Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-04-05

Contents

Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a new type of coronavirus: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak first started in Wuhan, China in December 2019. The first kown case of COVID-19 in the U.S. was confirmed on January 20, 2020, in a 35-year-old man who teturned to Washington State on January 15 after traveling to Wuhan. Starting around the end of Feburary, evidence emerge for community spread in the US.

We, as all of us, are indebted to the heros who fight COVID-19 across the whole world in different ways. For this data exploration, I am grateful to many data science groups who have collected detailed COVID-19 outbreak data, including the number of tests, confirmed cases, and deaths, across countries/regions, states/provnices (administrative division level 1, or admin1), and counties (admin2). Specifically, I used the data from these three resources:

- JHU (https://coronavirus.jhu.edu/)
    - The Center for Systems Science and Engineering (CSSE) at John Hopkins University.
    - World-wide counts of coronavirus cases, deaths, and recovered ones.
    - https://github.com/CSSEGISandData/COVID-19

- NY Times (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html)
    - The New York Times
    - "cumulative counts of coronavirus cases in the United States, at the state and county level, over time"
    - https://github.com/nytimes/covid-19-data

- COVID Trackng (https://covidtracking.com/)
  - COVID Tracking Project
  - "collects information from 50 US states, the District of Columbia, and 5 other US territories to provide the most comprehensive testing data"
 - https://github.com/COVID19Tracking/covid-tracking-data

JHU

Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

Here is the list of 10 records with the largest number of cases or deaths on the most recent date.
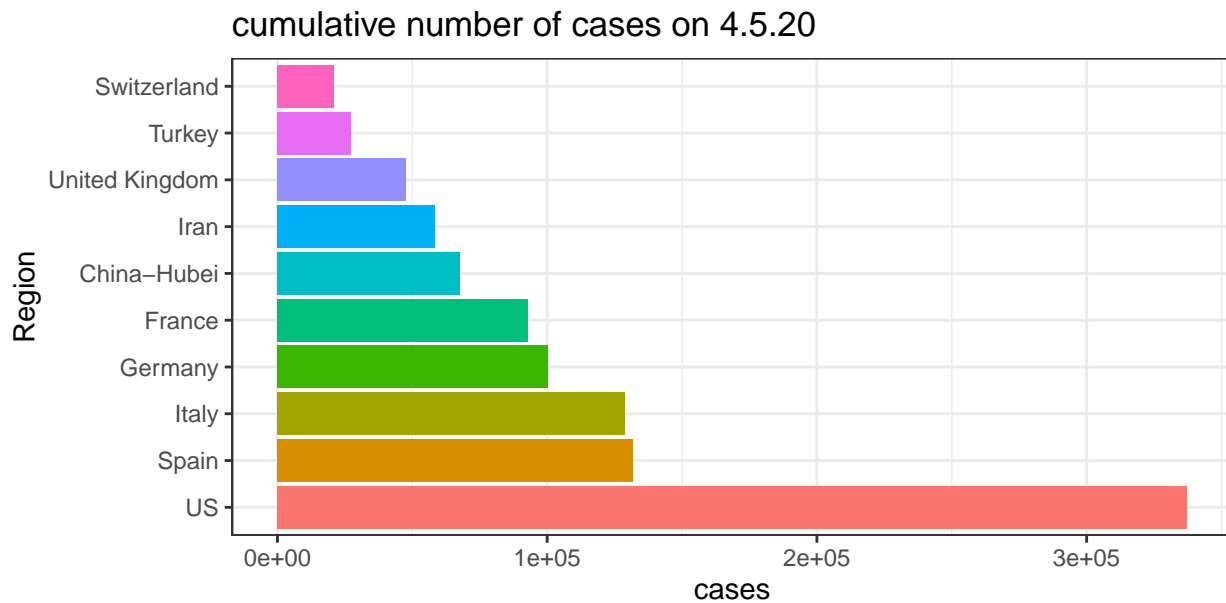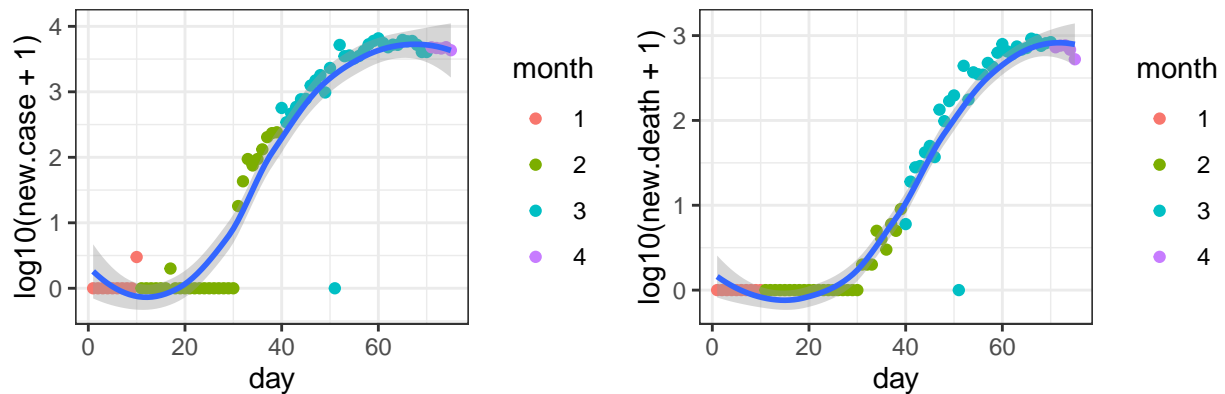
cumulative number of deaths on 4.5.20
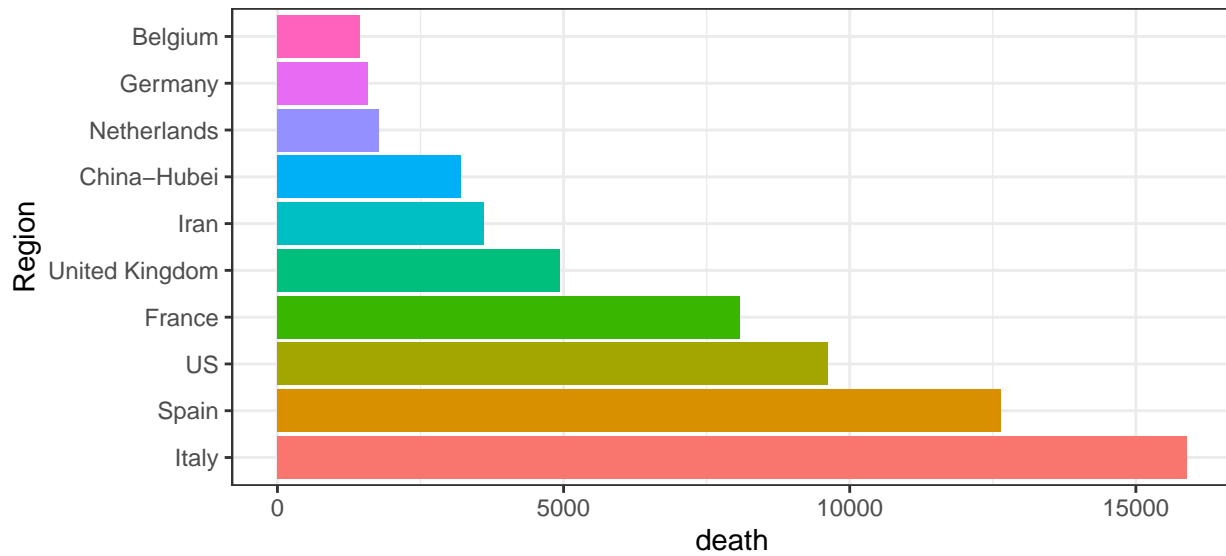
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.
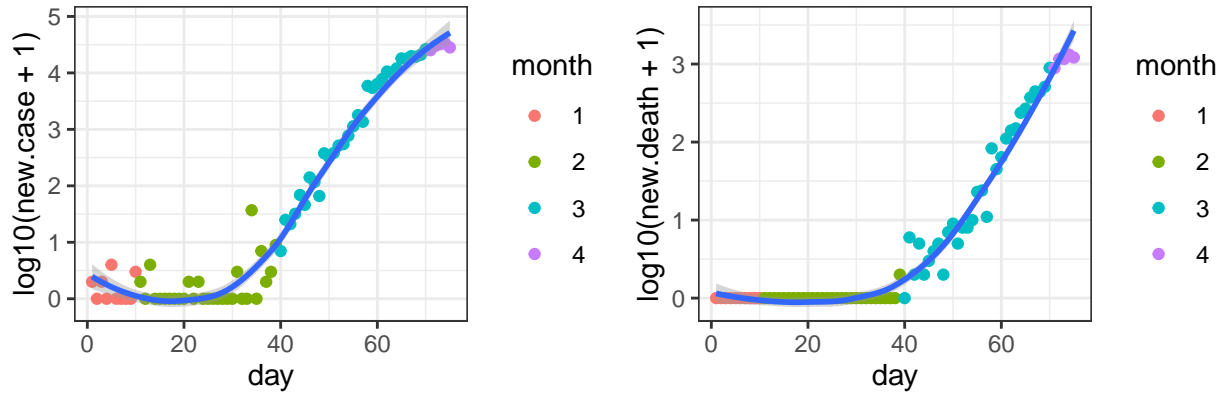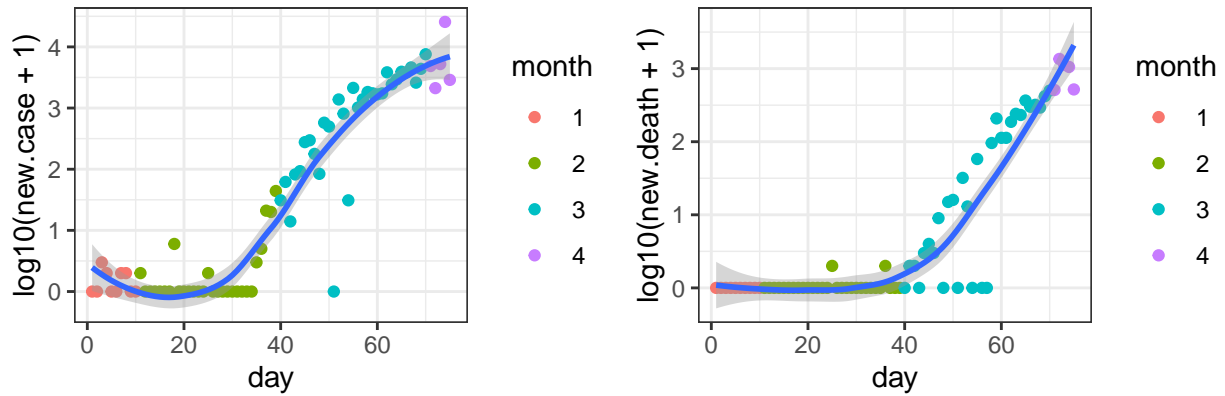
Italy

