DATA512 A7 Final Report

Ryan Williams

Introduction

The global COVID-19 pandemic has disrupted the lives of nearly everyone in the world. Many people have died, lost loved ones, lost their jobs or sources of income, or otherwise had their lives disrupted severely. The scale of disruption caused by the pandemic, both globally and locally, will likely continue to be studied for many years to come.

This analysis focuses specifically on how Cuyahoga County, Ohio has been disrupted by the pandemic. In this analysis, I seek to understand two broad topics: how mask mandates impacted the spread of COVID-19 infections in Cuyahoga, and how the spread of infections impacted unemployment rates in Cuyahoga.

Background & Related Work

I have 3 primary research questions driving this work. The first question was the original scope of the analysis, questions 2 and 3 were added afterward as an extension:

- **Question 1**: How did masking policies effect the spread of COVID-19 in Cuyahoga county?
    - **Hypothesis:** Enacting the masking mandate, and later removing it, had a statistically significant impact on the spread of COVID-19 infections
- **Question 2**: Is unemployment in Cuyahoga County correlated to changes in COVID-19 infection rate?
    - **Hypothesis:** Unemployment rate and COVID-19 infection rate will have a high correlation (>0.5), but there will probably be a stronger relationship between infection rate and unemployment when introducing a time lag for unemployment to "catch up" to infections.
- **Question 3**: Was the unemployment rate in Cuyahoga significantly different than the unemployment rates in the other counties in the Cleveland-Elyria metropolitan area?
    - **Hypothesis:** Cuyahoga, being the most populous county in the metropolitan area, and containing by far the largest city (Cleveland) will show a statistically significant difference in values and rate of changes in unemployment compared to the rest of the metropolitan area.

There are already an enormous amount of studies on infection modeling, masking, and unemployment over the course of the pandemic. However, as far as I can tell, there is no other research that looks into these topics specifically for Cuyahoga County. Below are some examples of related research, which informed my hypotheses (see 'References' section for full citations).

Infection Modeling

A variety of methods have been used to model infection rates during the pandemic. Early on I considered using a SIRD model like the one documented here, however decided that such a model would be too complex for the time constraints in this assignment. Instead I decided to use the Causal Impact package (documented here) to create a more simplistic time series model for counterfactual analysis.

Impact of Masking

Studies on the impact of masking (like this one published in PNAS), and mask mandates (like these published by the CDC and PLOS), have shown that the requirement and use of masks generally leads to fewer infections – this is why it seems reasonable to hypothesize that mask mandates in Cuyahoga will have an impact on infection rates.

Impact of COVID-19 on Unemployment

Research into the impact of COVID-19 on unemployment has shown how not everyone is impacted equally. An article published by Pew Research shows that some groups of people – like women and people of color – have faced higher rates of unemployment. Additionally, The Center on Budget and Policy Priorities did research showing that people with low-wage jobs were more likely to become unemployed impacted. Given that Cuyahoga is the largest, most urban county in the Cleveland-Elyria metropolitan area, and therefore has a unique mix of demographics and jobs, I hypothesize that unemployment rates in Cuyahoga may be impacted differently by COVID-19 than rates in the surrounding counties.

Methodology

Each of my three research questions required a different methodology:

How did masking policies effect the spread of COVID-19 in Cuyahoga county?

I use a counterfactual approach to determine whether enacting the Ohio mask mandate, and later removing it, influenced the spread of COVID-19. To do this analysis I use the Causal Impact package in Python, which models future predictions using a Bayesian structural time-series model, and then makes the counterfactual comparison of modeled values to actual values for statistical significance. I use two models to make comparisons: one is a model of what would happen if the mask mandate was never enacted, modeled on the pre-mandate data from 2020-04-10 to 2020-07-07. The other is a model of what would happen if the mask mandate never ended, modeled on data during the mask mandate (2020-07-08 to 2020-06-01).

I chose to do a counterfactual analysis because it's a common approach for making causal inferences in absence of a controlled test environment. I chose the Causal Impact model for my prediction because of its simplicity of implementation, though it also comes with a host of issues as a tradeoff (detailed later in this report).

Is unemployment in Cuyahoga County correlated to changes in COVID-19 infection rate?

For this question I calculate the Pearson correlation between monthly unemployment rate and monthly infection rate for the entire timespan of our infection data (2020-04-10 to 2021-08-15). Given that there may be a time lag in unemployment adapting to changes in infection rates, I also check whether introducing a lag of 1 to 6 months makes a difference in correlation.

Since the question is asking whether two series move together, correlation is a natural choice for a metric. However, recognizing that it takes time for people to adapt and make decisions, I decided to introduce the lags to see if those make a difference.

