Homework 1: Estimating Covid-19 Deaths

Neeraj Sharma

4/7/2020

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)?
- You don't have to provide exact numbers, but discuss what you think the standard errors associated with your estimates might be, and your rationale for thinking those would be the standard errors.
- 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.

```
library(tidyverse)
library(readr)
library(modelr)
library(broom)
library(here)
library(lubridate)

# For security reasons, my personal API key is hidden. Permission to access Census/ACS data
# to reproduce my results can be granted here: https://api.census.gov/data/key_signup.html
library(tidycensus)
```

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

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 four sources spread across three general categories. The categories I analyze are:

1. COVID
    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
2. Demographics
    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 with people under 18
        f) Number of Households with people over 60
        g) Number of Households with more than 1 person per room on average (Census definition of overcrowded)
3. Hospitals
    i. County-level capacity, load and utilization data aggregated to the county level from the Illinois Health Facilities and Services Review Board

```
# Pulls in dataset produced by dataset_creat.R

full <- read_csv("20200412_combined_covid_demos_hosp.csv") %>%
  mutate(days_since_first_case = date - ymd("2020-01-24")) %>%
  mutate(days_since_sheltering = date - mdy("3/20/2020"))
```

```
# Hospital data
# https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0?selectedAtt
# https://www.chicagobusiness.com/static/section/hospital-beds-database.html

# mobility data
#https://www.google.com/covid19/mobility/
#https://github.com/vitorbaptista/google-covid19-mobility-reports
#https://ai.googleblog.com/2019/11/new-insights-into-human-mobility-with.html
#https://www.nature.com/articles/s41467-019-12809-y

# date implemented social distancing in each county maybe?
# https://www.finra.org/rules-guidance/key-topics/covid-19/shelter-in-place
```

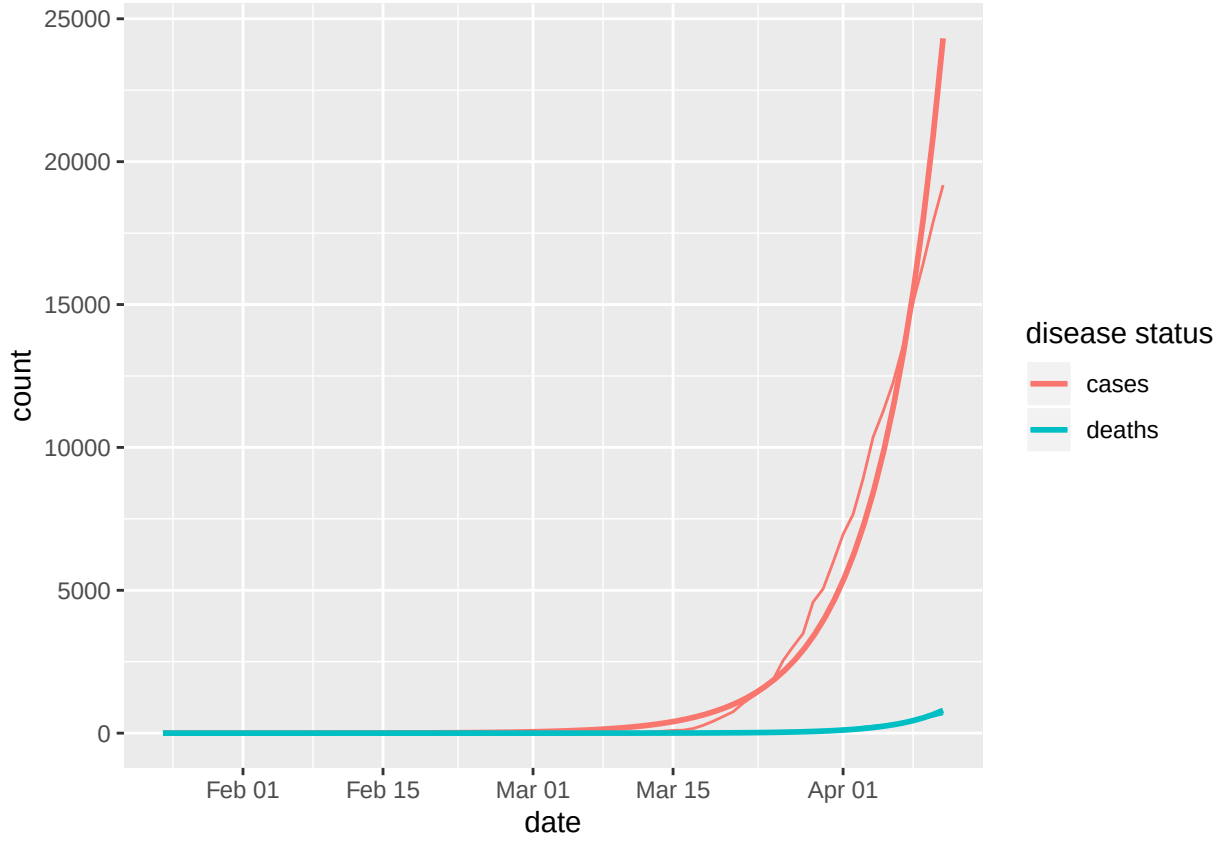
Describe the model(s) that you settled on for estimation. What was your logic for using that/those particular models?

```
regres <- glm(deaths ~ date, data = full, family = gaussian)
summary(regres)
```

