STAT 310 Final Report

Devin Barry, Ayushi Joshi, Mila Lewis-Peinado, Sophia Maalouly

December 2022

1 Introduction

Illinois is a state located in the mid-western region of the United States. The state consists of many small towns and cities built on flat prairies, which are mainly located in central and south Illinois. Chicago, which is one of the United States' larges cities, is located in northern Illinois. Lake Michigan brings seasonal tourism, which contributes to more facilitation of travel. Illinois is the eighth highest ranked state when it comes to income inequality and the demographic of the population is made up of 69.79 percent white, 14.13 percent black, 5.58 percent Asian, and 5.59 percent other races. We were interested in looking at COVID-19 in Illinois because of the strong contrast between the rural areas and the greater Chicago area.

The first case of COVID-19 in Illinois was confirmed by health officials on January 24th, 2020. In February, cases began to increase. A majority of the initial cases occurred in Chicago, which is located in Cook county. Chicago is also the location of two busy international airports. We will see later that Cook county stands out in our data. On the ninth of March, Governor J.B. Pritzker declared a state of emergency and on the 17th, statewide school, restaurant, and bar closures began. Non-essential businesses were closed under the statewide "stay-at-home order". The end of the order was pushed from April 7th to the 30th, then to May 29th, and finally to June 26th. On June 26th, large gatherings were allowed and restaurants and bars began to open again with social distancing and mask-wearing guidelines in effect. At the beginning of the spread of COVID-19, testing was not readily available due to the lack of definitive COVID-19 tests, however, Illinois was one of the first states to be able to test for COVID-19. Asymptomatic testing became especially available in the Cook county area. In December of 2020, Illinois experienced a very high

number of deaths per week and the fifth highest number of confirmed cases in the United States.

2 Questions of Interest and Hypotheses

After doing background research about Illinois and its COVID-19 timeline we developed some questions to try and answer. The overarching interest we have is to find out what factors most heavily influenced the spread of the virus throughout the state, or more specifically which variables are the best at predicting COVID-19 deaths. This question helped us to determine which variables and data to collect as we want to try and capture the differences between the Illinois counties. We also have a couple more specific questions we want to answer which are to determine how big of a role unemployment plays in deaths and how the vaccine affected cases. We want to look at unemployment because Illinois ranks number 8 in the country for income inequality, so we think it is a unique characteristic of the state that could play an interesting role. As for the vaccine, Illinois was one of the first states to be able to test for COVID-19 so it we want to see how that affected cases.

We are expecting to see unemployment have a negative relationship with deaths, as richer areas have more resources to combat the virus and can afford to stay home and not work, lowering their risk of contracting the virus and dying. As for the effect of the vaccine, new cases should be lower after the vaccine was available.

3 Exploratory Data Analysis

We used maps to conduct part of our exploratory data analysis. We plotted maps by county of each of our explanatory variables, and also grouped all of our variables by region because Illinois has 102 counties and splitting them by regions made it easier to model data. When looking at the summary statistics, the mean and lower quartiles are rather low, but there is a really large range.

There is a large range because we have an outlier, Cook county which includes Chicago. Cook county is found to have the largest numbers when it comes to overall population variables (total population, white male, white female, total male, and total female). Cook county has one of the lowest percentages of white people and the counties that surround Cook county rank the in the

White Male, White Female, Total Male, Total Female, Total Population (Respectively)

```
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd   p0  p25   p50   p75    p100 hist
1 data                  0             1 50968. 184463. 1931 6831 13537 24645 1693430 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd   p0  p25   p50   p75    p100 hist
1 data                  0             1 52014. 188023. 2002 7219 13806 25404 1726849 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd   p0  p25   p50   p75    p100 hist
1 data                  0             1 64818. 269629. 2147 7085 14519 26582 2514314 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd   p0  p25   p50   p75    p100 hist
1 data                  0             1 67362. 286812. 2114 7343 14359 27106 2680361 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate    mean      sd   p0   p25   p50   p75    p100 hist
1 data                  0             1 132180. 556431. 4320 14665 29718 53873 5194675 ▇▁▁▁▁
```

Percent of the Population That Is White, Percent of the Population That Is Not White, People Per Housing, Unemployment Rate

```
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd   p0  p25  p50  p75 p100 hist
1 data                  0             1 92.6 7.76 61.9 89.3 95.9 98.2 99.1 ▁▁▁▃▇
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd    p0  p25  p50  p75 p100 hist
1 data                  0             1 7.45 7.76 0.884 1.82 4.11 10.7 38.1 ▇▂▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean    sd   p0  p25  p50  p75 p100 hist
1 data                  0             1 2.20 0.220 1.60 2.08 2.17 2.29 2.93 ▁▃▇▂▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean    sd  p0 p25 p50 p75 p100 hist
1 data                  0             1 4.25 0.762 2.4 3.8 4.1 4.8  6.3 ▂▇▃▁▁
```

Figure 1: Summary Statistics for Explanatory Variables

highest numbers for the variable "people per house".