Was the unemployment rate in Cuyahoga significantly different than the unemployment rates in the rest of the Cleveland-Elyria metropolitan area?

"Different" can mean two things in this case: different values for unemployment, or different rates of change for unemployment. To answer this question I fit OLS regression models to the unemployment time series for each of the counties in the Cleveland-Elyria metro area:

$$Unemployment_{County} = Intercept + Time$$

To make sure that OLS regression does a good job of summarizing the time series data, I look for statistical significance of the time variable to determine that unemployment does vary with time, and an R^2 value of 0.5 or higher to determine that OLS fits the data reasonably well. Then I compare regressions to determine whether there is a statistically significant difference in values or slope, at a 0.05 alpha level, between Cuyahoga and each of the other 4 counties.

To do the tests for statistical significance, I actually use a different set of regressions. Determining whether the regressions have statistically significant differences in values can be reinterpreted as determining whether 'County' as a categorical variable has a statistically significant impact on a regression model for unemployment. I use this model, where 'County'' is an indicator variable representing either Cuyahoga or the comparison county:

$$Unemployment_{Counties} = Intercept + Time + County$$

Where the p-value of 'County' in the regression tells us whether regression lines for Cuyahoga and the comparison county have significantly different unemployment values. To compare slopes I use the same principle with a slightly different regression:

$$Unemployment_{Counties} = Intercept + Time + County + (Time * County)$$

Where the p-value of the interaction term, (Time * County), tells us whether the regression lines for Cuyahoga and the comparison county have significantly different slopes, i.e. a difference in rate of change of unemployment.

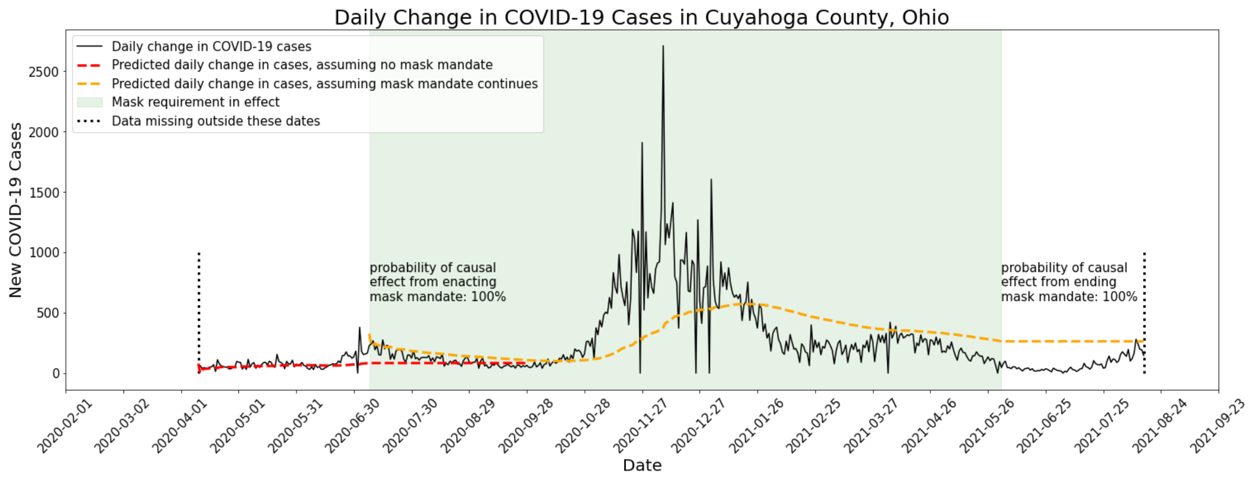
Since I do 4 statistical tests at once in each of these cases, I use a Bonferroni correction and divide the alpha level by 4 to account for the increased possibility of type 1 error - effectively using a 0.0125 alpha for each test.

I chose to use OLS models for this question because the time series for unemployment looks relatively linear, and because OLS summarizes the two metrics I'm interested in (values and rate of change).

Findings

How did masking policies effect the spread of COVID-19 in Cuyahoga county?