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020
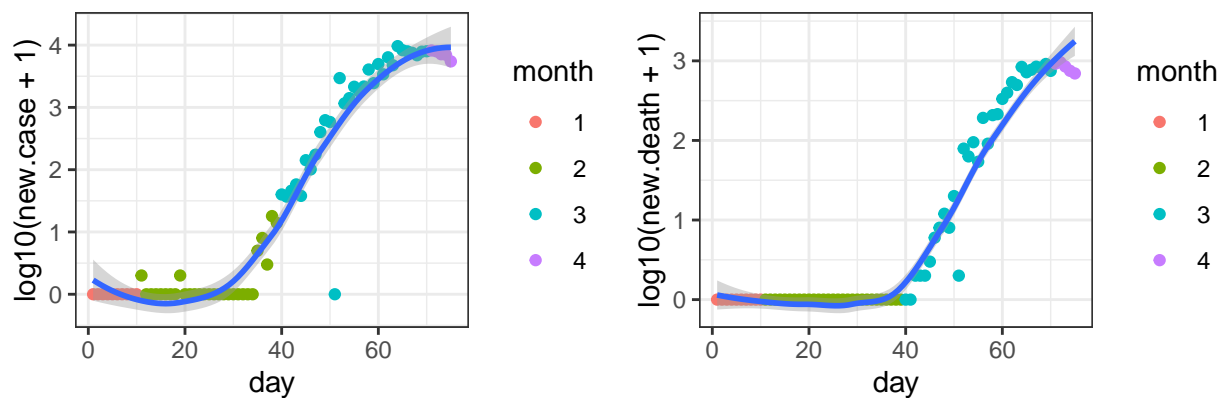
Spain

