Section_8_Team1_Lab2_Report

Jonas Degnan, Autumn Rains, Lucy Wu

8/5/2021

Contents

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 outbreak (COVID-19) in early 2020. The global travel industry has experienced sharp declines. Specifically in the United States (US), "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. As infection and vaccination rates oscillate regionally along with government policies for business operation restrictions in the United States, our team of data scientists will look to answer the following research question in more detail using linear regression techniques:

Do COVID-19 infection rates impact vacation travel distance for U.S. residents in the state of Texas? How does income, population density, and covid regulations in each county impact traveling?

We will analyze data pertaining to these variables for Texas counties and create three models to understand if causal relationships exist between the number of long distance trips, population income, population density (rural or urban counties). Our hypothesis is that an there is a causal relationship between """ [summer travel]

–>Need to explain distance of travel with bs TRANSPORT data[1] survey –>Need to add more to causal theory to justify variable selection

Causal theory needed to identify why these particular variables below were chosen

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 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.

For our investigation, we selected on cases and deaths in the month of June 2021, to best operationalize summer. The case and death counts were aggregated by county to mitigate time-series affects on modeling. The death counts were identified as being directly related to the COVID-19 infection and not from other instances (e.g., homicide), or unrelated comorbidities (e.g., cancer).

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

[1] https://www.bts.gov/statistical-products/surveys/national-household-travel-survey-long-distance-travel-quick-facts

[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

available documentation to map columnar data back to its primary source. We selected five measurements to support our causal theory: county population and density, median household income, number of individuals in poverty, and the unemployment rate. The economic measurements of poverty and employment were hypothesized to have an inverse relationship with enabling people to take long vacation trips, where median income should have a more direct relationship. 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. This variable was important since we hypothesized that rural counties may experience higher long trip counts due to their geographies.

Our final course were data produced by the US Department of Transportation's Bureau of Transportation Statistics (USBTS). These data are experimental *estimates* of mobility gathrered 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 analyis, we were interested in only trips longer than 100 miles in Texas in June of 2021 as this best operationalized the concept of a vacation[6].

Exploratory Data Analysis

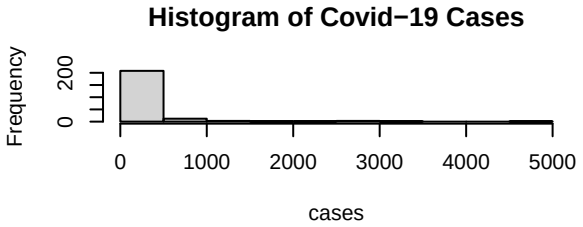
Table 1 below displays some information about the mean, median, minimum and maximum values for each variable. Figure 1 below depicts the distribution and correlation of each variable in our dataset by county type (where '0' indicates an urban county). From Figure 1, it can be observed that some of the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

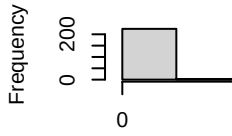
Below are histograms of variables that did not follow normal distribution and required transformation. Each

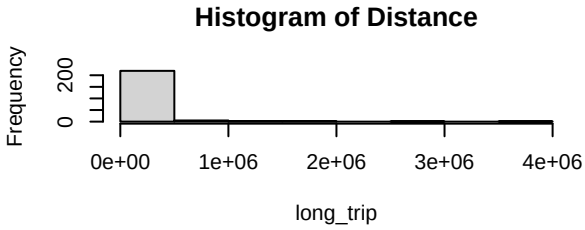
[6]https://www.bts.gov/statistical-products/surveys/national-household-travel-survey-long-distance-travel-quick-facts

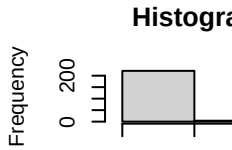
Table 1: Table 1: Texas County Variable Statistics

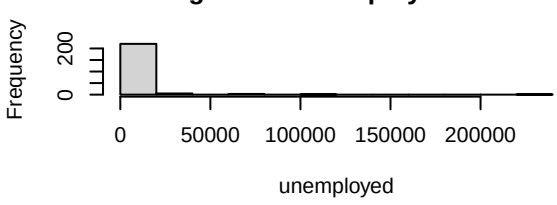
	Rural	Urban	Total
	(N=178)	(N=52)	(N=230)
County Population Density			
Mean (SD)	22.4 (23.0)	399 (575)	107 (315)
Median [Min, Max]	15.1 [0.630, 122]	201 [10.0, 2720]	23.0 [0.630, 2720]
County Covid-19 Cases			
Mean (SD)	55.4 (100)	695 (981)	200 (542)
Median [Min, Max]	18.5 [0, 752]	284 [1.00, 4580]	35.5 [0, 4580]
County Unemployment Rate			
Mean (SD)	933 (1110)	19000 (37100)	5030 (19100)
Median [Min, Max]	602 [31.0, 7560]	6020 [290, 236000]	809 [31.0, 236000]
County Median Income ($)			
Mean (SD)	51800 (10100)	66000 (16100)	55000 (13100)
Median [Min, Max]	50600 [31400, 93800]	61700 [40900, 106000]	53100 [31400, 106000]
County Age < 17 years (%)			
Mean (SD)	4580 (4830)	116000 (212000)	29700 (110000)
Median [Min, Max]	3050 [198, 23900]	43400 [1840, 1250000]	4220 [198, 1250000]
County Age > 65 years (%)			
Mean (SD)	4020 (3940)	59900 (98900)	16700 (52300)
Median [Min, Max]	2610 [207, 20400]	25300 [1850, 567000]	4010 [207, 567000]

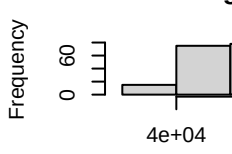












of these variables was logarithmically transformed.

