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)?
- 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.

```
library(tidyverse)
library(readr)
library(modelr)
library(curl)
library(broom)
library(here)
library(lubridate)
library(fable)
library(tsibble)
```

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

```r
# 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)
```

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
        f) Number of Households with people under 18
        g) Number of Households with people over 60
        h) Number of Overcrowded Households
3. Hospitals
    i. County-level capacity, load and utilization data aggregated to the county level from the Illinois Health Facilities and Services Review Board
4. Mobility
    i. Median distance traveled by cellphone pings.

```r
# Pulls in dataset produced by dataset_creat.R
full <- read_csv("20200414_combined_covid_demos_hosp_mobility.csv") %>%
  mutate(per_capita_deaths = deaths/`population-E`,
         `socialdistancing?` = if_else(date > as.Date("2020-03-20"), 1, 0),
         overcrowding = `num_hh_morethan1_person_perroom-E`/`num_hh-E`,
         at_icu_cap = if_else(cases > bedsIntensiveCare, 1, 0),
         at_hospital_cap = if_else(cases > bedsTotalCONAuthorizedBeds, 1, 0),
         delta_cases = if_else(is.na(lag(county)) | county != lag(county),
                               0, as.double(cases - lag(cases))),
         delta_deaths = if_else(is.na(lag(county)) | county != lag(county),
                                0, as.double(deaths - lag(deaths))),
         cases_rate = if_else(is.na(lag(county)) | county != lag(county),
                              0, as.double(delta_cases/lag(delta_cases))),
         deaths_rate = if_else(is.na(lag(county)) | county != lag(county),
                               0, as.double(delta_deaths/lag(delta_deaths))),
         days_since_first_case = date - ymd("2020-01-24"),
         days_since_first_death = date - ymd("2020-03-17"),
         days_since_sheltering = date - mdy("3/20/2020"))

# Note: April 21, 2020 is 88 days after the first IL COVID case was recorded on 2020-01-24
# Note: May 31, 2020 is is 128 days after the first IL COVID case was recorded on 2020-01-24
```

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

The model I ultimately settled on relates the log of deaths at a given time to the number of cases at that time, the date, a colinearly-related social distancing binary variable and median mobility, and a binary variable indicating if the hospital ICU cap has been reached yet. I settled on this model as I believe it accurately captures an important behavioral change that occured part-ways through the COVID pandemic that is difficult to account for while also including for an important motivator behind deaths in certain instances. Accounting for social distancing and the corresponding mobility reduction through including them as colinearly-related variables is justified as mobility remains relatively steady depending on if social distancing is in place or not. Including these variables in my model is a straightforward, 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 independently 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.

In order to turn my 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. 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.

```
apr <- full %>%
  select(date, county, cases, deaths, delta_cases, delta_deaths, cases_rate,
         deaths_rate, days_since_first_case, days_since_first_death, days_since_sheltering,
         `num_hh-E`, `num_hh_morethan1_person_perroom-E`, bedsIntensiveCare,
         bedsTotalCONAuthorizedBeds, location_samples, m50, `socialdistancing?`, overcrowding, at_icu_ca
         at_hospital_cap) %>%
  filter(deaths != 0) %>%
  as_tsibble(index = date, key = county)

lm.weight.function <- function(x) {10 / (1 + exp(-x))}
lm.weights <- lapply(seq(-13, 14, length.out = 27), lm.weight.function)


regres1 <- model(apr %>% index_by(date) %>% summarize(deaths = sum(deaths)), lm = ETS(deaths ~ trend("N
```

```
## Warning: 1 error encountered for lm
## [1] .data contains implicit gaps in time. You should check your data and convert implicit gaps into
```

```
regres2 <- model(apr %>% filter(!(is.na(m50))), lm = TSLM(log(deaths) ~ cases + `socialdistancing?`*dat
```

```
## Warning: 3 errors (1 unique) encountered for lm
## [3] 0 (non-NA) cases
```