After looking at our maps, we begun to look at histograms for our variables. The histograms for the regions vs. Illinois for white men, total females, and total population are very similar in terms of distribution.

The South region is lower than the Central region and the North region has the highest numbers because that is the region that includes Chicago. When looking at the regions vs percent of the population that is white, people per housing, and unemployment rate, the trend in data is also very similar.

The values for the North region is the lowest while the Central and South region values are higher and very similar to each other. Cook county is harder to differentiate from the other counties in these plots which shows that Cook county has a low percent of the population that is white, people per housing, and unemployment rate. This makes sense because urban areas are more racially diverse and have more issues regarding housing, so there are more people to one

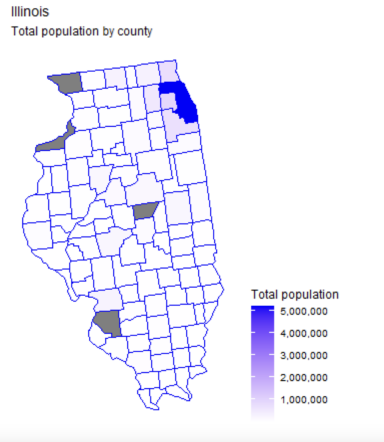

Figure 2: Population in Illinois by County

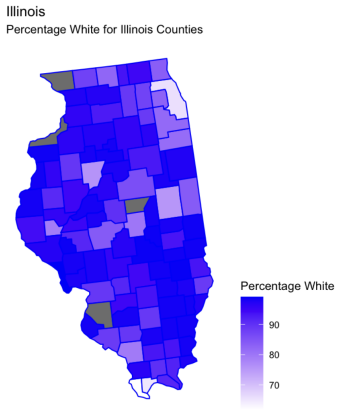

Figure 3: Percentage of White People in Illinois by County

housing unit. There are also more employment opportunities in urban regions than rural areas. In the regions vs percent of the population that is not white, all of the regions are very similar to each other.

We also plotted each of our outcome variables and took note of which counties stood out. The top five counties for total cases, total deaths, December 2020 deaths, and December 2020 deaths per 100000 were the same (Cook, DuPage, Will, Lake, Kane) and are all located in the Cook county area in the Northern part of Illinois. It took Cook county the longest amount of time out of all of the counties for COVID-19 infections to reach 5 percent of the population. In

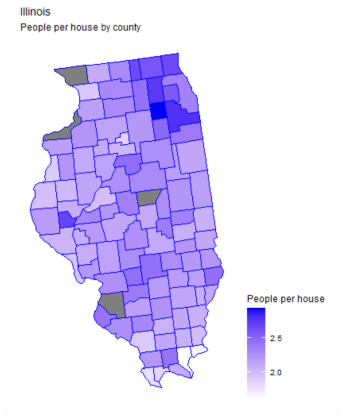

Figure 4: People per Housing Unit in Illinois by County

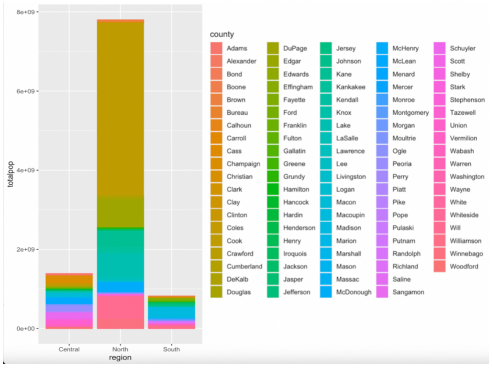

Figure 5: Total Population in Illinois by Region Colored by County

the summary statistics, we can see again that the mean and lower quartiles for our variables are rather low, but there is a really large range which is because of our outlier, Cook county.