Table 2: Accounting Table

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

Table 3: Table 2: Final Samples by County Type

County Type	Number of Samples
Rural	178
Urban	52

Figure 1: Variable Characterstics

Accounting Tables

To demonstrate the operationalization of the data, Tables 1 and 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 responses were analyzed as shown in Table 1. Table 2 depicts the county type breakdown for rural or urban classification as well.

Additional Changes 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 capture the changes made through cleaning and filtering as indicated previously. Table 2 above also displays the subset per county type for the overall dataset.

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 is limited to study only key variables of interest for measurement. The second model expands upon the first model to also include key explanatory variables and covariates that are hypothesized to improve performance of the previous model. The third model is all-inclusive of the variables available to the team at the time of data gathering. Details of each model analyzed can be found below.

$Model_1$: $f(\log(\text{Number of Long Trips})) = \beta_0 + \beta_1 g(\log(\text{Covid-19 Cases})) + \beta_2 h(\text{Rural County}) + \beta_3 h(\text{Median Income})$

$Model_2$: $f(\log(\text{Number of Long Trips})) = \beta_0 + \beta_1 g(\log(\text{Covid-19 Cases})) + \beta_2 h(\text{Rural County}) + \beta_3 h(\text{Median Income}) + \beta_4 h(\text{Unemployment Rate}) + \beta_5 h(\log(\text{Population Density})$

$Model_3$: $f(\log(\text{Number of Long Trips})) = \beta_0 + \beta_1 g(\log(\text{Covid-19 Cases})) + \beta_2 h(\text{Rural County}) + \beta_3 h(\text{Median Income}) + \beta_4 h(\text{Unemploymed}) + \beta_5 h(\log(\text{Population Density}) + \beta_6 h(\log(\text{Age} < 17) + \beta_7 h(\log(\text{Age} > 65) + \beta_7 h(\log(\text{In Poverty})$

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)               0.421***                0.156***                0.149***
##                              (0.036)                 (0.035)                 (0.035)
##
## log(deaths + 1)              0.502***                0.230***                0.255***
##                              (0.078)                 (0.071)                 (0.073)
##
## is_rural                    -0.455***                -0.093                  -0.080
##                              (0.162)                 (0.131)                 (0.131)
##
## log(median_income + 1)        0.272
##                              (0.223)
##
## median_income                                        -0.00000                -0.00000
##                                                     (0.00000)               (0.00000)
##
## unemployed                                           0.00000                 -0.00002
##                                                     (0.00000)               (0.00001)
##
```

```
## log(density)                                          0.561***                0.584***
##                                                       (0.046)                 (0.047)
##
## lt17                                                                          0.00001**
##                                                                               (0.00001)
##
## gt65                                                                          -0.00002*
##                                                                               (0.00001)
##
## Constant                          5.914**             8.146***                8.155***
##                                   (2.482)             (0.259)                 (0.262)
##
## ------------------------------------------------------------------------------------------------
## Observations                        230                 230                     230
## R2                                 0.796               0.879                   0.881
## Adjusted R2                        0.792               0.875                   0.876
## Residual Std. Error          0.683 (df = 225)    0.529 (df = 223)        0.526 (df = 221)
## F Statistic            219.132*** (df = 4; 225) 268.850*** (df = 6; 223) 204.039*** (df = 8; 221)
## ================================================================================================
## Note:                                                              *p<0.1; **p<0.05; ***p<0.01
```

Additionally, to indicate if any of the models provide a better fit to the dataset than a model that contains no independent variables, an F test was performed. The results for each model are shown below:

```
## Analysis of Variance Table
##
## Model 1: log(long_trip) ~ log(cases + 1) + log(deaths + 1) + is_rural +
##     log(median_income + 1)
## Model 2: log(long_trip) ~ log(cases + 1) + log(deaths + 1) + is_rural +
##     median_income + unemployed + log(density)
## Model 3: log(long_trip) ~ log(cases + 1) + log(deaths + 1) + is_rural +
##     median_income + unemployed + log(density) + lt17 + gt65
##   Res.Df     RSS Df Sum of Sq       F Pr(>F)
## 1    225 104.825
## 2    223  62.329  2    42.497 76.7351 <2e-16 ***
## 3    221  61.196  2     1.133  2.0455 0.1318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the F-test, we can see the the addition of the variables into model 2 and model 3 do contribute signficant information to our model based off of the derived p-values.

