# 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(curl)
library(modelr)
library(broom)
library(here)

# 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.

I draw on data provided by the Johns Hopkins Center for Systems Science and Engineering (CSSE). The CSSE provides detailed data on COVID cases in the United States all the way down to the City level. The CSSE stopped providing data for recoveries from COVID for the United States specifically as they found "no reliable data source reporting recovered cases for many countries, such as the US."[2]

```
# Pulls in the most recent version of COVID 19 cases and deaths from CSSE github repo.
# Static (April 9) versions of both of these datasets  are saved in "Raw Data."
cases_raw <- read_csv(curl("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid
deaths_raw <- read_csv(curl("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covi

# Create tidy DFs of COVID data for IL only
cases_counties <- cases_raw %>%
  filter(Province_State == "Illinois") %>%
  select(-UID, -iso2, -iso3, -code3, -FIPS, -Province_State, -Country_Region, -Lat, -Long_) %>%
  pivot_longer(c(-Admin2, -Combined_Key), names_to = "date", values_to = "cases") %>%
  mutate(date = as.Date(date, "%m/%d/%y"))

deaths_counties <- deaths_raw %>%
  filter(Province_State == "Illinois") %>%
  select(-UID, -iso2, -iso3, -code3, -FIPS, -Province_State, -Country_Region, -Lat, -Long_) %>%
  pivot_longer(c(-Admin2, -Combined_Key, -Population), names_to = "date", values_to = "deaths") %>%
  mutate(date = as.Date(date, "%m/%d/%y"))

# Summarizes data to reflect statewide trends
cases_IL<- cases_counties %>%
  group_by(date) %>%
  summarize(cases = sum(cases))

deaths_IL <- deaths_counties %>%
  group_by(date) %>%
  summarize(deaths = sum(deaths))

# Group together cases and deaths
cases_deaths_IL <- left_join(cases_IL, deaths_IL) %>%
  pivot_longer(c(cases, deaths), names_to = "disease status", values_to = "count")
```

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

ACS_vars_18 <- load_variables(2018, "acs5", cache = TRUE)
# Collect demographic data on IL counties.
get_acs(geography = "county",
        variables = c(medincome = "B19013_001"),
        year = 2018, state = "Illinois")
```