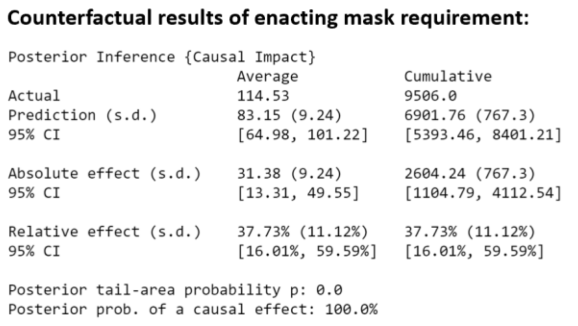
The results of my counterfactual analysis are summarized by this figure:



This chart shows the daily change in COVID-19 cases in Cuyahoga February 1, 2020 through October 15, 2021, along with the time range of the county's mask mandate, and a prediction of the impact that the mask mandate had on COVID-19 infections. The x-axis represents individual days from 2020-02-01 to 20201-10-15. The y-axis represents the count of new COVID-19 cases. The shaded green area shows the time range for which the mask mandate was in effect (2020-07-08 to 2020-06-01). There are 3 lines plotted on this chart:

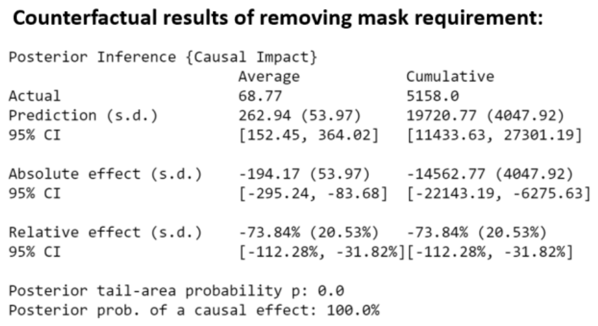
- The solid black line shows the change in daily new COVID-19 cases over time (see note below on the underlying data)
- The dashed red line is a prediction made by the Causal Impact model, based on data *prior* to the mask mandate from 2020-04-10 to 2020-07-07, of what the daily new infections would be in the 3 months following the mask mandate from 2020-07-08 to 2020-10-08.
- The dashed orange line is a different prediction made by the Causal Impact model, based on data *during* the mask mandate from 2020-07-08 to 2021-06-01, of what the daily new infections would be in the time following the end of mask mandate from 2021-06-02 to 2021-08-15.

The Causal Impact model produced the following results:



Counterfactual results of enacting mask requirement:

```
Posterior Inference {Causal Impact}
                          Average            Cumulative
Actual                    114.53             9506.0
Prediction (s.d.)         83.15 (9.24)       6901.76 (767.3)
95% CI                    [64.98, 101.22]    [5393.46, 8401.21]

Absolute effect (s.d.)    31.38 (9.24)       2604.24 (767.3)
95% CI                    [13.31, 49.55]     [1104.79, 4112.54]

Relative effect (s.d.)    37.73% (11.12%)    37.73% (11.12%)
95% CI                    [16.01%, 59.59%]   [16.01%, 59.59%]

Posterior tail-area probability p: 0.0
Posterior prob. of a causal effect: 100.0%
```



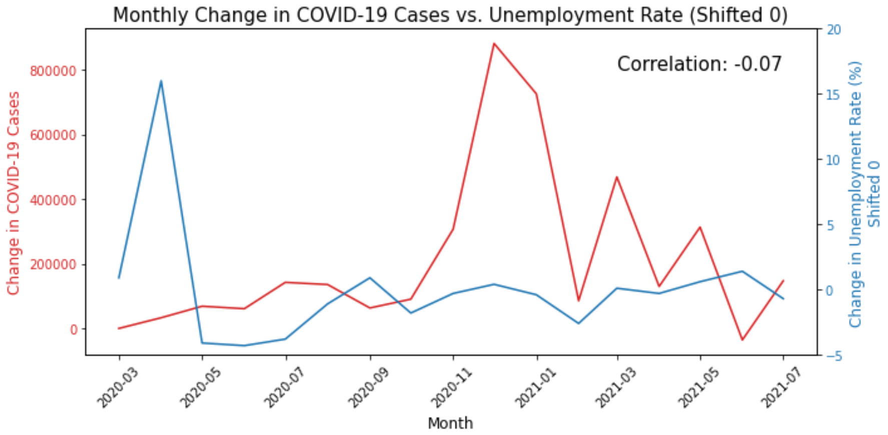
Counterfactual results of removing mask requirement:

```
Posterior Inference {Causal Impact}
                          Average            Cumulative
Actual                    68.77              5158.0
Prediction (s.d.)         262.94 (53.97)     19720.77 (4047.92)
95% CI                    [152.45, 364.02]   [11433.63, 27301.19]

Absolute effect (s.d.)    -194.17 (53.97)    -14562.77 (4047.92)
95% CI                    [-295.24, -83.68]  [-22143.19, -6275.63]

Relative effect (s.d.)    -73.84% (20.53%)   -73.84% (20.53%)
95% CI                    [-112.28%, -31.82%][-112.28%, -31.82%]

Posterior tail-area probability p: 0.0
Posterior prob. of a causal effect: 100.0%
```

We can see the posterior probability of a causal effect in both cases is calculated to be 100%. However, this of course assumes that the model is a good fit for the data – and even just inspecting the visualization shows significant differences between the models and the data they were trained on.