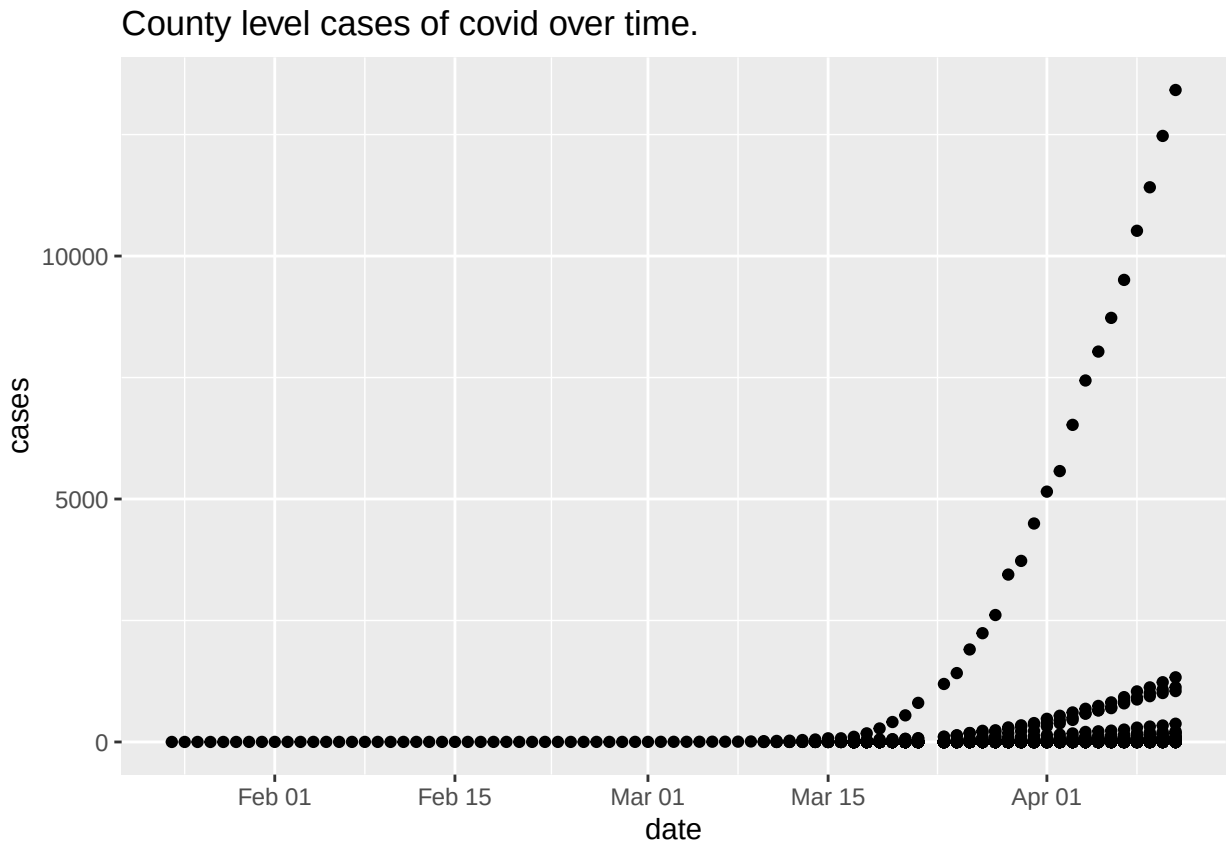
```
##
## Call:
## glm(formula = deaths ~ date, family = gaussian, data = full)
##
## Deviance Residuals:
```

```
##    Min      1Q  Median      3Q     Max
##  -3.70   -2.46   -1.53   -0.45  449.30
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.837e+03  6.454e+02  -4.396 1.15e-05 ***
## date         1.547e-01  3.517e-02   4.399 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 308.3144)
##
##     Null deviance: 823615  on 2653  degrees of freedom
## Residual deviance: 817650  on 2652  degrees of freedom
## AIC: 22746
##
## Number of Fisher Scoring iterations: 2
```