On average, there were 48 deaths in December of 2020 and it took 241 days for for COVID-19 infections to reach 5 percent of the population for each county. The distributions for the histograms regarding regions vs. total cases, total deaths, deaths in December of 2020, and deaths per 100000 in December of 2020 were very similar to each other.

The distribution of the histogram showing regions vs. number of days it took COVID-19 infections to reach 5 percent of the population is lower for the North region while the Central and South regions are close to each other, which is due to Cook county accounting for less of the values since it took the longest

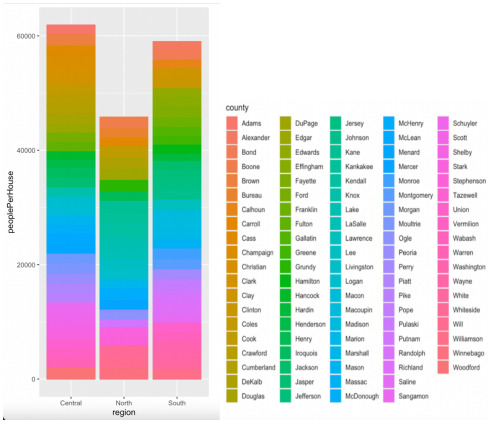

Figure 6: People per Housing in Illinois by Region Colored by County

Cases, Deaths, Total Cases, Total Deaths, Deaths in December of 2020, Deaths per 100000 in December of 2020, Number of Days to hit 5% (Respectively)

```
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean     sd p0 p25  p50  p75    p100 hist
1 data                  0             1 12775. 63509.  1 495 2219 6145 1193914 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean    sd p0 p25 p50 p75  p100 hist
1 data                  0             1 201. 1021.  0   9  43 105 14936 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean      sd  p0  p25  p50   p75    p100 hist
1 data                  0             1 33046. 128301. 960 4089 7509 14285 1193914 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean    sd p0 p25 p50 p75  p100 hist
1 data                  0             1 367. 1577.  8  49 106 176 14936 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd p0 p25 p50 p75 p100 hist
1 data                  0             1 48.7 168.  0   8  19  33 1590 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate     mean      sd p0     p25     p50     p75   p100 hist
1 data                  0             1 0.000487 0.00168  0 0.00008 0.00019 0.00033 0.0159 ▇▁▁▁▁
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate mean   sd  p0 p25 p50 p75 p100 hist
1 data                  0             1 241. 18.5 158 231 240 250  298 ▁▁▇▁▁
```

Figure 7: Summary Statistics for Outcome Variables

amount of days for COVID-19 infections to reach 5 percent of the population.

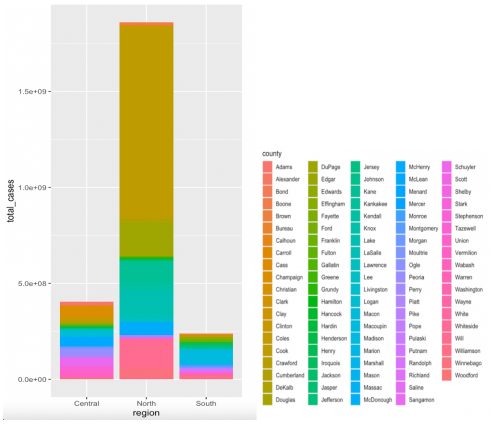

Figure 8: Regions in IL vs. Total Cases

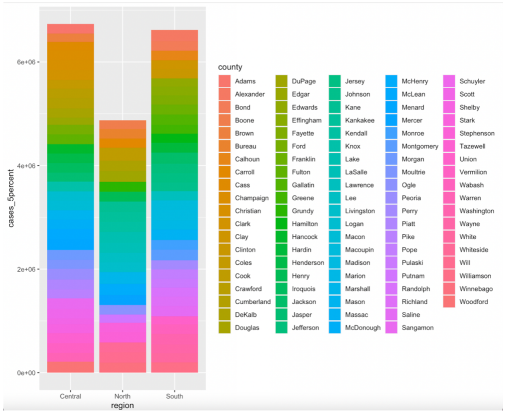

Figure 9: Regions in IL vs. Number of days that it took for cases to hit 5 percent of the population

4 Bivariate Testing

After conducting our exploratory data analysis, we decided to focus on our explanatory variables regarding unemployment rate, people per housing, and total population, and our outcome variables regarding the number of days it took COVID-19 infections to reach 5 percent of the population and deaths per 100000 in December of 2020. We modeled 2 variables against each other at a time.