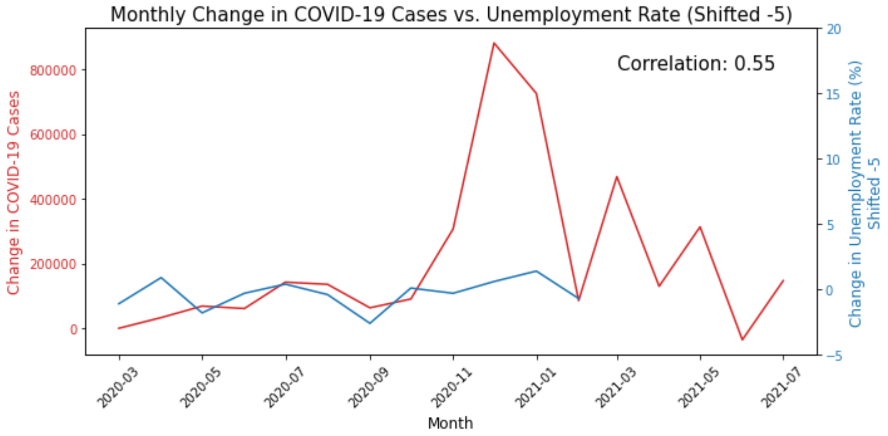
Is unemployment in Cuyahoga County correlated to changes in COVID-19 infection rate?

Calculating the Pearson correlation between monthly change in infections and unemployment rate yielded an unimpressive value of -0.07.



However, part of my hypothesis was that COVID-19 infections may be a leading indicator of unemployment, so I introduced incremental time lags of one month at a time and observed the correlation. Correlation increases between COVID-19 cases and unemployment rate as lags increase, maximizing at a lag of 5 months with a correlation of 0.55, and then starting to decrease with further lags.

Number of 1-month lags between infection rate and unemployment rate	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Correlation	-0.07	-0.14	0.15	0.32	0.29	0.55	0.3



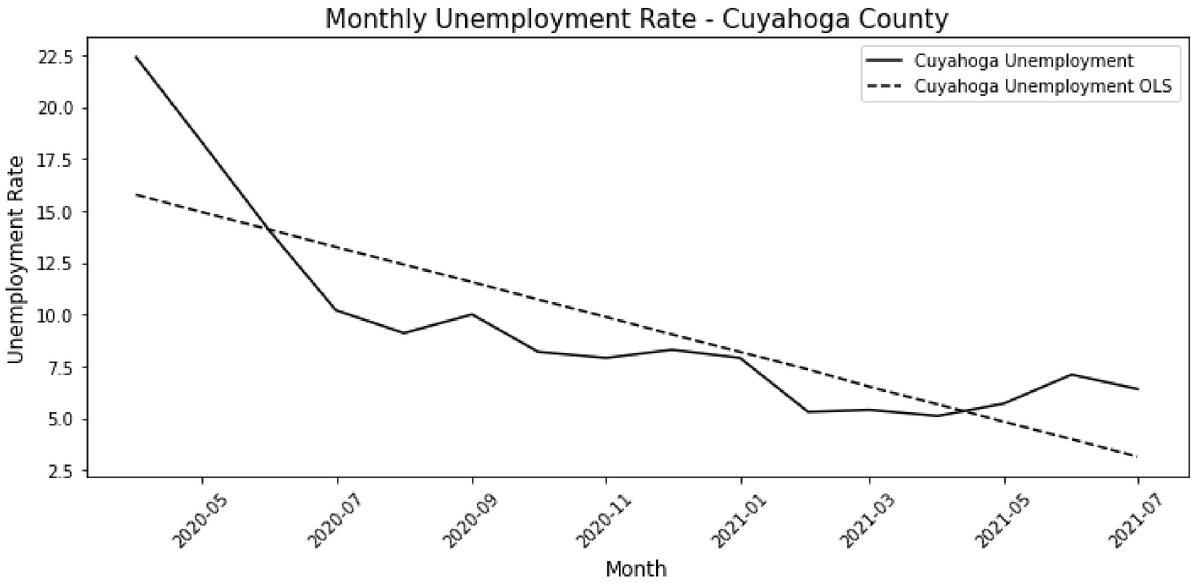
Correlation does maximize at a higher point than the 0.5 I originally hypothesized, though I would still call 0.55 a weak correlation. There’s also an implication from the lags that changes in unemployment take about 5 months to catch up to changes in COVID-19 infections.