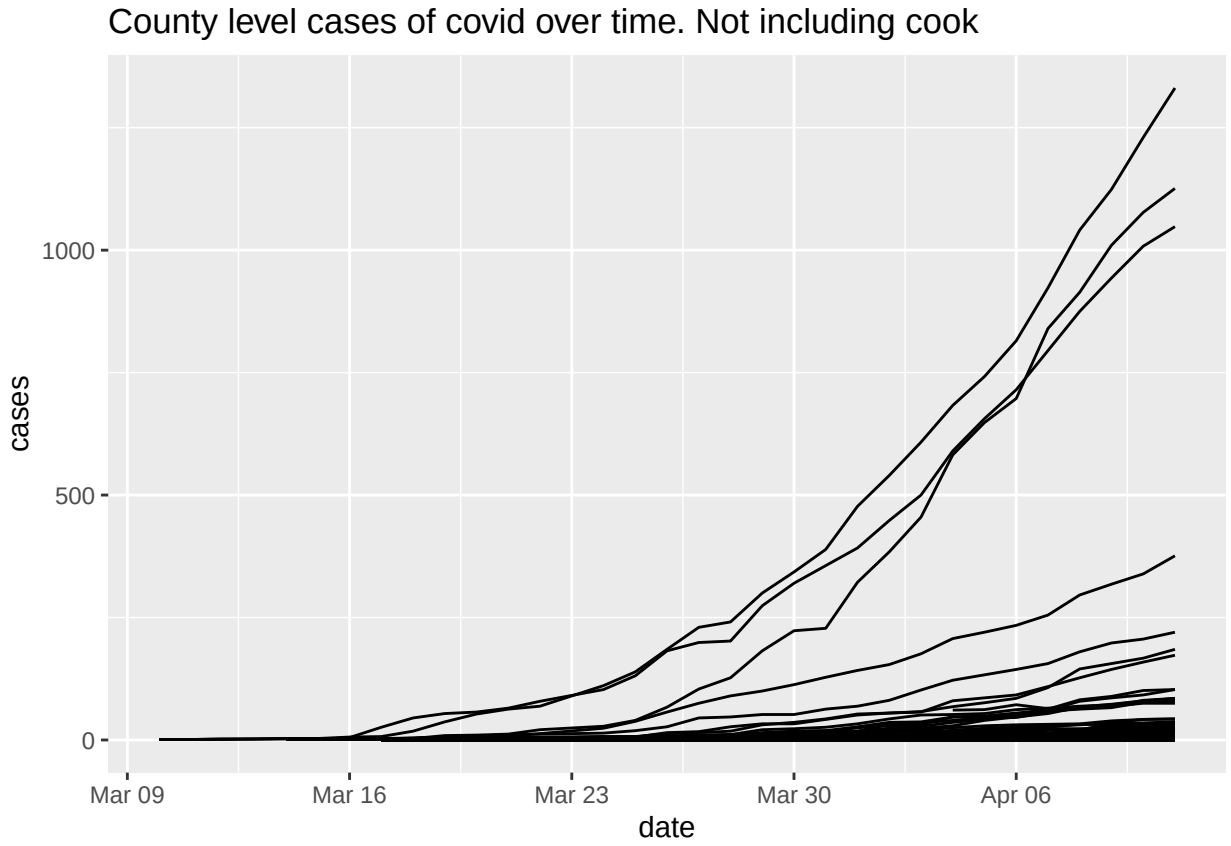
```r
ggplot(full %>% group_by(date) %>% summarise_at(c("cases", "deaths"), sum) %>% pivot_longer(-date, names
  geom_line() +
  geom_smooth(method = "glm",
              method.args = list(family = "poisson"),
              se = FALSE)
```




```r
ggplot(full, aes(x = date, y = cases, group = `county`)) +
  geom_point() +
  labs(title = "County level cases of covid over time.")
```




```
ggplot(full %>% filter(county != "Cook"), aes(x = date, y = cases, group = `county`)) +
  geom_path() +
  labs(title = "County level cases of covid over time. Not including cook")
```


County level cases of covid over time. Not including cook
cases
1000
500
0
Mar 09
Mar 16
Mar 23
Mar 30
Apr 06
date