect data are not observed upon final inspection. The Accounting Table captures the changes made through cleaning, filtering, and transforming as indicated previously. With a sample size over 100, a large-sample linear model can be built.

3 A Model Building Process

To investigate the causal relationships between distance traveled and variables within the dataset, a linear regression was performed in R Studio for three models. The first model includes only our the key variables associated with our primary hypothesis: that COVID-19 cases decrease vacation trip counts. In the second model, we added additional county econometric and demographic characteristics. In the final model, we added additional demographic variables for age groups.

3.1 Model Equations

Model 1:

$(\log(\text{Long Trip Count})) = \beta_0 + \beta_1(\log(\text{COVID-19 Cases})) + \beta_2(\text{Rural County}))$

Model 2:

$(\log(\text{Long Trip Count})) = \beta_0 + \beta_1(\log(\text{COVID-19 Cases})) + \beta_2(\text{Rural County})) + \beta_3(\log(\text{Median Income})) + \beta_4(\log(\text{County Density}))$

Model 3:

$(\log(\text{Long Trip Count})) = \beta_0 + \beta_1(\log(\text{COVID-19 Cases})) + \beta_2(\text{Rural County})) + \beta_3(\log(\text{Median Income})) + \beta_4(\log(\text{County Density})) + \beta_5(\log(\text{Population under 17})) + \beta_6(\log(\text{Population over 65}))$