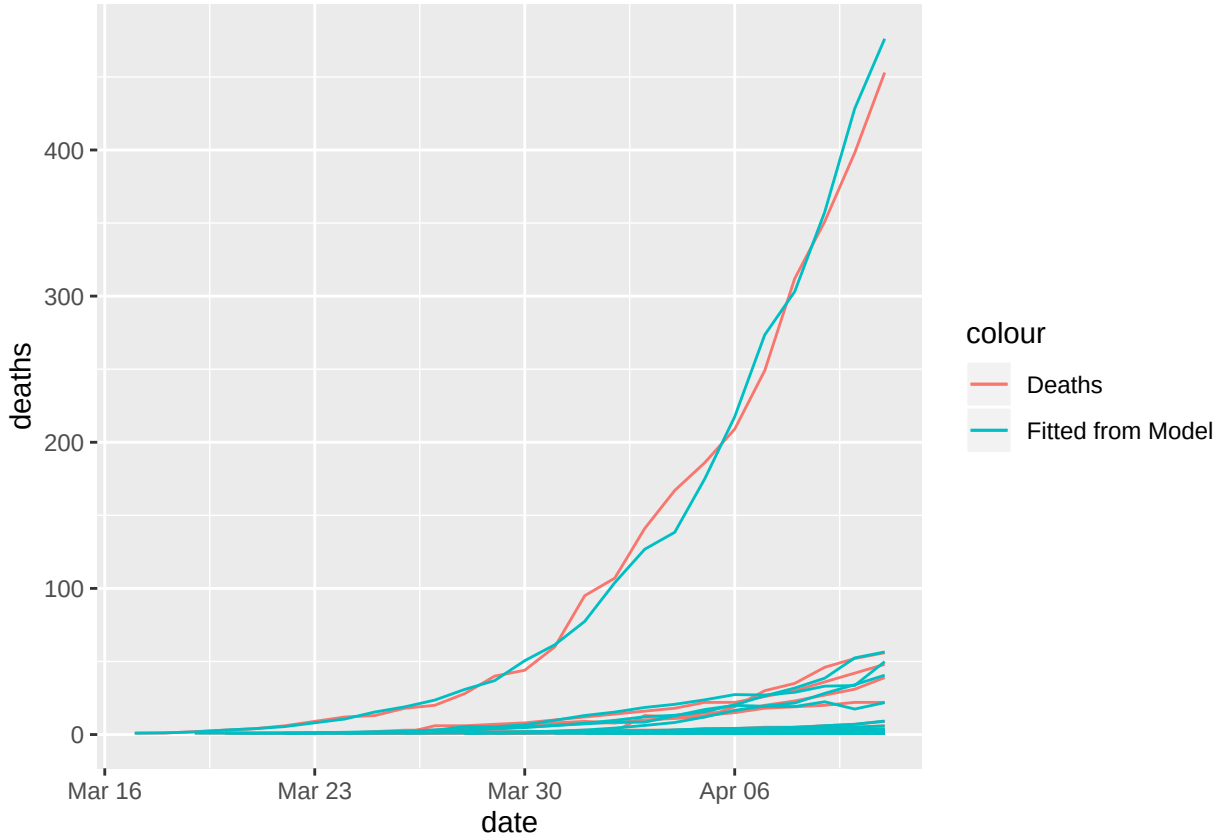
```
# going forward to plug new_data into forecast, date is just date, socialdistancing? is 1, m50 will be

regres2 %>% filter(county == "Cook") %>% report()
```

```
## Series: deaths
## Model: TSLM
## Transformation: log(.x)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16956 -0.07334 -0.01829  0.07963  0.20464
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -7.969e+03  1.036e+03  -7.695 4.25e-07 ***
## cases                     -2.330e-04  5.158e-05  -4.516 0.000267 ***
## `socialdistancing?`        1.096e+03  1.203e+03   0.912 0.374076
## date                       4.345e-01  5.646e-02   7.695 4.24e-07 ***
## m50                        1.941e-01  1.298e-01   1.496 0.152106
## at_icu_cap                -9.127e-02  1.351e-01  -0.675 0.507985
## `socialdistancing?`:date -5.975e-02  6.557e-02  -0.911 0.374199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1246 on 18 degrees of freedom
## Multiple R-squared: 0.9969,  Adjusted R-squared: 0.9958
## F-statistic: 959.1 on 6 and 18 DF, p-value: < 2.22e-16
```

```
ggplot(augment(regres2), aes(x = date, group = county)) +
  geom_line(aes(y = deaths, color = "Deaths")) +
  geom_line(aes(y = .fitted, color = "Fitted from Model"))
```

```
## Warning: Removed 30 rows containing missing values (geom_path).
```