Was the unemployment rate in Cuyahoga significantly different than the unemployment rates in the rest of the Cleveland-Elyria metropolitan area?

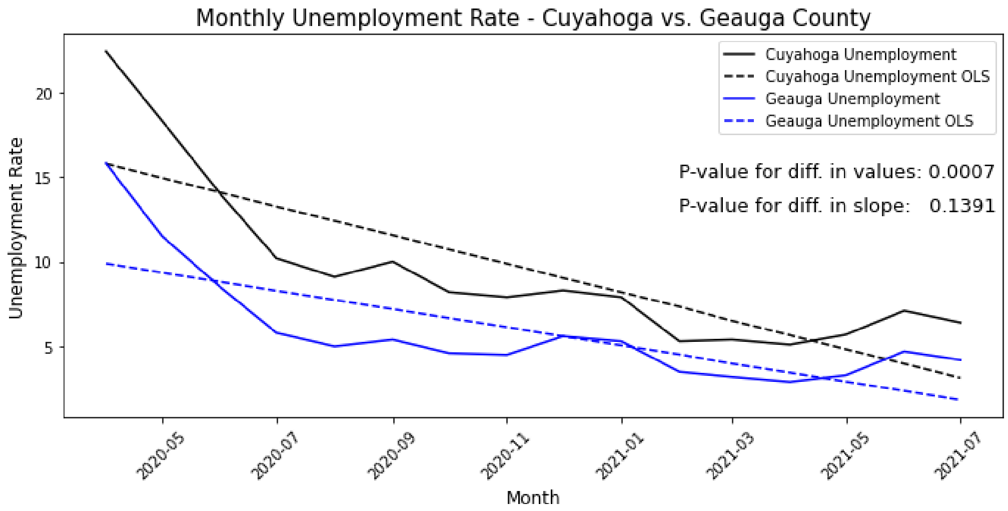
Fitting OLS regressions to unemployment rate over time in each of the 5 counties produces results that give confidence in using OLS to model rate of change of unemployment in these counties, in terms of p-values for the effect of time on unemployment, and the R^2 value for model fit:

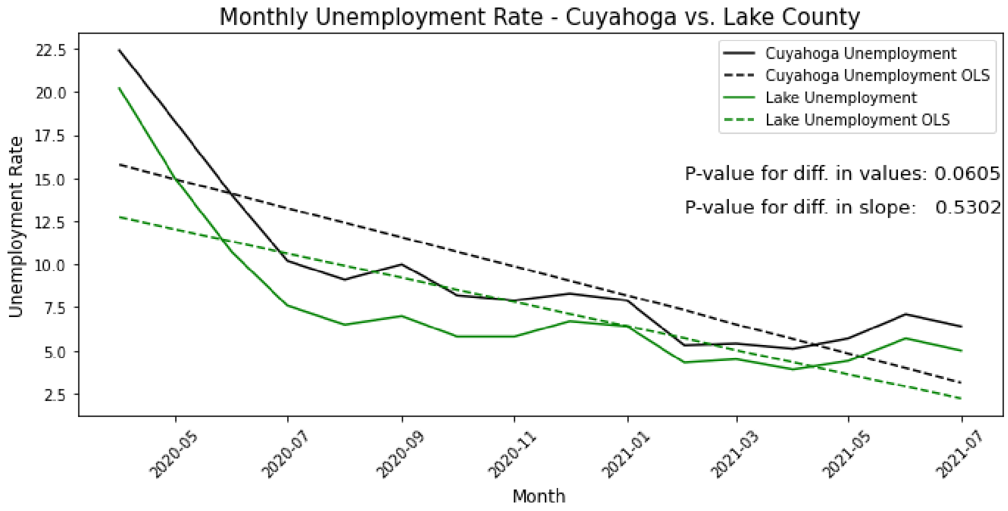
Results of OLS Regression: Unemployment ~ 1 + Time		
County	**p-value of Time**	**R^2 value**
Cuyahoga	< 0.000	0.676
Geauga	0.001	0.560
Lorain	0.001	0.579
Lake	0.001	0.587
Medina	0.001	0.568