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

US

France

United Kingdom
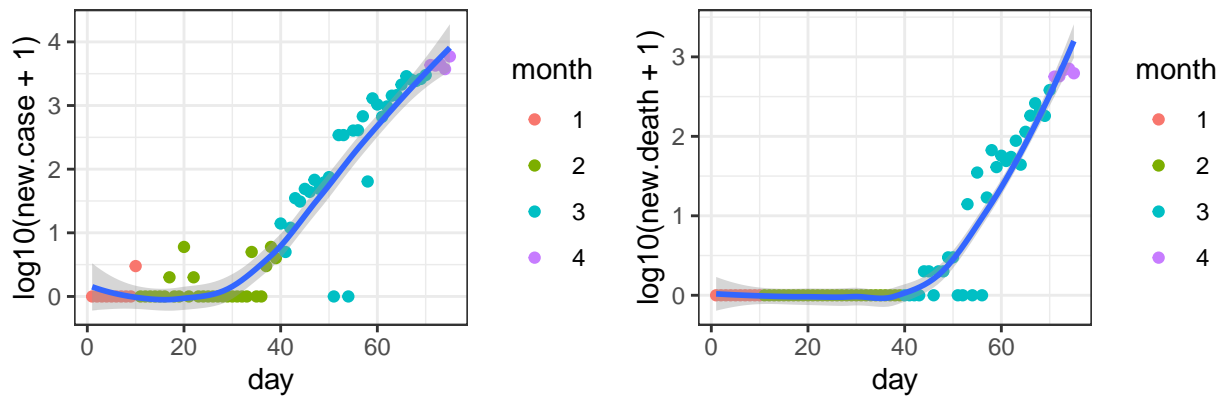
Iran
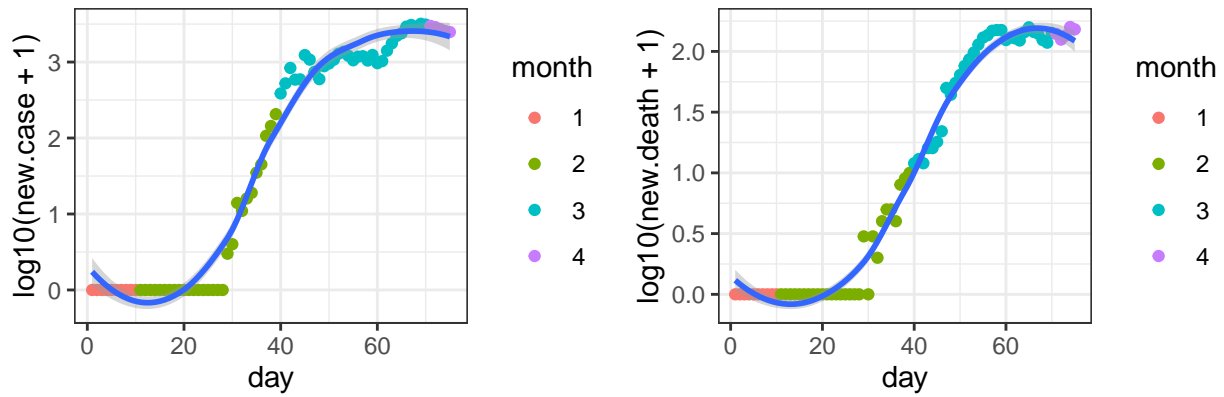
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

China-Hubei
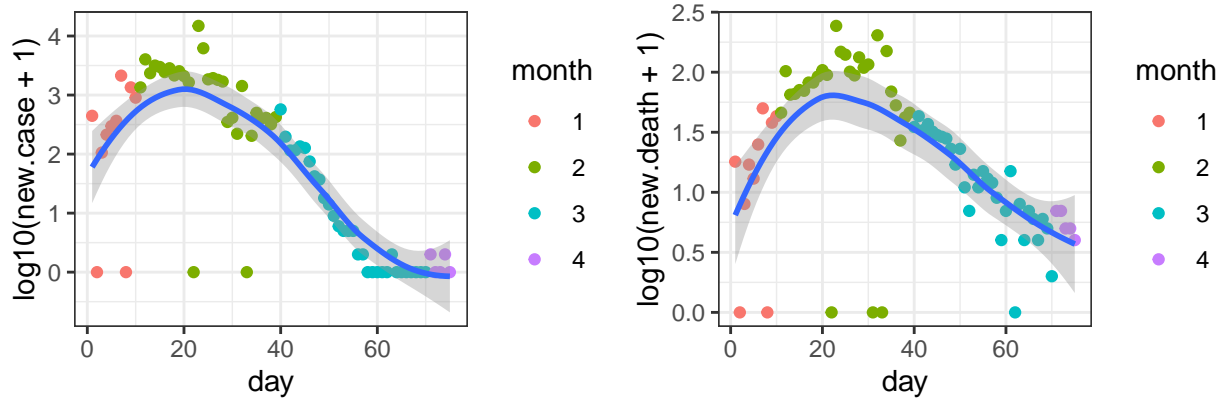
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