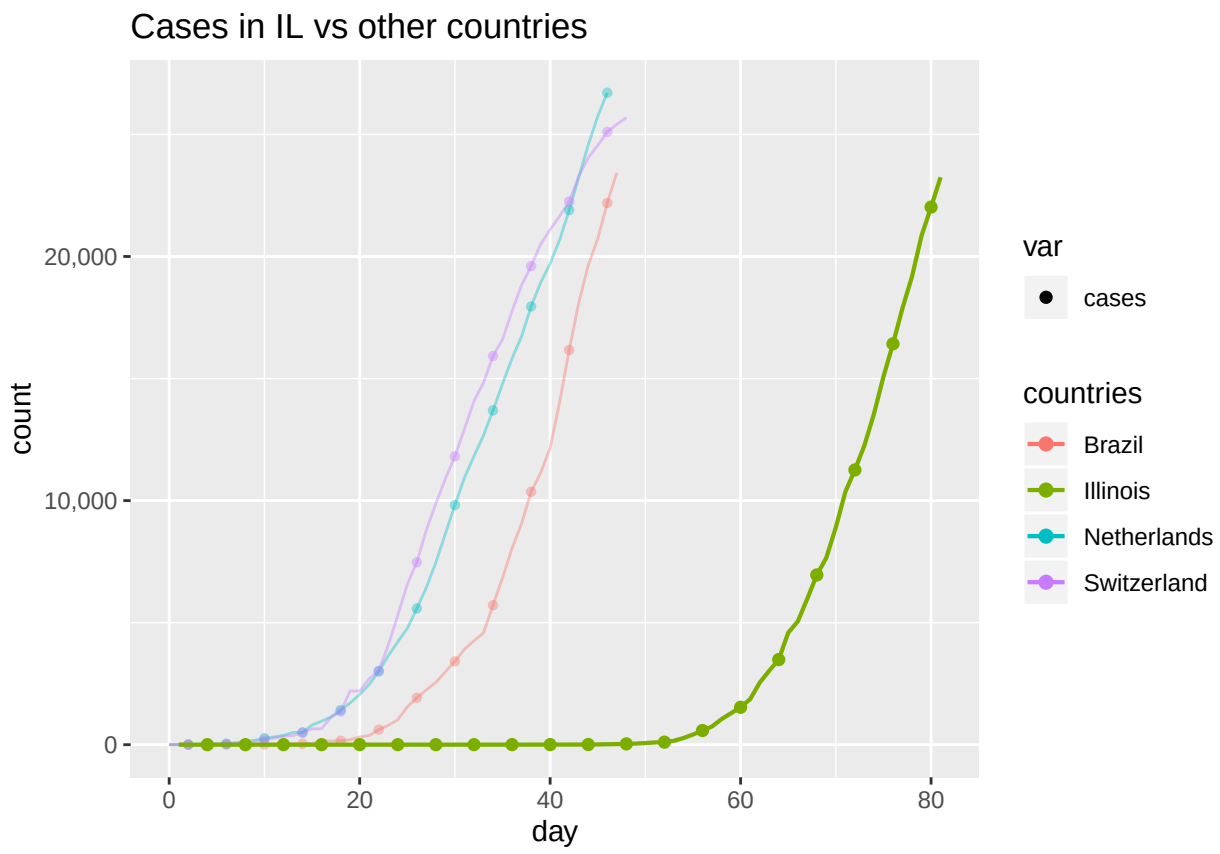



```
global_cases_raw <- read_csv(curl("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse
global_deaths_raw <- read_csv(curl("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/css
```