4 Linear Regression Results

The results from the linear regression analysis can be seen below for each model:

```
##
## Results
## ================================================================================================
##                                              Dependent variable:
##                          ------------------------------------------------------------------------
##                                                   log(long_trip)
##                                (1)                      (2)                      (3)
## ------------------------------------------------------------------------------------------------
## log(cases + 1)              0.546***                 0.188***                 -0.007
##                             (0.033)                  (0.036)                  (0.024)
##
## is_rural                   -0.974***                 -0.252**                  0.097
##                             (0.147)                  (0.126)                  (0.079)
##
## log(median_income + 1)                               -0.213                  -0.342***
##                                                      (0.182)                  (0.112)
##
## log(density)                                         0.620***                 -0.076
##                                                      (0.045)                  (0.048)
##
## log(lt17)                                                                     0.887***
##                                                                               (0.065)
##
## log(gt65)                                                                     0.121*
##                                                                               (0.071)
##
## Constant                    9.197***                10.295***                5.724***
##                             (0.214)                  (2.005)                  (1.278)
##
## ------------------------------------------------------------------------------------------------
## Observations                  230                      230                      230
## R2                           0.758                    0.868                    0.952
## Adjusted R2                  0.755                    0.866                    0.951
## Residual Std. Error    0.740 (df = 227)         0.548 (df = 225)         0.331 (df = 223)
## F Statistic           354.725*** (df = 2; 227) 371.006*** (df = 4; 225) 741.836*** (df = 6; 223)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

4.1 Results Discussion

From this output table, we can discern quite a bit of information.

In `Model 1`, all of our variables are statistically significant, though the `cases` variables is in the opposite direction as expected: our primary hypothesis expected an inverse relationship between COVID-19 case counts. The `is_rural` variable coefficient directionality aligns our primary hypothesis that living in a rural county increases frequency of vacation trips.

In `Model 2`, when county density is factored in, the COVID-19 case counts have a direct relationship on vacation trip frequency. The relationship between median income and vacation trip frequency is not statistically significant.

In `Model 3` with age factored in, the statistically significant relationships profile changes. Only age demographics and income now have significance on vacation trip frequency.

Each successive models improves R^2 and residual standard errors, indicating that variance estimates improve with each model resulting in improve model performance.

4.2 Coefficient Interpretation

In `Model 1`, for every percent increase in cases, there is a 0.0055% increase in trip frequency and being in a rural county increases trip frequency by 62.2%.

In `Model 2`, for every percent increase in county density, there is a 0.0062% increase in trip frequency. This provides evidence to support the causal theory that more dense counties would have high rates of vacation travel. The strength of the causal relationship between the case count and `is_rural` indicator has decreased with this new covariate.

In `Model 3`, the causal relationships between COVID-19 cases and county density are no longer statistically significant. The prevelance of COVID-19 cases no longer has a causal relationship with trips. This can be interpreted as Texans are experiencing pandemic fatigue and may no longer be influenced by health data when making travel plans. A county being rural or dense no longer a causal relationships with vacation trip counts and can be interpreted as age demographics have a accurately explain vacation trip counts.

Median income's relationship is now statistically significant, as well as both age quantiles. Median income has a negative causal relationship (0.003%) with vacation counts. This evidence does not support our hypothesis that higher median incomes result in an increase in the ability to take vacation trips. This is explored in our Classical Linear Model (CLM) assumptions related to collinearity. The population under age 17 now has the strongest causal relationship (0.008%) with vacation trip counts and due to the time frame of our analysis, this makes sense because June is the start of summer break when many families take vacations together. The population over age 65 is typically retired and more available to take vacations and has a positive (0.001%) causal relationship.

5 Model Limitations

Below is an analysis of the Classin Linear Model assumptions for `Model 2` and `Model 3`:

1. **IID Sampling:** There isn't a direct database check for this unfortunately. Looking into how the data was generated, we believe this is IID generally. The data was gathered from independent surveys within the Texas population for all models generated across all counties. There may be instances of sample clustering given that individuals in counties may have influenced individuals in neighboring counties as well as being subject to statewide restrictions on implementation of COVID transmission

mitigration strategies (e.g., mask mandates)[9]. However, the sample size for the analysis is sufficient to minimize this possible impact.

2. **No Perfect Colinearity:** This can be assessed through examination of the variance inflation factor (VIF). The results for the VIF analysis are seen in the table below. Running a VIF in R for `Model 2` and `Model 3`, we see the VIF scores are relatively low (below 5) for the predictor variables in `Model 2`. However, upon review of `Model 3`, this is not the case. With VIF factors greater than 5, the both age variables would be candidates for removal outright. When reviewing the population density variable in `Model 3`, it remains significant for `Model 2`. Additionally, when viewing the variable relationships between the two age variables and population in Figure 3, strong linear relationships can be viewed which would also indicate almost near perfect colinearity and thus should be removed from the model.

Table 3: Model 2 Covariate Variance Inflation Factors

Variable	VIF
log(cases + 1)	3.38
is_rural	2.14
log(median_income + 1)	1.27
log(density)	4.32

Table 4: Model 3 Covariate Variance Inflation Factors

Variable	VIF
log(cases + 1)	4.10
is_rural	2.26
log(median_income + 1)	1.31
log(density)	13.40
log(lt17)	24.48
log(gt65)	22.94

3. **Linear Conditional Expectation:** For this assumption, the residuals for linear relationships must be evaluated. By plotting the residuals, as seen below in the first scatter plot for both `Model 2` and `Model 3` ('Residual vs Fitted'). We generally see a linear line as we move across predictor values with the residuals with no strong pattern. By transforming the variables as indicated previously logarithmically, we are able to see this linear relationship across all predictor values.

4. **Homoskedastic Errors:** This assumption is examined by looking for constant error variance across the entire range of the x's (homoskedastic errors across the range of the x's). By plotting fitted values vs. the square root of the standardized residuals for the `Model 2` and `Model 3` ('Scale-Location' figure), there is an even or equal spread dispersion across the x values. Therefore `Model 2` and `Model 3` both meet this condition when including the transformed variables.

5. **Normally Distributed Errors:** When investigating a normal distribution of the residuals, the Q-Q plot is helpful to investigate this assumption. For `Model 2` and `Model 3`, a linear relationship between theoretical quantities and the standardized residuals ('Normal Q-Q') can be observed. In conclusion, we meet this assumption with both `Model 2` and `Model 3`.

[9]https://gov.texas.gov/news/post/governor-abbott-issues-executive-order-prohibiting-government-entities-from-mandating-masks

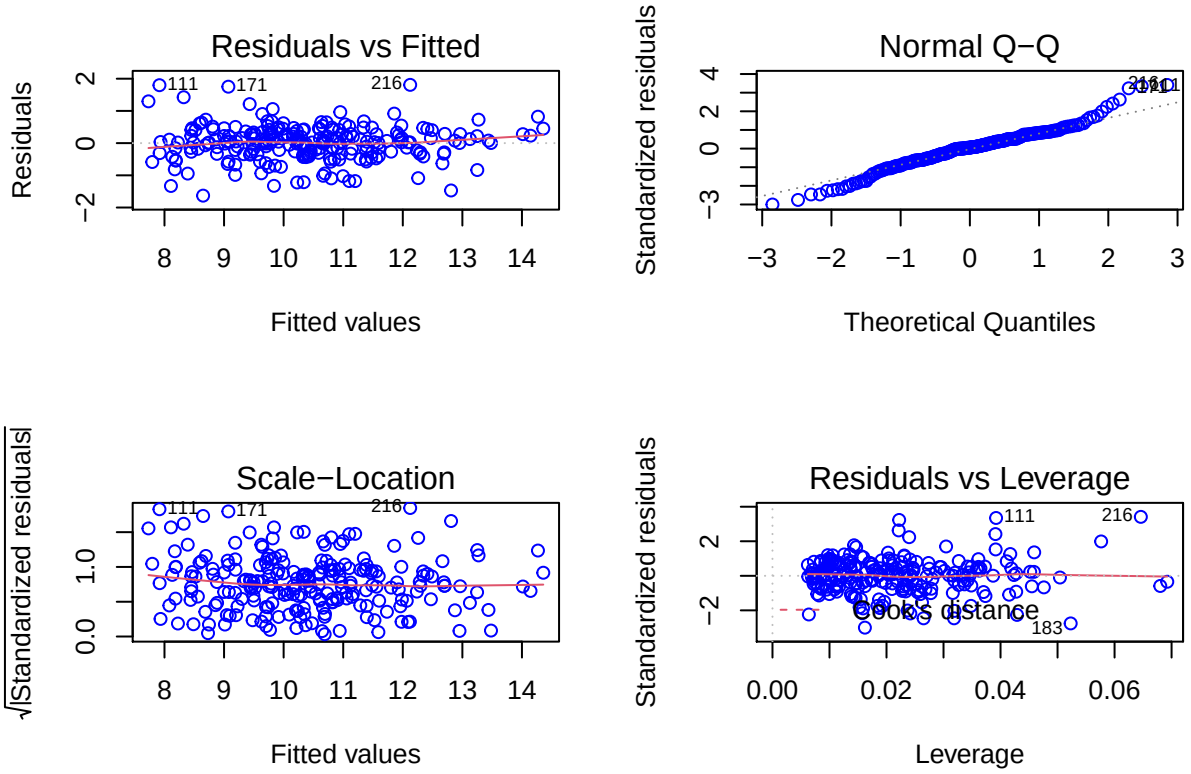

Figure 4: Model 2 Q-Q

6 Omitted Variable

6.1 Vehicle Ownership

One example of an omitted variable that may cause bias in our model is that of vehicle ownership among the population. It can certainly be expected that individuals that own a vehicle would be able to more easily travel short or long distances more frequently with relative ease. If an individual owned a vehicle, this would influence our model by driving the bias further from zero. From the data available, there would no immediate proxy for this omitted variable.

6.2 Fuel Prices

Another example of an omitted variable that may also cause bias in our model is fuel prices. Whether this is jet fuel or gasoline, fluctuations in this commodity would certainly impact travel costs. If fuel prices increase, it is generally observed that costs for travel increase. The opposite is also generally true. In the instance of increased fuel prices, this would influence our model by driving the bias closer to zero as distance traveled would drop. Conversely, if fuel prices for the time period of the sample for the population were lower, the bias would would be farther from zero.

7 Conclusion

The goal of this analysis was to determine if there existed a causal relationship between distance traveled and COVID-19 infection rates for residents of the state of Texas. Based upon the linear regression analysis above, model 2 would indicate there may be a causal relationship between distance traveled (outcome variable) and the variables found within that model, though with all variables considered, our results do not indicate vacation and COVID-19 cases are necessarily related.

7.1 Next Steps

For future analysis, it is recommended that COVID-19 policy adherence data per county be incorporated into the model to increase the accuracy. While the Governor of Texas issued detailed protocols or mandates for the state of Texas, the degree to which each county (and thus the individual residents) complied would be useful data to have if such data were to exist. Additionally, incorporating vaccination rates could aid in development of a model with even higher accuracy. Generally, vaccinations could increase confidence among individuals interested in traveling, especially if they have been in long periods of quarantine or government mandated lockdowns. Last, it is highly recommended that any model generated in the future continue to use the latest data available given the oscillation in vaccination rates, government policy, and new or known variants of the COVID-19 virus.