Netherlands

data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

Germany

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium
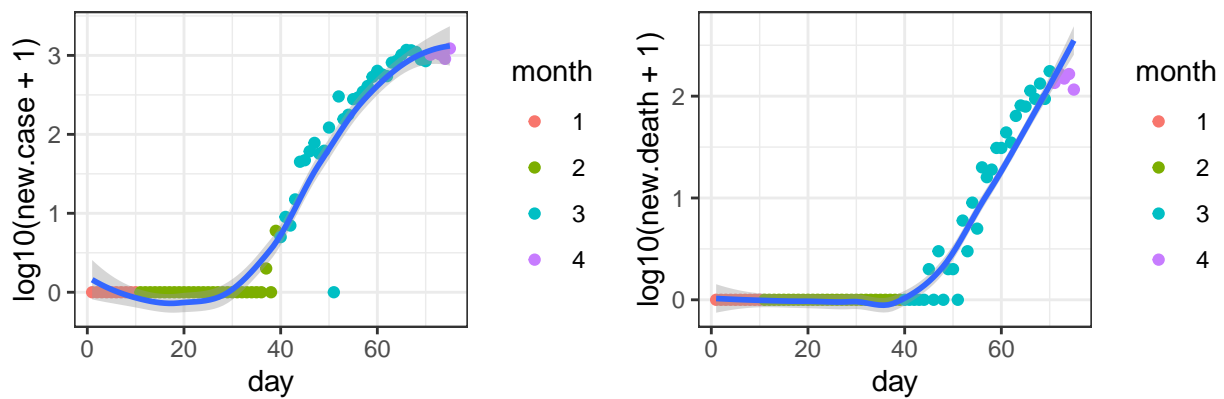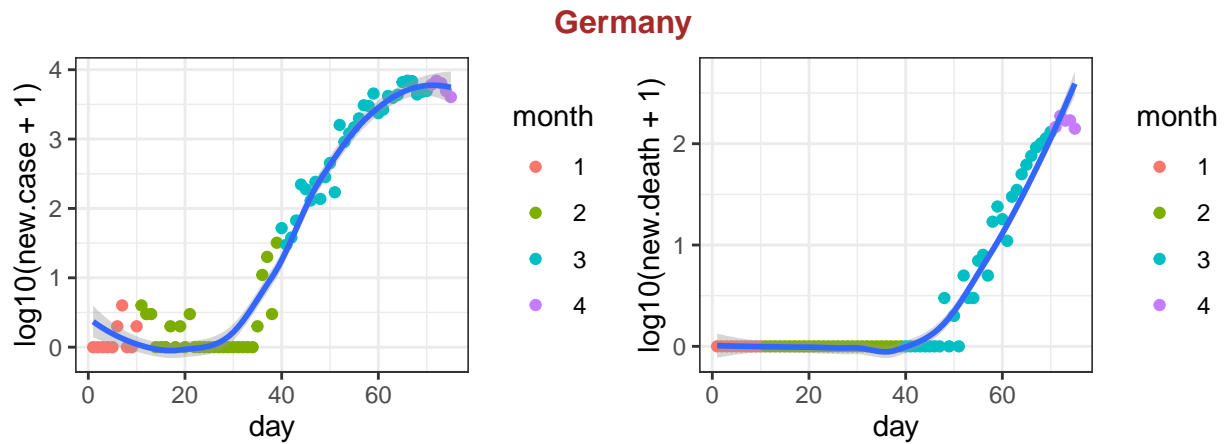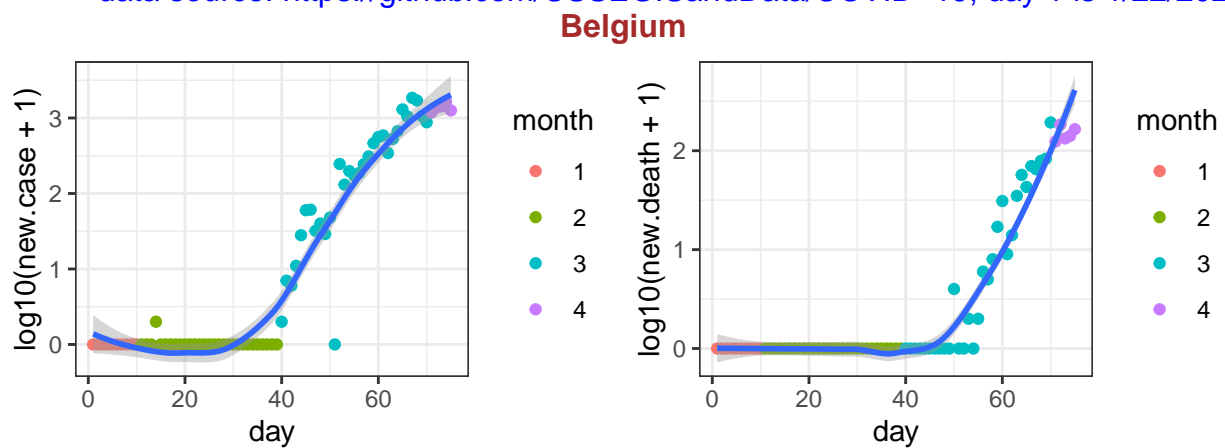
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) inlcude information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for informatoin of individual states or counties.

number of records in Hopkins daily reports
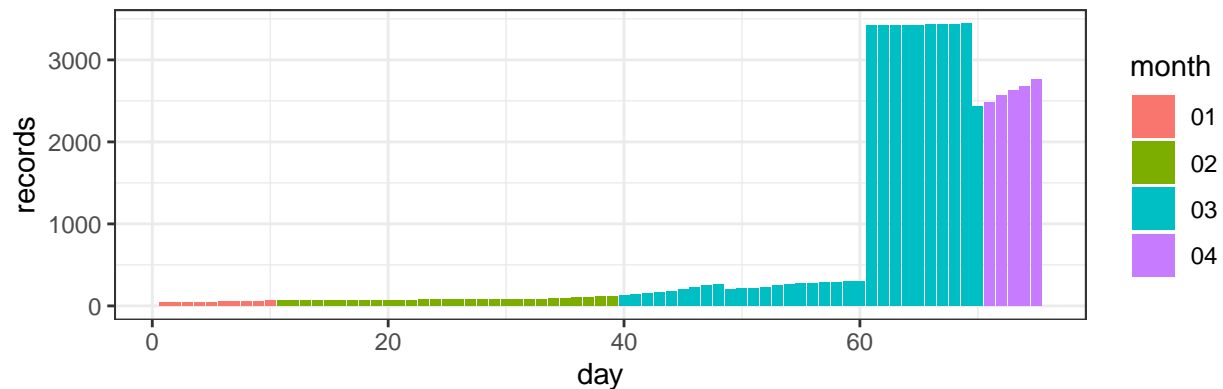
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

NY Times

The data from NY Times are saved in two text files, one for state level information and the other one for county level information.
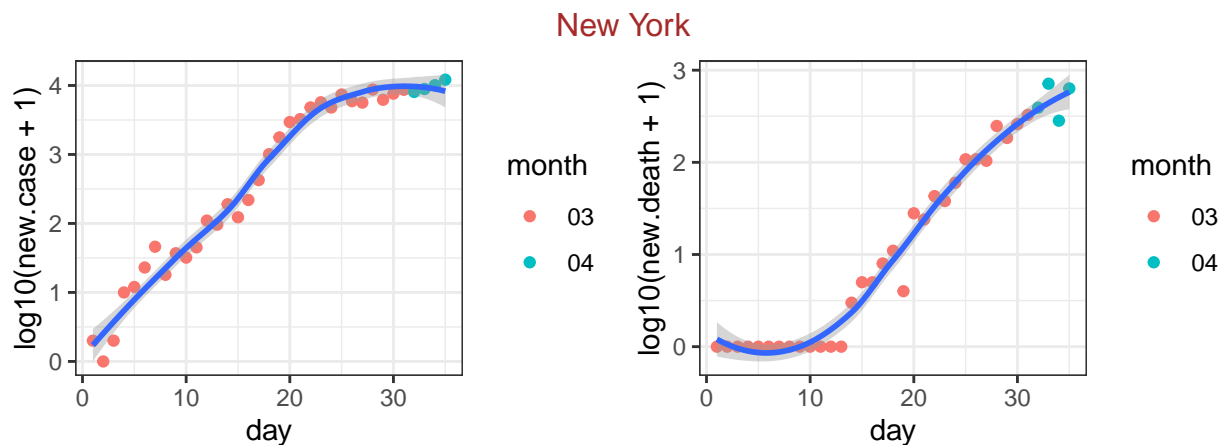
The currente date is

