DATA 200 Graduate Project
Topic 1: Dataset A

Anya Michaelsen (3034964414)

Contents

1 Background **2**
- 1.1 The COVID-19 Pandemic . . . 2
- 1.2 COVID-19 Data Tracking . . . 2
- 1.3 Research Questions . . . 3
- 1.4 Literature Review . . . 3

2 Methodology **3**
- 2.1 The Data . . . 3
  - 2.1.1 COVID-19 Cases Data . . . 4
  - 2.1.2 Weather Data . . . 4
  - 2.1.3 Vaccination Data . . . 4
- 2.2 Data Processing . . . 4
  - 2.2.1 Data Cleaning . . . 4
  - 2.2.2 Feature Engineering . . . 4
  - 2.2.3 Causal Inference . . . 4
  - 2.2.4 Forecast Modeling . . . 5

3 Results **5**
  - 3.0.1 Dataset Findings . . . 5
  - 3.0.2 Model Findings . . . 5
  - 3.0.3 Limitations and Future Research . . . 5

4 Appendices **5**

5 References **5**

1 Background

1.1 The COVID-19 Pandemic

COVID-19 is airborne respiratory disease caused by SARS-CoV-2, which originated in China and has spread to a global pandemic. Symptoms range from mild or even asymptomatic to fatal. While mortality rates for COVID-19 are still being estimated, scientists believe COVID-19 to be substantially more deadly than most strains of flu.As an airborne illness, COVID-19 spreads through droplets in the air, making it highly contagious, and asymptomatic infection combined with up to two week incubation time before symptoms arise make slowing the spread of the virus a public health challenge.

In December of 2019, the first cases of COVID-19 were detected in Wuhan, China. About twenty days later, the Center for Disease Control (CDC) confirmed the first case in the United States, with the first death following about a month after that.

Initially, there was no vaccine for the novel coronavirus, and public health measures included mask wearing and social distancing from others to prevent transmission. At a national level, travel bans were implemented to reduce transmission between countries, in particular slowing the spread from countries with high case rates.

In the United States, measures such as social distancing and masking were quickly politicized, slowing their adoption and mitigating their effectiveness. Mid-March of 2020, the US declared a state of national emergency and some states, such as California, issues Stay-at-Home orders which required people to stay home unless necessary. Over the next year states implemented a variety of measures to stop the spread, including similar stay-at-home orders, mask mandates in public spaces.

In December of 2020, the the first COVID-19 vaccine was approved for emergency use by the Food and Drug Administration (FDA) in the United States. This vaccine, by Pfizer, used a novel vaccination approach that had been developed over years prior to the COVID-19 pandemic. This method required two doses for full vaccination, spaced several weeks apart, and the vaccine itself required special handling that increased distribution challenges. A week later another vaccine by Moderna, applying a similar inoculation strategy was approved for emergency use. A third vaccine, by Johnson & Johnson was later approved in March that required only a single shot and more typical storage requirements. During vaccine roll-outs, approval and recommendations were often stratified by age, medical conditions, and exposure risks.

1.2 COVID-19 Data Tracking

Tracking case numbers, hospitalizations, deaths, and symptom severity has been critical for political bodies making public health decisions as well as for scientists and health care professionals treating and combating the virus. Reporting systems vary globally, both in which metrics are tracked and often how they are defined. Within the United States, case numbers were not centrally tracked at a national level and left to states, which created further disparities in data reporting. Journalists, data scientists, and health institutions took up the mantle of aggregating COVID-19 case data until a national framework could be put in place.

One such database was created and maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and posted publicly on GitHub, which combines state level data for COVID-19 cases from April of 2020 through March of 2021.

By the time vaccines had been developed and approved for use in the United States, the CDC was prepared and able to track roll out in a centralized manner.

1.3 Research Questions

While tracking COVID-19 cases has been crucial for public health policy, it is also important to both predict cases going forward to implement preventative measures, such as mask wearing, social distancing, and increased vaccination, as well as understand the *causes* of transmission to create effective measures and loosen ineffective restrictive ones. These aims would simultaneously save lives, health care costs, and limit unnecessary restrictions on people's lives as much as possible.

Throughout the course of the pandemic in the United States, there have been several significant spikes in COVID-19 cases. Possible causes include variants of the virus that are either more transmissible, more deadly, or both, increased travel during holiday months, anti-masking and anti-vaccine rhetoric and mentality in some regions/populations, changes in weather affecting social gathering patterns, and more.