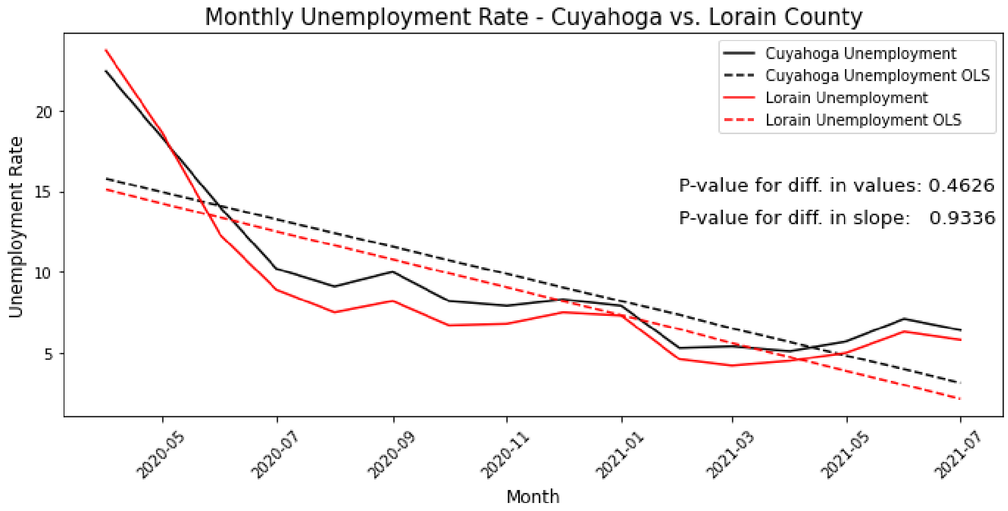
Additionally, visualizing the modeled vs. actual values shows that while the data doesn't exactly follow a linear trend, a linear regression looks like a reasonable way to summarize the values and rate of change for comparisons between counties:

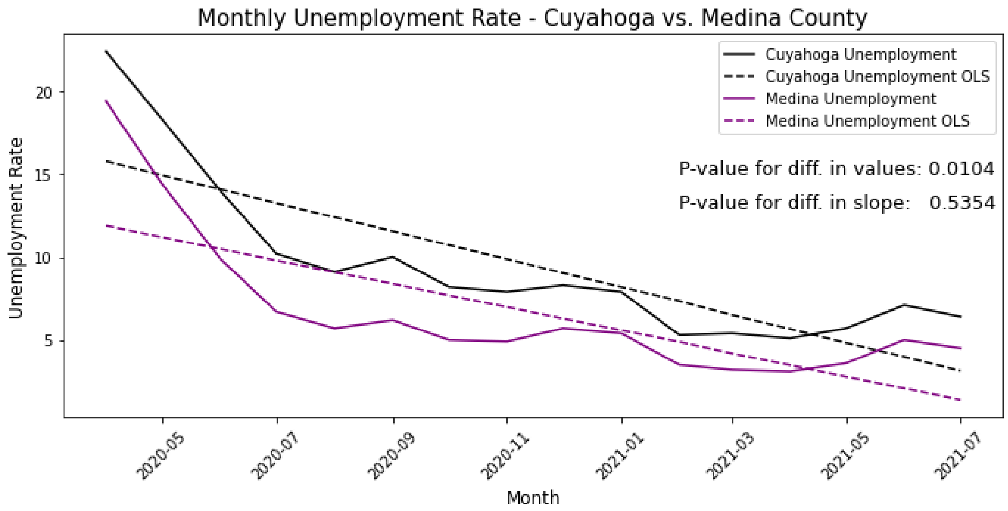


Given that linear regression seems like an appropriate model, I moved forward with comparing regressions between counties. The charts below show the difference in actual values, difference in modeled values, and statistical significance of differences between values & slope for each pairwise comparison:


Monthly Unemployment Rate - Cuyahoga vs. Geauga County
Cuyahoga Unemployment
Cuyahoga Unemployment OLS
Geauga Unemployment
Geauga Unemployment OLS
P-value for diff. in values: 0.0007
P-value for diff. in slope: 0.1391
Unemployment Rate
20
15
10
5
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
Month


Monthly Unemployment Rate - Cuyahoga vs. Lake County
Cuyahoga Unemployment
Cuyahoga Unemployment OLS
Lake Unemployment
Lake Unemployment OLS
P-value for diff. in values: 0.0605
P-value for diff. in slope: 0.5302
Unemployment Rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
Month


Monthly Unemployment Rate - Cuyahoga vs. Lorain County
Cuyahoga Unemployment
Cuyahoga Unemployment OLS
Lorain Unemployment
Lorain Unemployment OLS
P-value for diff. in values: 0.4626
P-value for diff. in slope: 0.9336
Unemployment Rate
20
15
10
5
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
Month



There are statistically significant differences in values for unemployment rate between Cuyahoga and Geauga and Medina county at the alpha = 0.0125 level (Bonferroni corrected from alpha = 0.05). There are no statistically significant differences in rate of change.