[1] "2020-04-04"

state level data

First check the 10 states with the largest number of deaths.
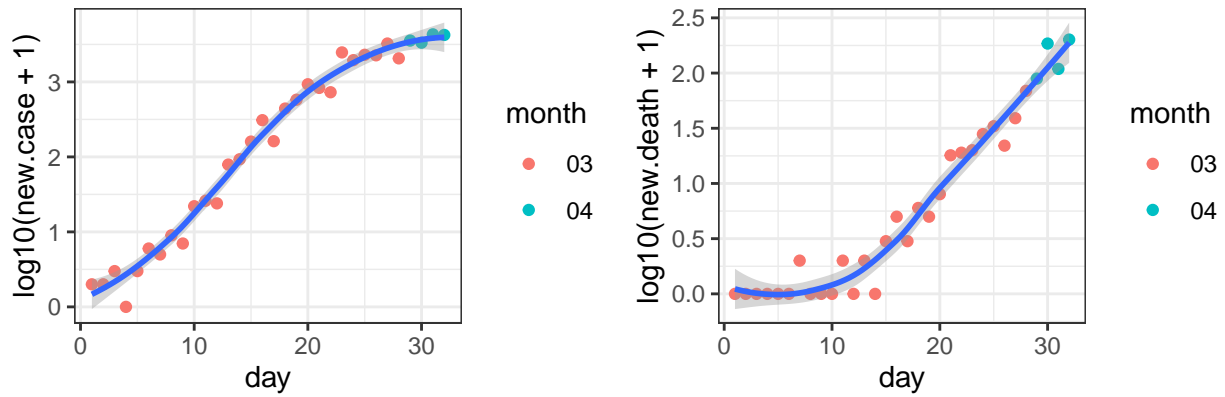
```
##              date         state fips  cases deaths
## 1808 2020-04-04      New York   36 114996   3568
## 1806 2020-04-04    New Jersey   34  34124    847
## 1798 2020-04-04      Michigan   26  14225    540
## 1794 2020-04-04     Louisiana   22  12496    412
## 1779 2020-04-04    California    6  13796    323
## 1826 2020-04-04    Washington   53   7498    318
## 1789 2020-04-04      Illinois   17  10357    248
## 1797 2020-04-04 Massachusetts   25  11736    216
## 1785 2020-04-04       Georgia   13   6383    208
## 1784 2020-04-04       Florida   12  11537    194
```

For these 10 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.
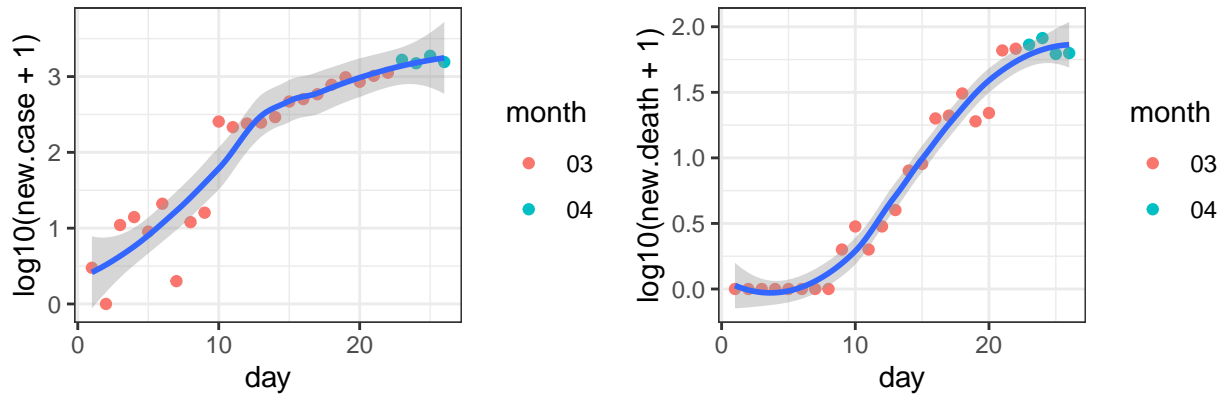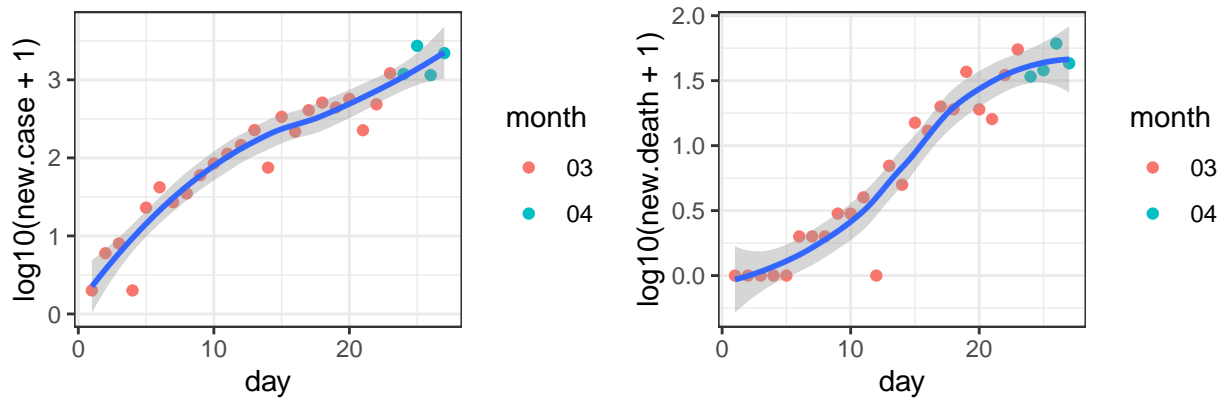

