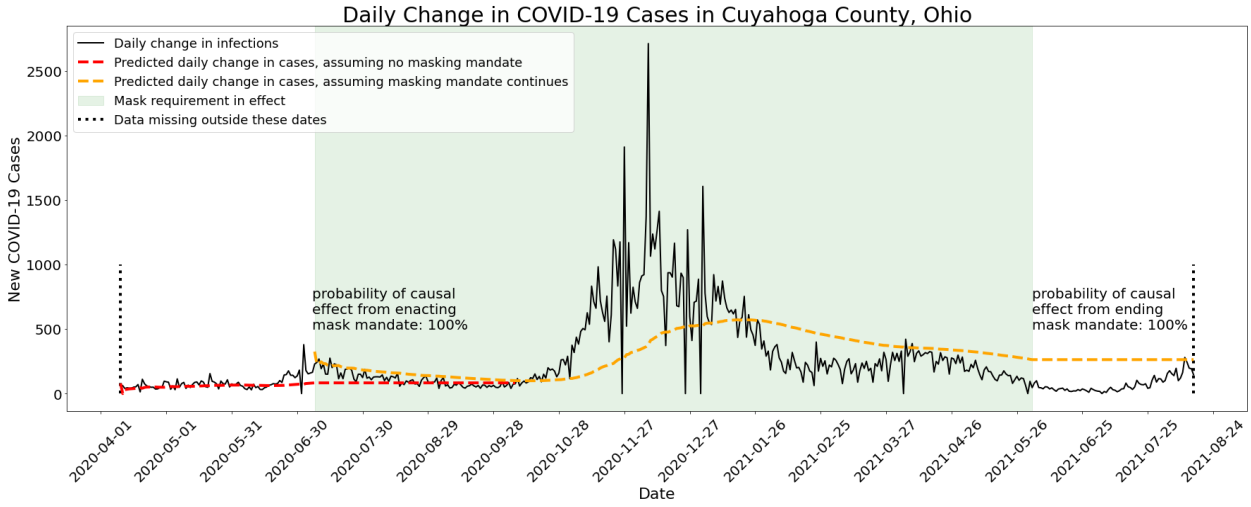
A4 - Common Analysis

Ryan Williams
11/04/2021

Visualization



Explanation of Visualization

This visualization shows the daily change in COVID-19 cases (the derivative of daily cumulative cases) in Cuyahoga County, Ohio from 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 (though only 30-day intervals are labeled to enhance readability). 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, 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, 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 underlying data comes from 2 sources:

- COVID-19 cases came from RAW_us_confirmed_cases.csv from the Kaggle repository of John Hopkins University COVID-19 data. Since this dataset only contains data from April 10, 2020 to August 15, 2021, there are dotted lines on the graph to show that data doesn't exist for the full time range being plotted.
- Dates when the mask mandate was in effect for Cuyahoga County came from the CDC dataset of masking mandates by county

A user of this chart should compare the black line (actuals) to the dashed lines (predictions) to see the apparent impact that the masking mandate had on daily COVID-19 infections. The beginning and ending of the masking mandate are annotated with an estimated probability of a causal effect from these two events - this estimate is the result of a hypothesis test comparing the actual values to the counterfactual predicted values. In both cases there is a statistically significant impact.

Reflection

I was surprised how much of the methodology I used for this assignment was shaped by the collaborative discussions, even though no code was shared in our public Slack channel. There were some very basic questions that turned out to be useful to me: Sabrina Wang asking what to do about dates extending beyond our dataset, and Emily Linebarger started a discussion on what metric to show (infections vs infection rate) which both led to guidance from the TAs that I incorporated. There was also discussion on how people were visualizing their mask mandate periods, which gave me a lot of ideas for how I wanted to create my visualization - I ended up using Emily Linebarger's method for showing the mask mandate by having a shaded region covering the time range.

The discussions about methods for showing the impact of the mask mandate took me down a path that eventually led to my final model. I knew early on that I wanted to create some kind of counterfactual prediction to compare our data to, and I started by looking into SIRD models to estimate what infections would look like without a mask mandate. Kevin Sweet posted a very useful paper for using SIRD to model COVID-19 (linked here). However, after reading the paper and trying to build my own SIRD model, I realized I'd need data from other sources which seemed outside the scope of the assignment. Thinking about the data available to us, which is essentially just daily infections over time with an intervention in the form of the mask mandate, I felt like the most appropriate model would be some kind of time series forecast. I read this Medium article to get an idea of how to do a time series intervention analysis, and I used the Causal Impact package documented here to implement the analysis.

Even though the modeling methodology I ended up using came from my own research, I learned a lot about creating an epidemiological SIRD model from the resources Kevin Sweet shared. My causal impact model is very naive, since it's just forecasting out the historical trend of daily infections. A SIRD model would be much better since it can factor in recoveries, deaths, and changes in contact rate over time. Similarly, using something like a

GLM (an idea proposed by Grant Savage) would allow me to include features like the delta variant and vaccine availability, which are ignored in my current model but should have a big impact on infection rate. Even though my current analysis doesn't incorporate these discussions, I think they were invaluable in helping me think about how I would improve on this analysis with a more robust model.