Discussion/Implications

There are three direct implications for Cuyahoga county if the findings above are taken at face value: that changes in masking policy had a significant impact on COVID-19 infection rates, that changes in infection rates are a weak leading indicator of changes in unemployment, and that unemployment rates and rates of change in unemployment are generally not significantly different between Cuyahoga and its neighboring counties. However, there is a lot to unpack with these implications.

Further research is absolutely needed on the impact of masking policies in Cuyahoga for the first set of findings to be conclusive. The time series model used to predict the impact of not changing mask requirements is very naïve, as I elaborate on in the 'Limitations' section. A more sophisticated model for epidemiology, like a SIRD model, could lead to much more convincing conclusions. A model that can take into account features like vaccine availability and introduction of new viral strains would also be welcome.

Additionally, this research only cracks the surface of what we could try to understand about the impact of COVID-19 on unemployment in Cuyahoga. I initially wanted to look at correlation between infection rates and unemployment rates because I was wondering how people and policymakers might adapt their employment policies to changes in infection rates. However, looking only at correlation doesn't tell us everything. What seems likely, especially given the weak correlations observed in my results, is that these decisions are made based on more factors than just discrete changes in infections. There are certainly more complex social factors that future research could account for, like examining whether there were specific policy changes in Cuyahoga related to unemployment. For example, some companies (and entire industries) were certainly required by the government to shut down for a period of time. It's also possible that Cuyahoga adapted to news of COVID-19 infections in other parts of the world rather than just examining its own infection rates.

I was interested in looking at comparisons between counties because I suspected that the counties would have very different industries and demographics that may be differently affected by COVID-19. Although my testing only showed that Cuyahoga differed significantly in unemployment rate from

Geauga and Medina, and differed from no counties in rate of change of unemployment, there is more research that could be done. I wasn't able to look at the very human-centric questions of how demographics compared, or how workers or industries differed, between Cuyahoga and its neighboring counties. Given the opportunity for further research, I would start by examining how the counties differ by demographics and industries, and then develop new research questions specifically around how unemployment in those demographics and industries was impacted by COVID-19.

Limitations

There are many limitations of this research, with some key limitations observed here in 4 sections:

The model used for the impact of mask requirements

As stated in earlier sections, the Causal Impact model used for making time series forecasts in Research Question #1 is very naïve. It only takes historical data into account in making its predictions; however, we know there are many interventions occurring with COVID-19 that the model can't account for. Some of these interventions include things like vaccination availability, quarantine mandates, and introduction of new COVID-19 strains. Additionally there's the nature of the epidemiology itself which isn't accounted for, like infected people getting well and becoming immune. A more sophisticated model is needed to reinforce the results of this analysis, potentially a SIRD model which would be better for an epidemiological event like this pandemic.

Using correlation to infer a causal relationship between infections and unemployment

While I claim the evidence of a weak relationship between changes in infection rates and, 5 months later, changes in unemployment rates, it's important to note a few caveats here. One is that we can't assume a causal relationship between infections and unemployment using correlation alone – it's entirely possible (even likely) that these two data points move together as a result of an external variable, like policy changes related to COVID-19. There's also a bit of cherry-picking going on by introducing the monthly lags. On one hand it does seem likely that a lag must be introduced to allow time for unemployment policies to react to changing infections, but this introduces the issue that by random chance alone some time series will correlate better than others, and our improvement in correlation might just be a result of random variance. We can even see in the results observed that correlation doesn't increase exactly linearly (4 lags had a slightly lower correlation than 3).

Use of OLS for comparing unemployment in different counties

The point of using OLS was to summarize the values and rate of change of unemployment in the Cleveland-Elyria metro area counties, but OLS isn't necessarily an ideal model for this. We can see from inspecting the graphs that the trend looks more logarithmic than linear, and a better fitting model might provide more conclusive results. The comparison between regressions also isn't the only way to tackle this question – there are a multitude of ways that could be used to compare unemployment values and rates of change between counties. For example, values could be compared directly with a different statistical test. Finally, even though I show a statistically significant difference in unemployment values between Cuyahoga and Geauga/Medina, looking at historical values for unemployment in these counties shows that Cuyahoga has also generally had much higher unemployment rates in the past – so the significant difference isn't necessarily an impact of COVID-19.