New York

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01

New Jersey

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04

Michigan

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

Louisiana

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09

California
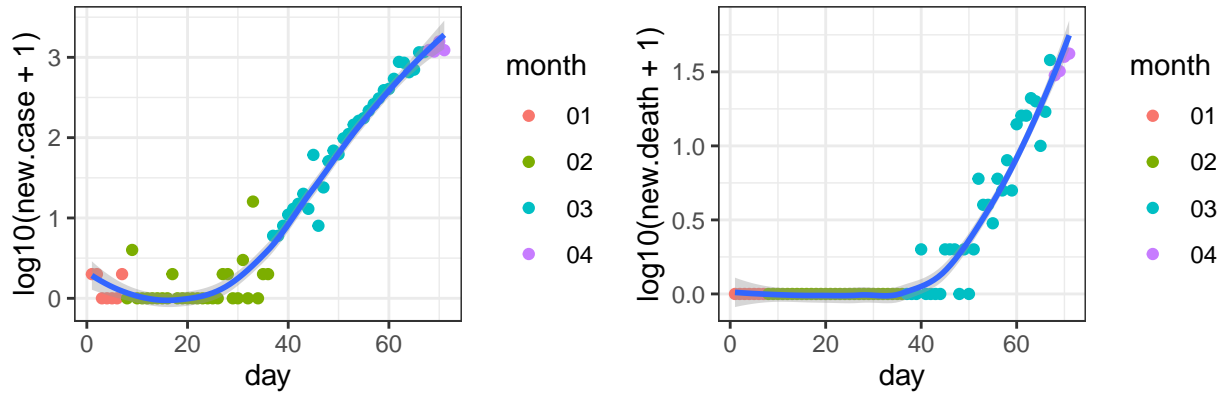
data source: https://github.com/nytimes/covid-19-data, day 1 is 01–25

Washington
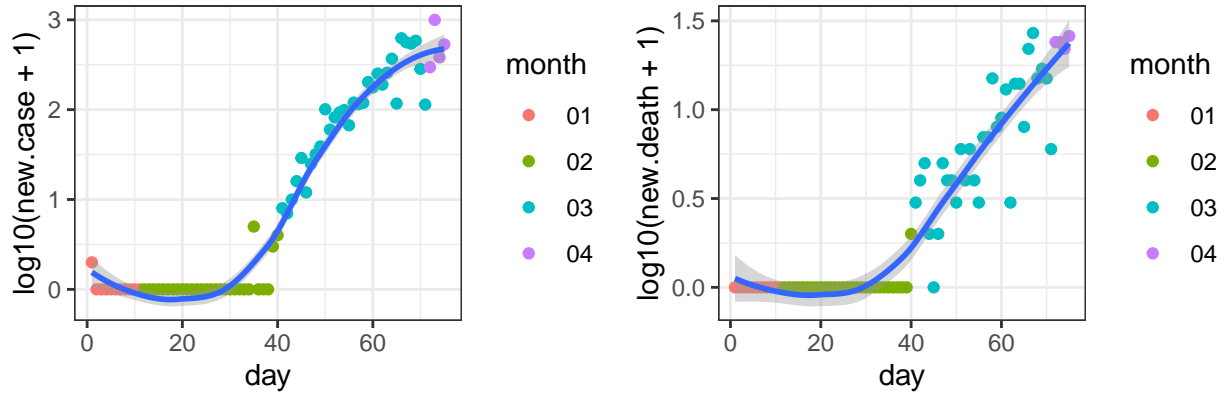
data source: https://github.com/nytimes/covid-19-data, day 1 is 01–21

Illinois
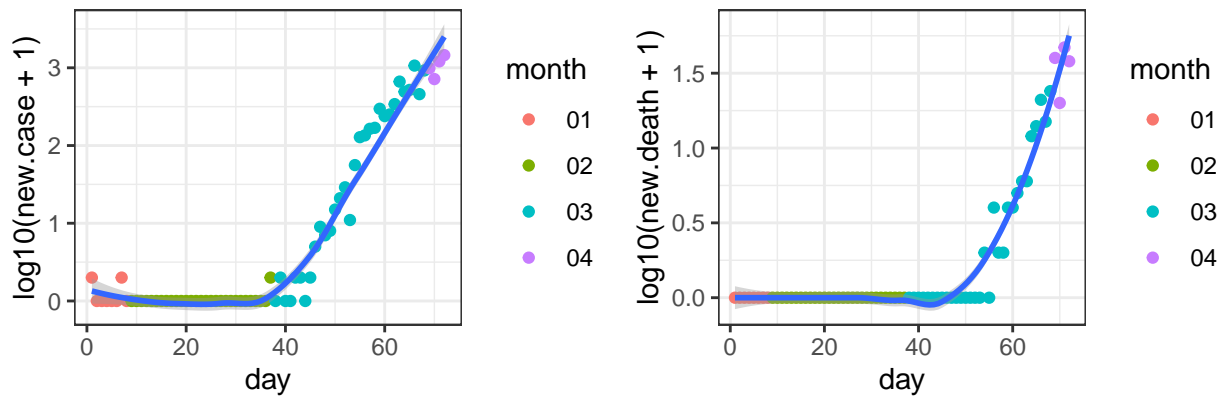
data source: https://github.com/nytimes/covid-19-data, day 1 is 01–24

Massachusetts
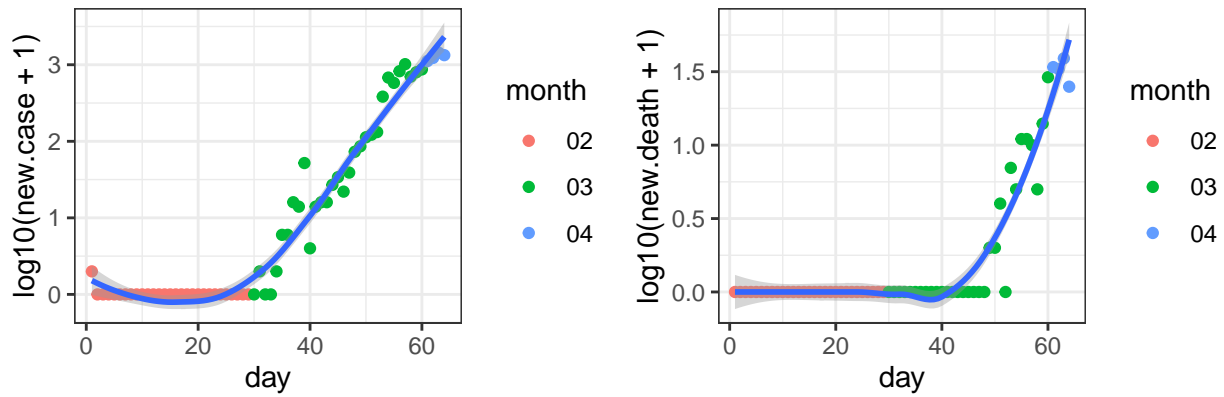
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01