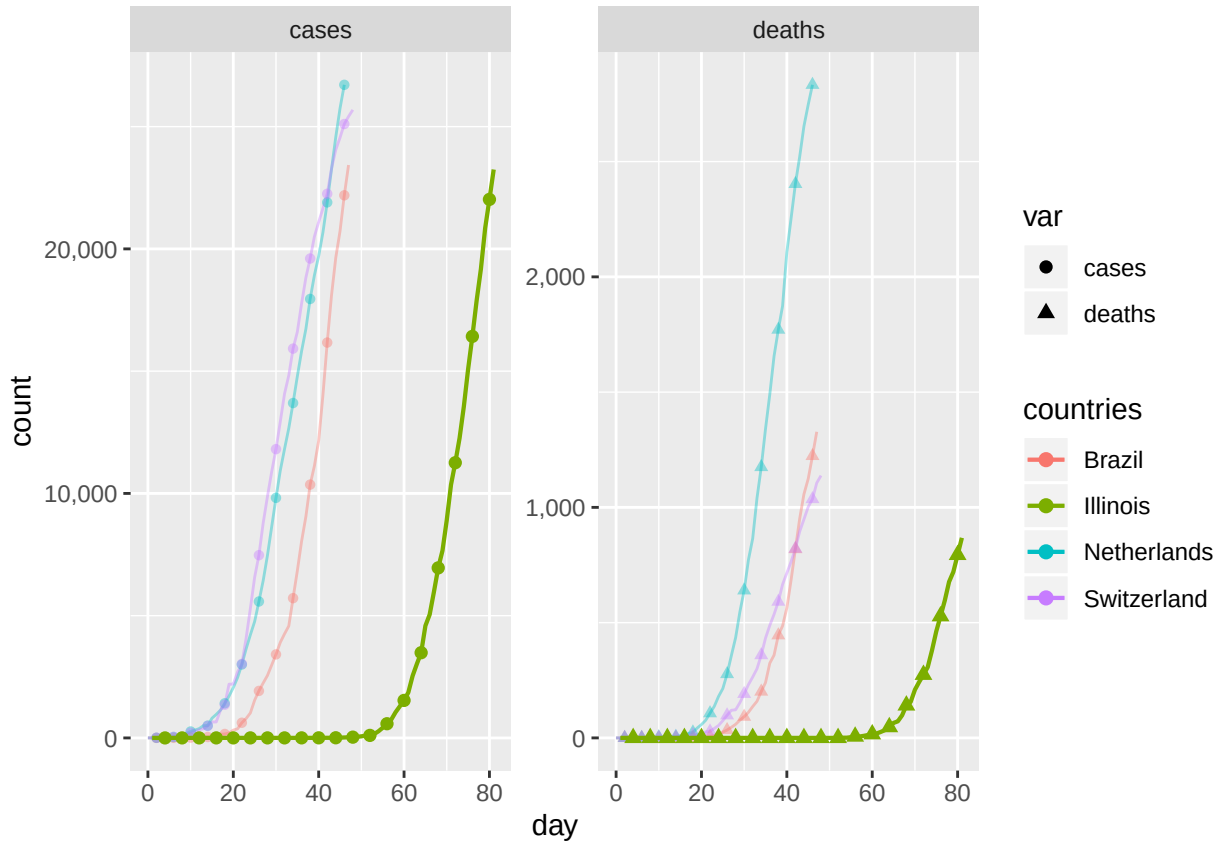
```r
first_reported_cases <- global_cases_raw %>%
  select(-`Province/State`, -Lat, -Long) %>%
  pivot_longer(-`Country/Region`, names_to = "date", values_to = "cases") %>%
  mutate(first_case_date = as.Date(date, "%m/%d/%y")) %>%
  group_by(first_case_date, `Country/Region`) %>%
  summarize(cases = sum(cases)) %>%
  arrange(-desc(`Country/Region`)) %>%
  ungroup() %>%
  filter((cases != 0 & lag(cases) == 0 & cases <= lead(cases)) | (first_case_date == as.Date("2020-01-2
  select(-cases, country = `Country/Region`)

il <- full %>%
  select(date, cases, deaths, delta_cases, delta_deaths, `socialdistancing?`) %>%
  group_by(date, `socialdistancing?`) %>%
  summarize(deaths = sum(deaths), cases = sum(cases)) %>%
  ungroup() %>%
  filter(cases != 0) %>%
  select(-`socialdistancing?`) %>%
  pivot_longer(-date, names_to = "var", values_to = "count")  %>%
  mutate(day = as.double(date - ymd("2020-01-24"))) %>%
  filter(day > 0)
```

```r
country <- c("Brazil", "Netherlands", "Switzerland")
# Growth curves for Germany, France, Iran(debatable), Italy, Korea, South, Spain are too steep

versus <- left_join(global_cases_raw %>%
                      filter(`Country/Region` %in% country) %>%
                      select(-`Province/State`, -Lat, -Long) %>%
                      pivot_longer(-`Country/Region`, names_to = "date", values_to = "cases") %>%