The goal of this research is to produce models for COVID-19 cases, as measured by 'Confirmed Cases' using state level data for COVID-19 metrics, and explore the effects of several possible variables, including weather temperature data, the distribution of cases by age, and adjacent state COVID-19 metrics. Specifically, the aims are:

Q1 Can weather data, both current and historical averages, be used to improve state-level models for the spread of COVID-19? Can we infer from these models whether extreme temperatures, (high, low, or either), affect the spread of COVID-19?

Q2 Does the ratio of COVID-19 cases by age have any significant effect in predicting COVID-19 deaths?

Q3 Does incorporating COVID data from adjacent states significantly improve our state level COVID models?

1.4 Literature Review

A brief survey of related work on the topic(s) of your analysis and how your project differs from or complements existing research.

2 Methodology

Methodology: carefully describe the methods you use and why they are appropriate for answering your search questions. It must include

- a brief overview of causal inference, which should be written in a way such that another student in Data 100 who has never been exposed to the concept can carry out the analyses involving the datasets in your project.

- a detailed description of how modeling is done in your project, including inference or prediction methods used, feature engineering and regularization if applicable, and cross-validation or test data as appropriate for model selection and evaluation.

2.1 The Data

The primary datasets for this analysis pertain to COVID-19 cases in the United States from April 2020 through March 2021.

2.1.1 COVID-19 Cases Data

2.1.2 Weather Data

2.1.3 Vaccination Data

2.2 Data Processing

Lorem Ipsum

2.2.1 Data Cleaning

Lorem Ipsum

2.2.2 Feature Engineering

Lorem Ipsum

2.2.3 Causal Inference

A black-box forecasting model that can predict COVID-19 cases based on data that is measurable prior to the time of forecasting can be used in allocating resources for hospitals such as ventilators or health care workers as well as vaccines. However such a model would not provide any underlying information about the *causes* of spikes in case number and may be limited in the public health policy ramifications.

In contrast, an interpretable forecasting model would provide future estimates as well as inference into the underlying factors that influence the spread of the virus or its fatality. A feature of a linear model will have an impact on the outcome if its coefficient in the model differs from 0, which would represent excluding the feature from the model. Given training data, a linear model can be fit and the coefficients determined for each feature. However variability in sampling as well as co-linearity can confound the effects of a feature, either by yielding an non-zero coefficient for an irrelevant feature or by weighting an influential feature with a close to zero coefficient, so we turn to bootstrapping methods to create confidence intervals for these coefficients.

Bootstrapping is based on the idea of resampling. If we had access to the overall population and could resample arbitrarily many times, we could construct numerous samples, and for each fit a linear model and look at the range of coefficients for each feature across the samples. A feature that significantly affects the outcome in the underlying population will yield a significant coefficient most of the time, although randomness may prevent it from being so in all samples. While this is unrealistic, if we can plausibly assume our sample is representative of the overall population, then we can perform a similar resampling method *on our sample*, and create confidence intervals for our coefficients using the resampled models.

One final issue that must be considered when attempting to infer causality from a linear model is the issue of colinearity. The meaning of a coefficient in a linear model is the effect of that feature *while holding all others constant*. In the case that two or more features are linearly dependent, i.e. one can be expressed as a linear combination of the others, then it will be impossible to vary one without changing the others as well according to their linear relationship. Thus interpreting the coefficient becomes meaningless. For example with two features that are multiples of each other, slight changes in the data could result in one feature having a positive coefficient and the other with a near-zero coefficient, while a different model may reverse these effects. As a result, the confidence intervals for *both* coefficients are likely to contain zero, since each variable is masking the effect of the other across different models.

ADD CAUSAL INFERENCE REMARKS FOR MY MODEL(S)!!

2.2.4 Forecast Modeling

ex-ante vs post-ante forecasting article:
https://otexts.com/fpp2/forecasting-regression.html#ex-ante-versus-ex-post-forecasts

3 Results

3.0.1 Dataset Findings

3.0.2 Model Findings

3.0.3 Limitations and Future Research

An evaluation of your approach and discuss any limitations of the methods you used.
Describe any surprising discoveries that you made and future work.

4 Appendices

5 References