```
## Getting data from the 2014-2018 5-year ACS
```

```
## # A tibble: 102 x 5
##    GEOID NAME                      variable  estimate   moe
##    <chr> <chr>                     <chr>        <dbl> <dbl>
```

[2] https://github.com/CSSEGISandData/COVID-19/issues/1250#issuecomment-602271179

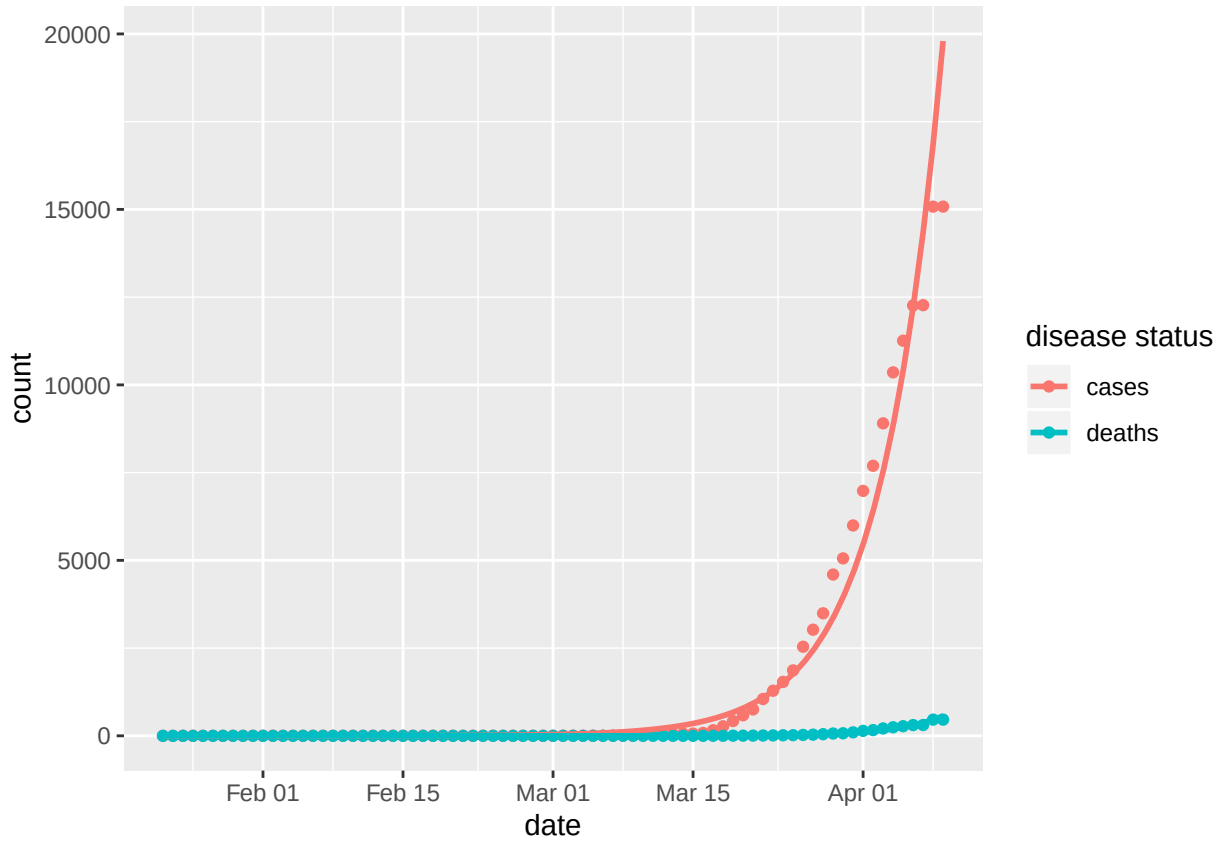
```
##  1 17001 Adams County, Illinois     medincome    51398  1583
##  2 17003 Alexander County, Illinois medincome    34709  3017
##  3 17005 Bond County, Illinois      medincome    58097  5803
##  4 17007 Boone County, Illinois     medincome    66898  4955
##  5 17009 Brown County, Illinois     medincome    58762  2302
##  6 17011 Bureau County, Illinois    medincome    55940  2114
##  7 17013 Calhoun County, Illinois   medincome    54392  6158
##  8 17015 Carroll County, Illinois   medincome    51228  3421
##  9 17017 Cass County, Illinois      medincome    51997  2526
## 10 17019 Champaign County, Illinois medincome    51692  1225
## # ... with 92 more rows
```

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

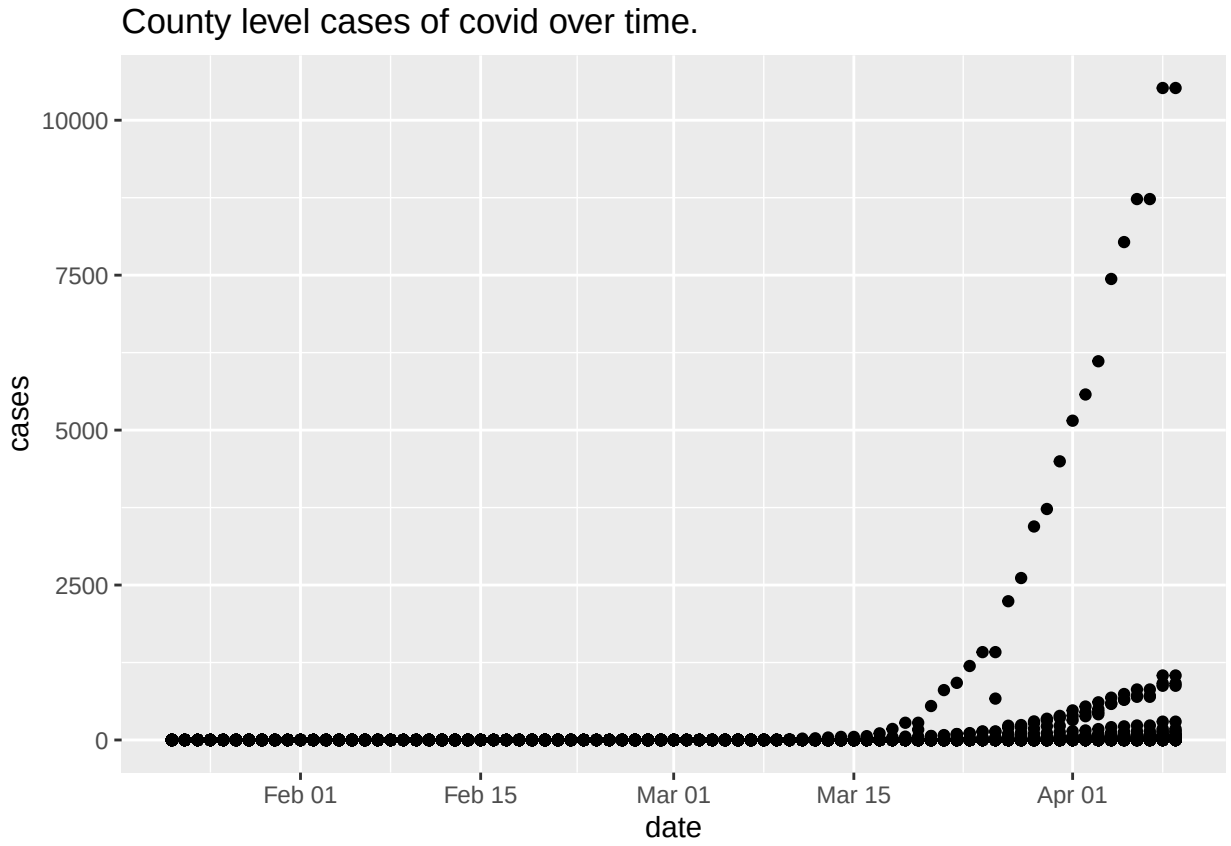
```
regres <- glm(deaths ~ date, data = deaths_IL, family = gaussian)
summary(regres)
```