Georgia
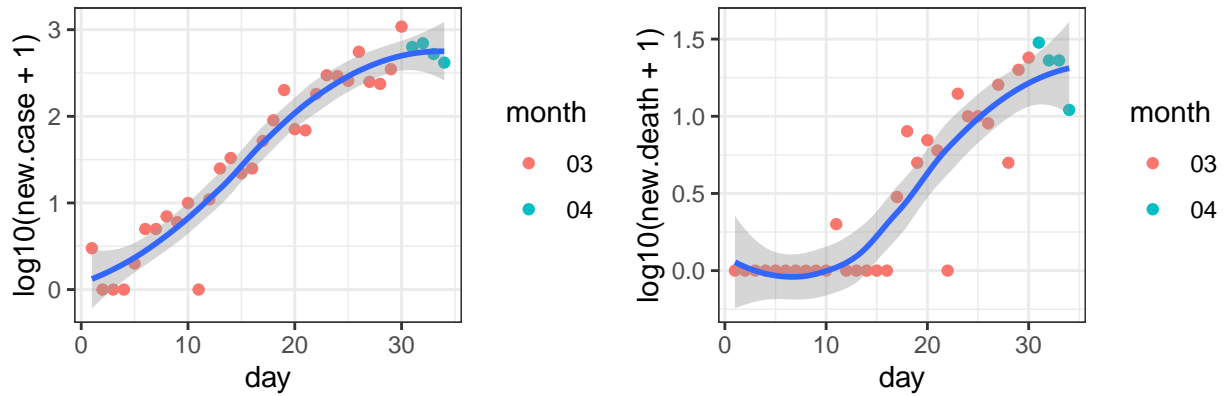
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02

Florida
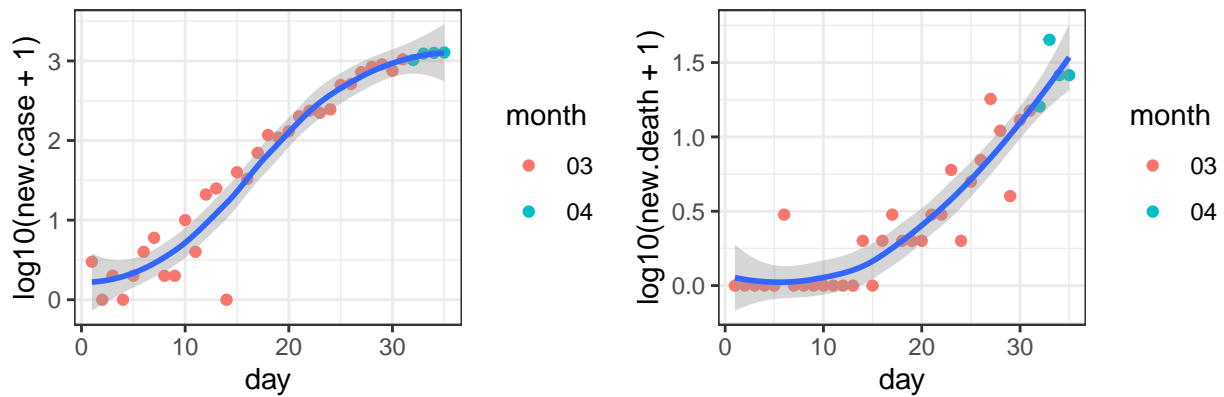
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March
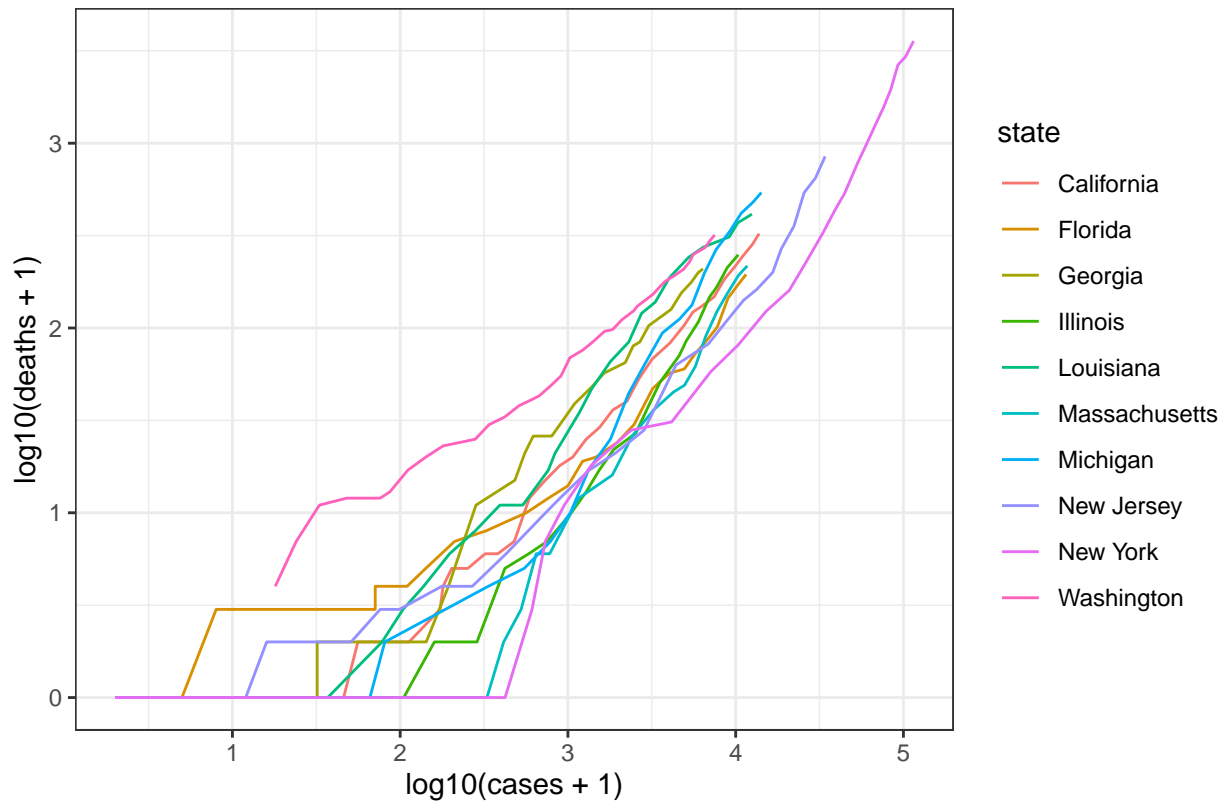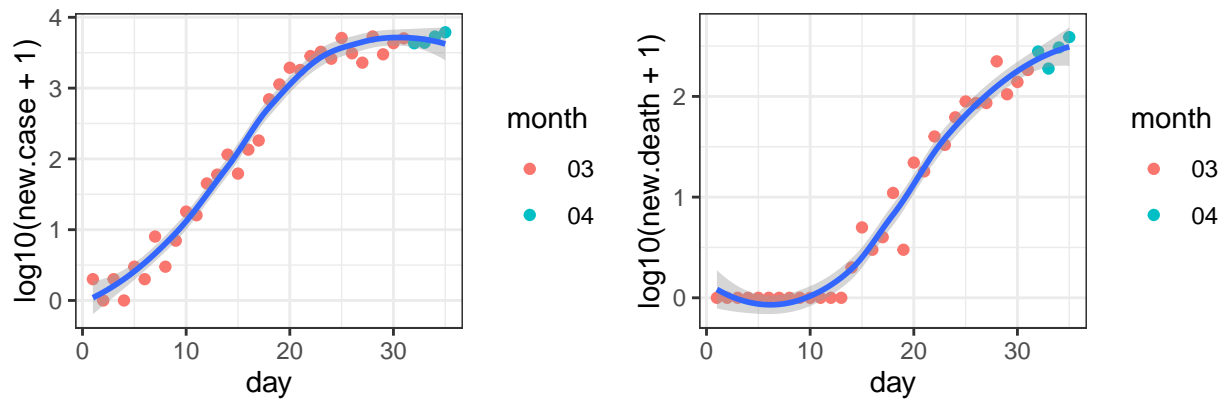
county level data