                      group_by(`Country/Region`, date) %>%
                      summarize(cases = sum(cases)),
                    global_deaths_raw %>%
                      filter(`Country/Region` %in% country) %>%
                      select(-`Province/State`, -Lat, -Long) %>%
                      pivot_longer(-`Country/Region`, names_to = "date", values_to = "deaths") %>%
                      group_by(`Country/Region`, date) %>%
                      summarize(deaths = sum(deaths))) %>%
  filter(cases != 0) %>%
  pivot_longer(-c(`Country/Region`, date), names_to = "var", values_to = "count") %>%
  mutate(date = as.Date(date, "%m/%d/%y")) %>%
  left_join(first_reported_cases, by = c("Country/Region" = "country")) %>%
  mutate(day = as.double(date - first_case_date)) %>%
  rename(countries = `Country/Region`) %>%
  group_by(countries, date, var, day) %>%
  summarize(count = sum(count))

forecast_newdata <- tsibble(date = seq(ymd("2020-04-14"), ymd("2020-04-21"), by = "1 day"), `socialdista

ggplot() +
  geom_line(versus %>% filter(var == "cases"), mapping = aes(x = day, y = count, group = interaction(cou
  geom_point(versus %>% filter(day %% 4 == 2, var == "cases"), mapping = aes(x = day, y = count, group =
  geom_line(il %>% filter(var == "cases"), mapping = aes(x = day, y = count, group = var, color = "Illi
  geom_point(il %>% filter(day %% 4 == 0, var == "cases"), mapping = aes(x = day, y = count, group = va
```

