Homework 1: Estimating Covid-19 Deaths

Neeraj Sharma

4/14/2020

Assignment

Please submit: the answers to the questions, your code, and your dataset. The code you provide should reproduce your model. All of these should be submitted via canvas.

The Governor of Illinois, J. B. Pritzker, has decided that a key input to public policy is knowing how many people will die from Covid-19 in the near future. He has asked you to estimate the total number of official Covid-19 deaths that will be officially recorded in the state of Illinois by April 21 and by May 31.

To fulfill that request, you will need to assemble a data set, do estimation based on that data, and have some sort of theoretical model in your mind to extrapolate out to the future.

- Describe the data set that you chose to assemble and the rationale behind the choices you made in deciding what data to use.
- Describe the model(s) that you settled on for estimation. What was your logic for using that/those particular models?
- Provide an exact number which is your prediction for cumulative official Illinois Covid deaths through April 21
- Provide an exact number which is your prediction for cumulative official Illinois Covid deaths through May 31
- How did you get from the estimates in (2) to the predictions in (3) and (4)?
- Provide an exact number of your best guess of the 90-10 confidence interval for your estimates in questions 3 and 4 (those confidence intervals obviously can be different from one another). A 90-10 confidence interval is the range that would encompass the true value 80 percent of the time. I don't want you necessarily to provide the standard error churned out by the computer, but something more thoughtful. Discuss the thought process/rationale underlying the standard errors you choose.
- Make exactly one pretty picture/graph/slide that you would show to the Governor to allow him to easily understand what he should be expecting in terms of Covid deaths.

Introduction

In December 2019, scientists in China reported the discovery of a novel coronavirus originating from a wild seafood and exotic animal market in the city of Wuhan, Hubei Provence, China. Over the subsequent months, the virus spread over the world, infecting individuals on all populated continents and in nearly every country.[1] The assignment given is to provide a prediction of deaths that might occur by April 21 and by May 31 in the state of Illinois for consideration by JB Pritzker.

Describe the data set that you chose to assemble and the rationale behind the choices you made in deciding what data to use.

My dataset pulls together data from fice sources spread across four general categories. The categories I analyze are:

1. COVID data
    i. January 24, 2020 to March 16, 2020 – Data on cases and deaths as reported by the New York Times
    ii. March 17, 2020 and onwards – Data on cases, deaths, tested, and negative results from the Illinois Department of Public Health

[1] https://www.nytimes.com/article/coronavirus-timeline.html

2. Demographics of IL
    i. County-level demographic data pulled from the 2018 5-year Census American Community Survery
        a) Population
        b) Population under 18
        c) Population enrolled in school
        d) Median Income
        e) Number of Households
        f) Number of Households with people under 18
        g) Number of Households with people over 60
        h) Number of Overcrowded Households
3. Hospital information in IL
    i. County-level capacity, load and utilization data aggregated to the county level from the Illinois Health Facilities and Services Review Board
4. Mobility data recorded in IL
    i. County level data about average mileage traveled by cellphones recorded by cell-tower pings aggregated to state level from Descartes Labs

I began with the mindset that I should attempt to collect as much data as possible and form a model out of what I observe. I realized that this led me down overcomplicated paths and wasted my time. I retained some of this data for several reasons:

First, I chose my sources of COVID data by first extracting the JSON from the Illinois Department of Health website. I figured that the IL DPH would be the most authoritative and accurate source for my data as it has minimal aggregation which causes loss in resolution. To get data prior to early March, I sourced from the New York Times. I first considered sourcing from the Johns Hopkins CSSE dataset, but found there to be moderate discrepancies in Illinois data specifically.

Secondly, I origially wanted to extract demographic data for Illinois to compare to other states like New York or California, but after doing exploratory analysis, I found that route to be too limited. The variables I selected (noted above), I chose because of the qualities and characteristics of the Coronavirus. I specifically was interested in metrics relevant to the elderly population, school age children, and home overcrowding as those metrics are repeadidly discussed in the context of COVID. Unfortunately, it was difficult to find a reasonable match for Illinois at the state level so I opted to compare against other countries.

Third, I got information on hospitals in Illinois to identify the number of ICU and general available beds to determine when the state would be overwhelmed with patients.

Finally, I extracted mobility data to attempt to factor in the impact of social distancing.

Describe the model(s) that you settled on for estimation. What was your logic for using that/those particular models?

I build two models: one for high population counties in Illinois and another for low population counties in Illinois. I chose to split the state up in this way because the spread of viruses in urban centers differs dramatically from in rural areas. Thus, I wanted to define two models to account for this difference.