All of the relationships we found were roughly linear, with total population and deaths per 100000 in December of 2020 being the most positively linear.

Examining these plots solidified our interest in exploring unemployment rate and people per housing.

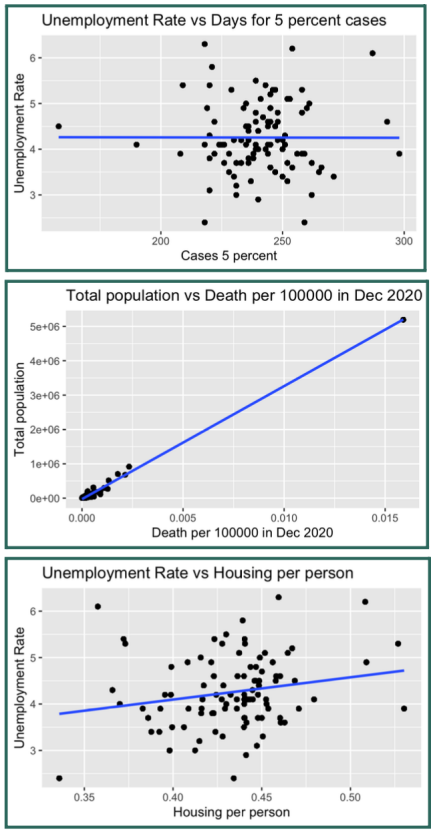

Figure 10: Regions in IL vs. Number of days that it took for cases to hit 5 percent of the population

5 Multivariate Modeling

To answer our overarching question of which variables most heavily influenced the total COVID-19 deaths in Illinois, we found the best fitting multivariate model for predicting deaths. This model shows us which predictor variables are the most influential, and allows us to quantify their influence by looking at their respective βs. We first created a model that included every predictor variable to get a baseline fit. The regression equation was $deaths \sim whitemale + whitefemale + totalpop + pctwhite + unemploymentRate + peoplePerHouse + region$. The resulting output is shown below.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.090e+02  4.376e+01  -4.777 1.78e-06 ***
whitemale       -5.327e-03  1.337e-03  -3.984 6.77e-05 ***
whitefemale      1.201e-03  1.312e-03   0.915 0.359977
totalpop         2.977e-03  5.616e-05  53.013  < 2e-16 ***
pct_white        1.591e+00  2.758e-01   5.771 7.93e-09 ***
unemploymentRate 1.025e+01  2.812e+00   3.645 0.000267 ***
peoplePerHouse   1.456e+01  1.104e+01   1.319 0.187268
regionNorth     -6.145e-01  5.411e+00  -0.114 0.909577
regionSouth      8.779e+00  4.266e+00   2.058 0.039612 *
```

Figure 11: Initial Multivariate Model Output

We next removed the predictor variables that had a p-value > 0.05, as those variables are not statistically significantly influential. Looking at the output above, the variables that had large p-values were *whitefemale, peoplePerHouse,* and *region*. Our new regression equation without these variables was $deaths \sim whitemale + totalpop + pctwhite + unemploymentRate$. The output from this regression is shown below.

Now all of the included predictor variables have a p-value < 0.05, meaning they are all statistically significant and should be left in the regression model. This shows that the best fitting predictor variables are white male, total population, percent white, and unemployment rate. The R-squared of this regression is 0.7715, which is fairly high, and supports this subset of variables being a good fit for predicting deaths. From analyzing the coefficients, we see that the largest β, and therefore most influential variable, is that of unemployment rate. It is also positive, meaning that as unemployment rate increases deaths also increases. The only variable with a negative coefficient is white male, which shows that areas with more white males are expected to have fewer deaths from

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.632e+02  3.032e+01  -5.382 7.40e-08 ***
whitemale       -4.098e-03  1.426e-04 -28.732  < 2e-16 ***
totalpop         2.975e-03  4.682e-05  63.529  < 2e-16 ***
pct_white        1.523e+00  2.710e-01   5.619 1.93e-08 ***
unemploymentRate 9.247e+00  2.469e+00   3.745 0.000181 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 488.2 on 75702 degrees of freedom
Multiple R-squared:  0.7715,    Adjusted R-squared:  0.7715
F-statistic: 6.39e+04 on 4 and 75702 DF,  p-value: < 2.2e-16
```

Figure 12: Subset Multivariate Model Output

COVID-19. Both of these findings support our initial hypotheses that deaths would be higher in counties that are poorer and/or have a larger non-white population.