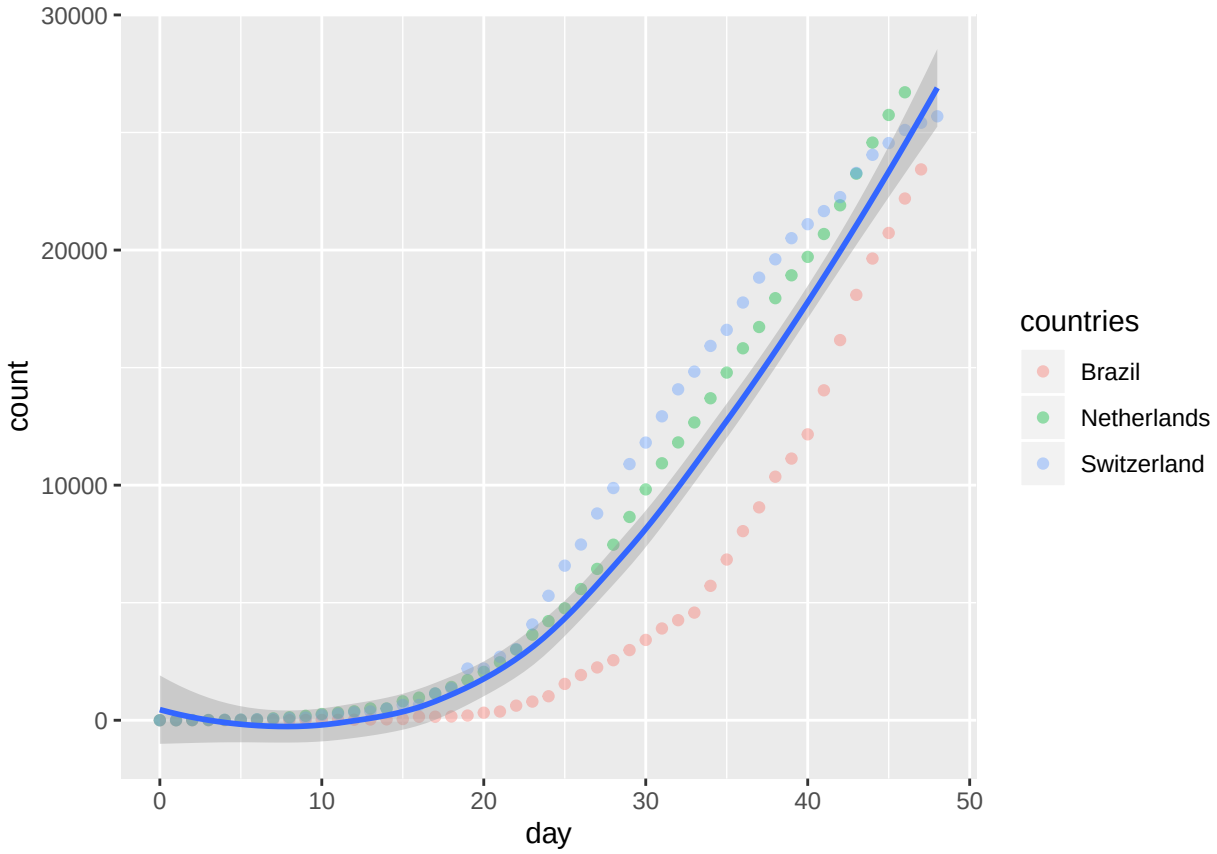
```
  scale_y_continuous(labels = scales::comma) +
  labs(title = "Cases in IL vs other countries")
```




```
ggplot() +
  geom_line(versus, mapping = aes(x = day, y = count, group = interaction(countries, var), color = coun
  geom_point(versus %>% filter(day %% 4 == 2), mapping = aes(x = day, y = count, group = interaction(co
  geom_line(il, mapping = aes(x = day, y = count, group = var, color = "Illinois"), size = .75) +
  geom_point(il %>% filter(day %% 4 == 0), mapping = aes(x = day, y = count, group = var, color = "Illi
  scale_y_continuous(labels = scales::comma) +
  facet_wrap(~var, scales = "free")
```




```
# geom_line(apr_highpop_forecast, mapping = aes(x = day, y = deaths, color = "prediction"))


ggplot() +
  geom_point(versus %>% filter(var == "cases"), mapping = aes(x = day, y = count, group = interaction(co
  geom_smooth(versus %>% filter(var == "cases"), mapping = aes(x = day, y = count), alpha = 0.4)
```



A Chicago resident contracted the first confirmed case of COVID-19 in Illinois. The Chicago case was announced Jan. 24, and the first cases outside of Chicago and Cook County were reported March 11, in Kane and McHenry counties. Governor Pritzker issued a disaster proclamation on March 9, which provides the state access to federal and state resources to mitigate the spread of COVID-19.

On Jan 24, IL had it's first COVID patients. Those two inviduals had literally recovered by the time another patient was diognosed in IL. I remove the days that there were those two random early stage cases and instead start my day 0 for Illinois on Feburary 29, 2020. On Feburary 29, 2020, the first COVID patient in IL not affiliated with the two early cases was diognosed.[2] I count that as the real first COVID case in IL and thus make it in the initial start date for IL.

Additionally, I also move Russia's curve to begin on March 2, 2020. Prior to that date, the only cases in Russia were Chinese Nationals who had recently entered Russia from China and eight Russians from the Diamond Princess Cruise Ship who were immediatly quarentined. All these individuals recovered prior to another case appearing in Russia. On March 2nd, a man was diognosed with COVID after vacationing in Italy, so I mark that as the first Russian COVID case.

[2]https://wgntv.com/news/illinois-health-officials-announce-new-case-of-coronavirus/