Last, the significance of individual regression coefficients was tested using robust standard errors in R studio for model 3. The results are shown below:

```
##
## t test of coefficients:
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)      8.1457e+00  2.8697e-01 28.3855 < 2.2e-16 ***
## log(cases + 1)   1.5606e-01  4.2200e-02  3.6981 0.0002736 ***
## log(deaths + 1)  2.3001e-01  6.6325e-02  3.4680 0.0006291 ***
## is_rural        -9.3138e-02  1.1985e-01 -0.7771 0.4378991
## median_income   -3.2818e-06  3.1318e-06 -1.0479 0.2958186
```

```
## unemployed       2.3798e-06  1.1426e-05  0.2083 0.8351921
## log(density)     5.6108e-01  6.6776e-02  8.4025 5.149e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the t-test, we can see that Covid-19 cases and median income are statistically significant. Additionally, the percentage of adults greather than 65, the population density, and the Covid-19 death variables are even more statistically significant.

Discussion

Model Limitations

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.

2. **No Perfect Colinearity:** This can be assessed through examination of the variance inflation factor (VIF). 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. Therefore, no predictor variables were removed from the model. The results for the VIF analysis are seen in the table below:

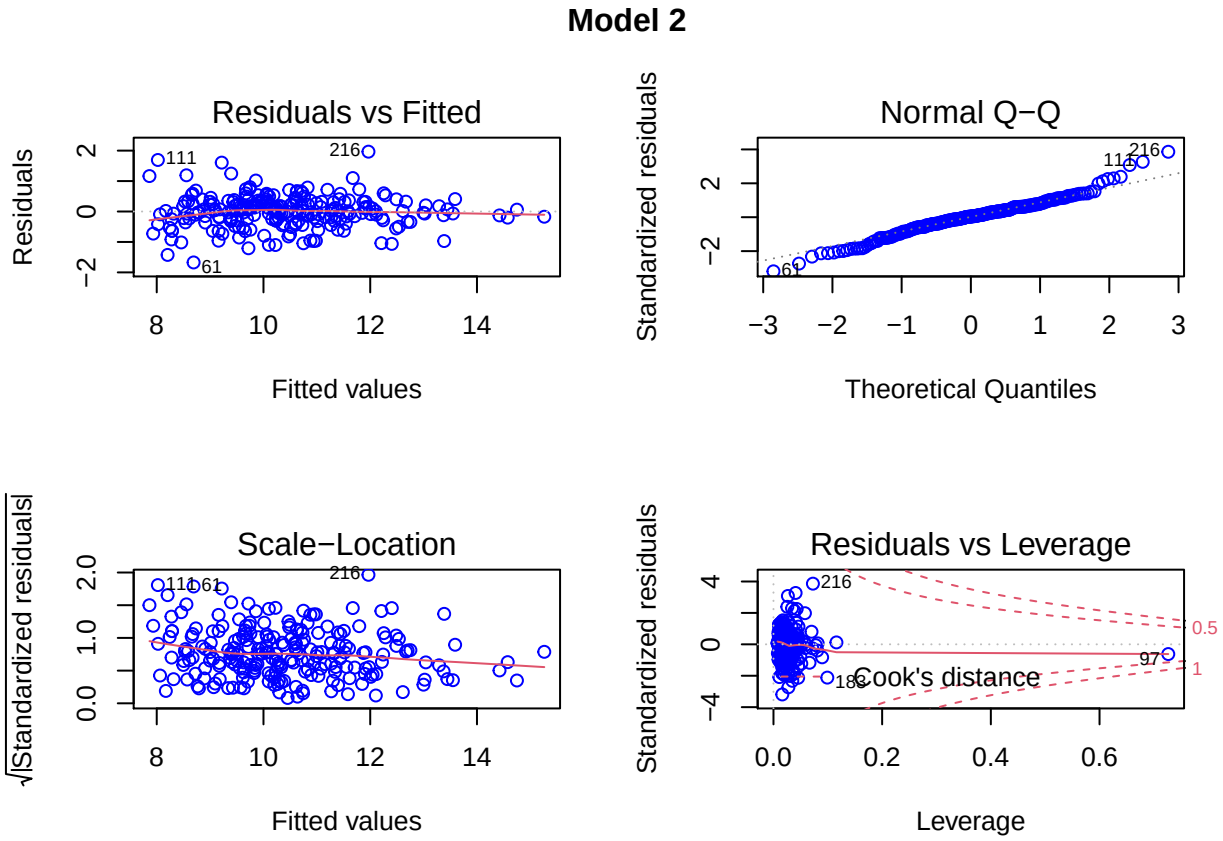
```
##  log(cases + 1) log(deaths + 1)        is_rural   median_income      unemployed
##        3.553546        3.778266        2.478667        1.301736        1.695943
##    log(density)
##        4.738804
```

```
##  log(cases + 1) log(deaths + 1)        is_rural   median_income      unemployed
##        3.588684        3.956202        2.486508        1.367379       52.455059
##    log(density)             lt17            gt65
##        5.136391      310.496444      187.629927
```

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 both meet this condition when including the transformed variables.

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



Omitted Variable

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.

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.

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.

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 Govenor 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.