General data limitations

The data used for this analysis was fairly straightforward, however there are some limitations:

- Data collected on infections is incomplete – there will always be some unreported infections, and the degree of this could vary regionally.
- Using monthly unemployment data didn't allow me to utilize the full set of daily infection data (which had to be summarized at a monthly level), though more granular data for employment is unrealistic.

Conclusion

I researched three COVID-19 related questions in Cuyahoga County, Ohio:

- **Question 1**: How did masking policies effect the spread of COVID-19 in Cuyahoga county?
- **Question 2**: Is unemployment in Cuyahoga County correlated to changes in COVID-19 infection rate?
- **Question 3**: Was the unemployment rate in Cuyahoga significantly different than the unemployment rates in the other counties in the Cleveland-Elyria metropolitan area?

What I found was that masking policies did show a significant impact on the spread of COVID-19, but the model I used rendered these results unconvincing. Changes in unemployment rates in Cuyahoga did seem to correlate weakly (correlation = 0.55) with infection rate after introducing a 5 month lag between infections and unemployment. The values for unemployment rates in Cuyahoga differed significantly from values in Geauga and Medina, but rates of change of unemployment were not significantly different between Cuyahoga and any other counties in the Cleveland-Elyria metropolitan area.

While not all of these results are useful by themselves, this research could serve as a foundation for the diving deeper into how Cuyahoga was impacted by the pandemic. Doing this kind of deep, local-level research is deceptively important. While the global impact of COVID-19 is being investigated at large, the local impact to regions like Cuyahoga is still relatively unresearched. By understanding the impact COVID-19 has had on issues like employment at a local level, we not only build on our existing knowledge, but we also create deeper empathy for those affected.

References

- Fernández-Villaverde, J., & Jones, C. (2020). *Estimating and Simulating a SIRD Model of COVID-19 for Many Countries, States, and Cities*. Published online 2020. Retrieved from https://web.stanford.edu/~chadj/sird-paper.pdf.
- Adjodah D, Dinakar K, Chinazzi M, Fraiberger SP, Pentland A, Bates S, et al. (2021) *Association between COVID-19 outcomes and mask mandates, adherence, and attitudes*. PLoS ONE 16(6): e0252315. https://doi.org/10.1371/journal.pone.0252315
- Guy GP Jr., Lee FC, Sunshine G, et al. *Association of State-Issued Mask Mandates and Allowing On-Premises Restaurant Dining with County-Level COVID-19 Case and Death Growth Rates — United States, March 1–December 31, 2020*. MMWR Morb Mortal Wkly Rep 2021;70:350–354. DOI: http://dx.doi.org/10.15585/mmwr.mm7010e3
- Jeremy Howard, Austin Huang, Zhiyuan Li, et al. *An evidence review of face masks against COVID-19*. Proceedings of the National Academy of Sciences Jan 2021. DOI: 10.1073/pnas.2014564118
- *Tracking the COVID-19 economy's effects on food, housing, and employment hardships*. Center on Budget and Policy Priorities. (n.d.). Retrieved December 10, 2021, from https://www.cbpp.org/research/poverty-and-inequality/tracking-the-covid-19-economys-effects-on-food-housing-and.

- Kochhar, R., & Bennett, J. (2021, April 14). *U.S. labor market inches back from the COVID-19 shock, but recovery is far from complete*. Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2021/04/14/u-s-labor-market-inches-back-from-the-covid-19-shock-but-recovery-is-far-from-complete/.
- Documentation for the Python Causal Impact package: https://pypi.org/project/pycausalimpact/

Data Sources

The data used in this analysis is available at the following links:

- The raw US confirmed cases file from the Kaggle repository of John Hopkins University COVID-19 data: https://www.kaggle.com/antgoldbloom/covid19-data-from-john-hopkins-university?select=RAW_us_confirmed_cases.csv
- The CDC dataset of masking mandates by county: https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i
- The New York Times mask compliance survey data (not used in final results): https://github.com/nytimes/covid-19-data/tree/master/mask-use
- Monthly unemployment rates for each of the counties in the Cleveland-Elyria metropolitan area:
    - Cuyahoga county: https://fred.stlouisfed.org/series/OHCUYA5URN
    - Geauga county: https://fred.stlouisfed.org/series/OHGEAU5URN
    - Lake county: https://fred.stlouisfed.org/series/OHLAKE2URN
    - Lorain county: https://fred.stlouisfed.org/series/OHLORA3URN
    - Medina county: https://fred.stlouisfed.org/series/OHMEDI0URN
    - Cleveland-Elyria metropolitan area (not used in final results): https://fred.stlouisfed.org/series/CLEV439URN