6 Hypothesis Testing

To find relations between the variable we found interesting, we conducted two hypothesis tests.

(i) **Unemployment Rate:**

We conducted a Welch two sample t-test to compare the means between the levels of unemployment rates in the state of Illinois.
The unemployment rates across the state varied from a wide range of 2.4 to 6.3. To help distinguish the unemployment rates, we divided the variable into two categories - **"low"** and **"high"**. For data values with unemployment rate greater than 4.3, it was classified into the "high unemployment rate" category while for all the data values with unemployment rate less than and equal to 4.3, it was classified into the "low unemployment rate" category. The high unemployment category implied lower income stability and the low unemployment category implied higher income stability.

Below are the null and alternate hypothesis for the test.
The Null hypothesis :
The difference between the means of death for low and high unemployment

is equal to 0.

$$H_0 : \mu_1 - \mu_2 = 0$$

The Alternative hypothesis :
The difference between the means of death for low and high unemployment is not equal to 0.

$$H_1 : \mu_1 - \mu_2 \neq 0$$

```
        Welch Two Sample t-test

data:  low_unemp and high_unemp
t = 29.417, df = 31124, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 209.8763 239.8405
sample estimates:
mean of x mean of y
296.68696  71.82853
```

Figure 13: Two sided hypothesis test result for Unemployment Rate

Upon conducting the two sided hypothesis testing on the above defined data, the p-value for the test came out to be $2.2e - 16$. As the p-value came out to be a very small value, we **rejected the null hypothesis** that the true difference in means between low and high unemployment rate deaths is not equal to 0. The 95% confidence interval for the data was lower ci = 209.8763 and higher ci = 238.8405 **Vaccine Availability:**

We conducted a Welch two sample t-test to compare the means between the new cases before and after the availability of COVID-19 vaccine to the public in the state of Illinois.
The vaccine was introduced to the public on the 29th of March, 2021. Since we were interested to compare the effect of vaccine on the new cases, we examined data for 180 days before and after the availability of vaccine.

Below are the null and alternate hypothesis for the test.

The Null hypothesis :
The difference between the means of new cases before and after the vaccine is equal to 0.

$$H_0 : \mu_1 - \mu_2 = 0$$

The Alternative hypothesis :
The difference between the means of new cases before and after the vaccine is not equal to 0.

$$H_1 : \mu_1 - \mu_2 \neq 0$$

```
        Welch Two Sample t-test

data:  beforevaccine and aftervaccine
t = 0.00015227, df = 30932, p-value = 0.9999
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1454.158  1454.384
sample estimates:
 mean of x  mean of y
0.20233397 0.08935618
```

Figure 14: Two sided hypothesis test result for Vaccine Availability

Upon conducting the two sided hypothesis testing on the above defined data, the p-value for the test came out to be 0.9999. As the p-value came out to be a very large value, we **failed rejected the null hypothesis** the means of new cases before and after the vaccine is equal to 0. The 95% confidence interval for the data was lower ci = 0.20233397 and higher ci = 0.08935618

7 Conclusion

After completing hypothesis tests, we found that employment rate does, in fact, play a significant role in COVID-19 deaths, with counties of higher employment rates having fewer deaths than counties of lower employment rates. This coincides with our hypothesis that lower employment rates means less income which leads to more deaths. We also found that the vaccine did not

affect new cases. This is the opposite of what we expected to see, and it may be due to the vaccine taking a long time to become effective in a population, or people not getting the vaccine right away. As for our overarching question of the best predictor variables, we found that white male, total pop, pct white, and unemployment rate were the best subset to predict deaths, and this regression subset yielded an R-squared value of 0.7715.

References

[1] https://www.chicagotribune.com/coronavirus/ct-viz-coronavirus-timeline-20200507-uvrzs32nljabrpn6vkzq7m2fpq-story.html

[2] https://www.thecentersquare.com/illinois/how-income-inequality-in-illinois-compares-to-other-states/article-76d83857-5bd0-5d22-ae4a-894935303b56.html

[3] https://www.census.gov/library/stories/state-by-state/illinois-population-change-between-census-decade.html