The model I ultimately settled on for higher population areas relates the deaths at a given time to the number of cases at that time, the date and a co-linearly-related social distancing binary variable, and a binary variable indicating if the hospital ICU cap has been reached yet. I chose this model for high population areas as I believe it captures the impact of social distancing independent from normal progressions in time relatively well. Social distancing is an important behavioral change as is decreases the transmission rate of the virus, so its implementation, or a lack thereof, is an important motivator behind deaths once cases begin to trend upwards exponentially. Distancing occurs part-ways through the COVID pandemic and disproportionately affects high population areas. Accounting for social distancing and the date through including them as co linearly-related variables is justified as movement is dependent on if social distancing is in place or not. Including these variables in my model is a straightforward, yet simple way to account for the effect social distancing has on deaths. I choose to include data on cases and if the ICU cap has been reached yet as

those each impact the death rate of infected patients. As cases increase, deaths will naturally increase as well. However, they will accelerate more dramatically once the ICU cap has been reached as hospitals will be unable to devote the resources necessary to meet the needs of each individual patient.

For low population areas, I identify a linear trend that I extrapolate outwards. If there will be any significant death, it will occur in population centers like Cook County or East Saint Louis. This linear trend is sufficient for the small number of deaths it accounts for.

Provide an exact number which is your prediction for cumulative official Illinois Covid deaths through April 21

I predict that there will be 1099 deaths through April 21.

Provide an exact number which is your prediction for cumulative official Illinois Covid deaths through May 31

I predict that there will be 1183 deaths through May 31.

How did you get from the estimates in (2) to the predictions in (3) and (4)?

In order to turn my high population regression into a predictive model I needed to feed the model data for the future in order to estimate deaths on April 21 and the end of May. The most complicated variable to predict data for is the cases, as that essentially is another recreation of this project as a whole. I approach this question in a simplified manner. First, I identify countries with similar demographic and growth curves to Illinois and base the future cases in Illinois on the progression of cases I observe in those other countries. I use a logisic model to estimate those curves using the method of non-linear squares to fit the data. I then use the equation produced by the logistic model to estimate the future progression of cases in Illinois. Once the cases data has been created, creating dummy data for the socialdistancing? binary variable, date, and at_icu_cap is relatively straightforward as social distancing will continue to remain in place, the date is known, and the ICU cap is dependent on cases which I have just explained my predictive model for.

I then re-run my regression with the extended dataset based on my predicted estimates for cases, date, socialdistancing?, and ICU capacity to identify what the deaths on April 21 and May 31 will be. My model eventually trends downwards as my dummy data for cases eventually remains constant given the logistic fitting I use to project cases. Thus, for May 31, I identify the maximum number of deaths that occured in my model and pull that number forward.

Provide an exact number of your best guess of the 90-10 confidence interval for your estimates in questions 3 and 4 (those confidence intervals obviously can be different from one another). A 90-10 confidence interval is the range that would encompass the true value 80 percent of the time. I don't want you necessarily to provide the standard error churned out by the computer, but something more thoughtful. Discuss the thought process/rationale underlying the standard errors you choose.

The computer tells me that the 90-10 prediction interval (+, -) for April 21 is (1128.7505, 1068.7173), and that the 90-10 prediction interval (+, -) for May 31 is (1,224.14027, 1,140.3893). That's 100% not true and my model is significantly less accurate than that. I think that the error is more likely closer to (1300, 1000) for April 21 and (1700, 1200) for May 31. The reason I believe this is that my model for estimating the total number of cases has an increasingly smaller standard deviation as the model continues on as the logistic curve flattens. This is not an accurate representation of reality. Because my model underestimates the standard deviation, the standard error and prediction interval becomes more fixed as the date approachs May 31. In reality, the model should include have increasing uncertainty as time progresses, and have a higher bound on the confidence interval than my model reports.

It's impossible for me to select a defined set of errors to square and include in this model because of the limitations of my prediction methodology. In the future, I will keep this in mind.

Make exactly one pretty picture/graph/slide that you would show to the Governor to allow him to easily understand what he should be expecting in terms of Covid deaths.

```
il_final <- il  %>%
  filter(var == "deaths") %>%
  pivot_wider(names_from = "var", values_from = "count") %>%
  bind_rows(forecast_total %>%
              mutate(hpop_deaths = if_else(date > ymd("2020-04-26"), 1140.1446, hpop_deaths)) %>%
              mutate(hpop_upper = if_else(date > ymd("2020-04-26"), 1143.8320, hpop_upper)) %>%
              mutate(hpop_lower = if_else(date > ymd("2020-04-26"), 1067.0260, hpop_lower)) %>%
  select(-deaths, -upper, -lower) %>%
  mutate(deaths = hpop_deaths + lpop_deaths, upper = hpop_upper + lpop_upper, lower = hpop_lower + lpop
  select(date, deaths, upper, lower)

ggplot(il_final, mapping = aes(x = date)) +
  geom_line(il_final %>% slice(1:45), mapping = aes(y = deaths, color = "Deaths to Date")) +
  geom_ribbon(il_final %>% slice(46:92), mapping = aes(ymin = lower, ymax=upper, fill = "Projected Deatl
  geom_line(il_final %>% slice(46:92), mapping = aes(y = deaths, color = "Projected Deaths")) +
  labs(title = "Current and Projected COVID deaths through May 31", x = "Date", y = "Number of Deaths")
```