```
##
## Call:
## glm(formula = deaths ~ date, family = gaussian, data = deaths_IL)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -82.35  -56.78  -14.61   31.62  328.18
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.624e+04  7.215e+03  -6.409 1.07e-08 ***
## date         2.526e+00  3.938e-01   6.415 1.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 6369.396)
##
##     Null deviance: 752521  on 78  degrees of freedom
## Residual deviance: 490444  on 77  degrees of freedom
## AIC: 920.15
##
## Number of Fisher Scoring iterations: 2
```

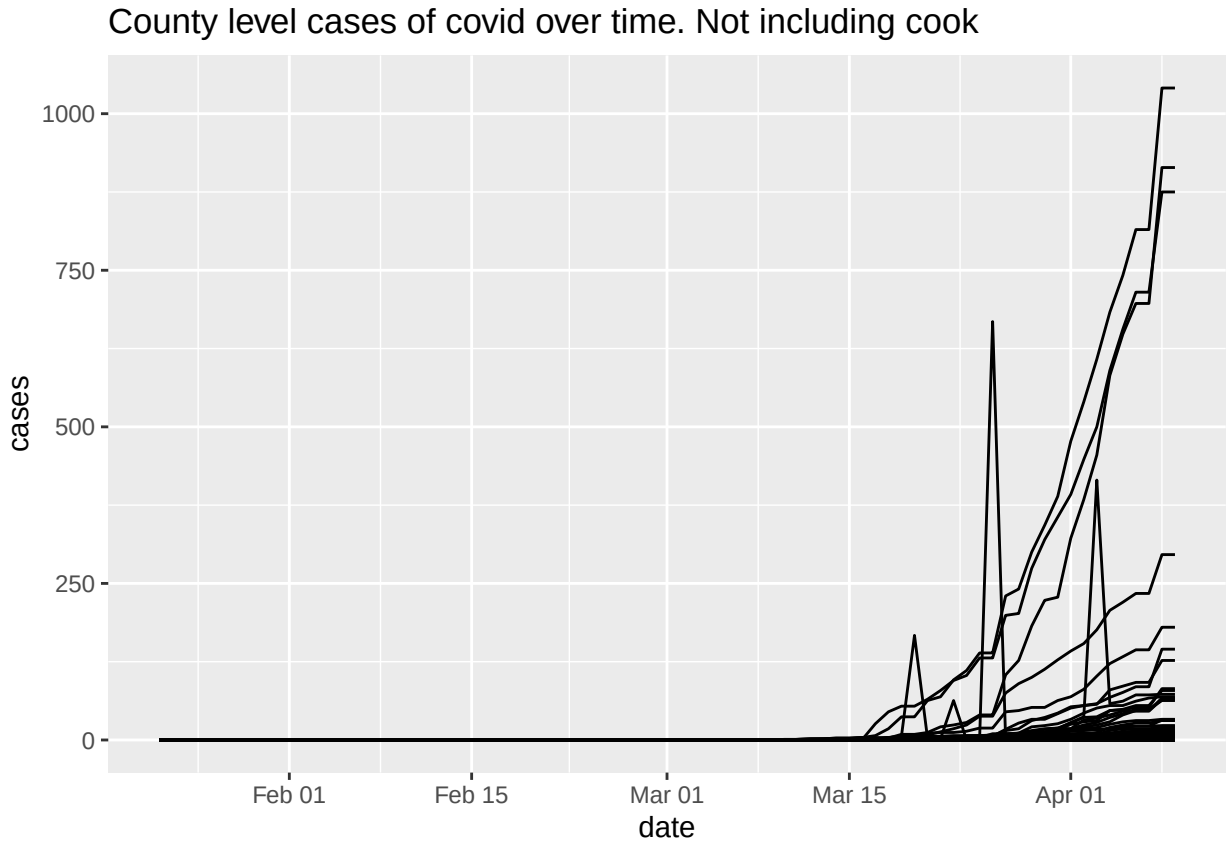
```
ggplot(cases_deaths_IL, aes(x = date, y = count, group = `disease status`, color = `disease status`)) +
  geom_point() +
  geom_smooth(method = "glm",
              method.args = list(family = "poisson"),
              se = FALSE)
```




```
ggplot(cases_counties, aes(x = date, y = cases, group = `Admin2`)) +
  geom_point() +
  labs(title = "County level cases of covid over time.")
```




```
ggplot(cases_counties %>% filter(Admin2 != "Cook"), aes(x = date, y = cases, group = `Admin2`)) +
  geom_path() +
  labs(title = "County level cases of covid over time. Not including cook")
```


County level cases of covid over time. Not including cook
1000
750
500
250
0
cases
Feb 01
Feb 15
Mar 01
Mar 15
Apr 01
date