First check the 10 counties with the largest number of deaths.

```
##               date        county       state  fips cases deaths
## 32269 2020-04-04 New York City    New York    NA 63307   2254
## 32268 2020-04-04        Nassau    New York 36059 13346    396
## 32295 2020-04-04       Unknown    New York    NA     0    309
## 31904 2020-04-04         Wayne    Michigan 26163  6762    252
## 33125 2020-04-04          King  Washington 53033  2900    204
## 32299 2020-04-04   Westchester    New York 36119 13080    197
## 32194 2020-04-04        Bergen  New Jersey 34003  5760    179
## 32290 2020-04-04       Suffolk    New York 36103 12328    175
## 31376 2020-04-04          Cook    Illinois 17031  7439    167
## 32199 2020-04-04         Essex  New Jersey 34013  3584    156
```
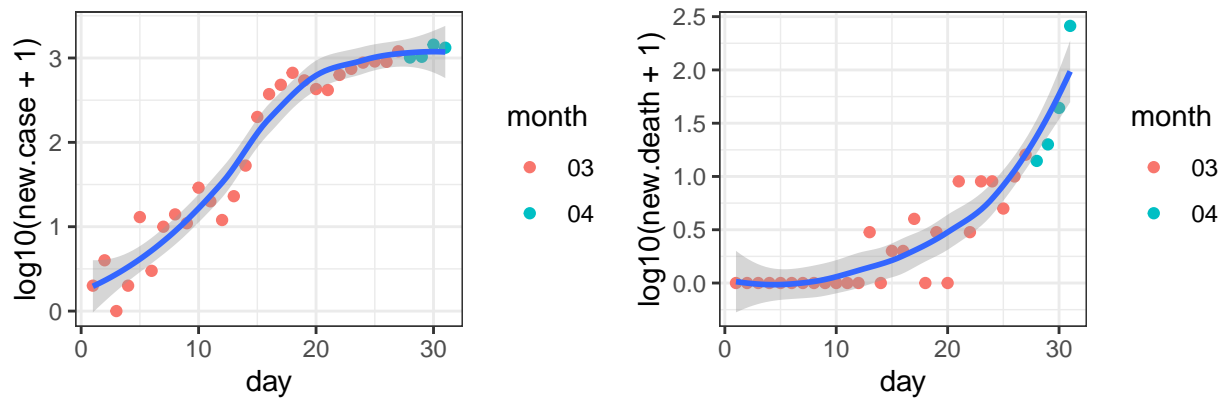
For these 10 counties, I check the number of new cases and the number of new deaths.
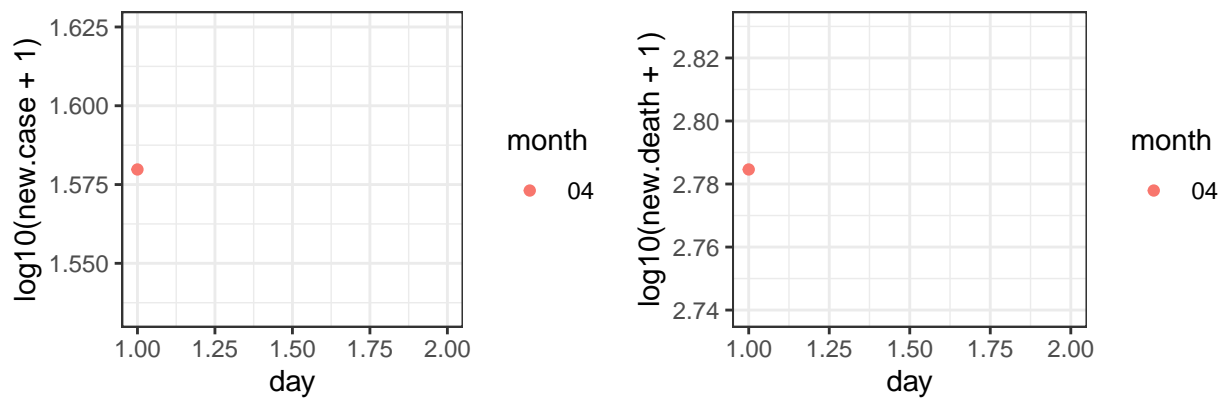
New York City_New York

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01
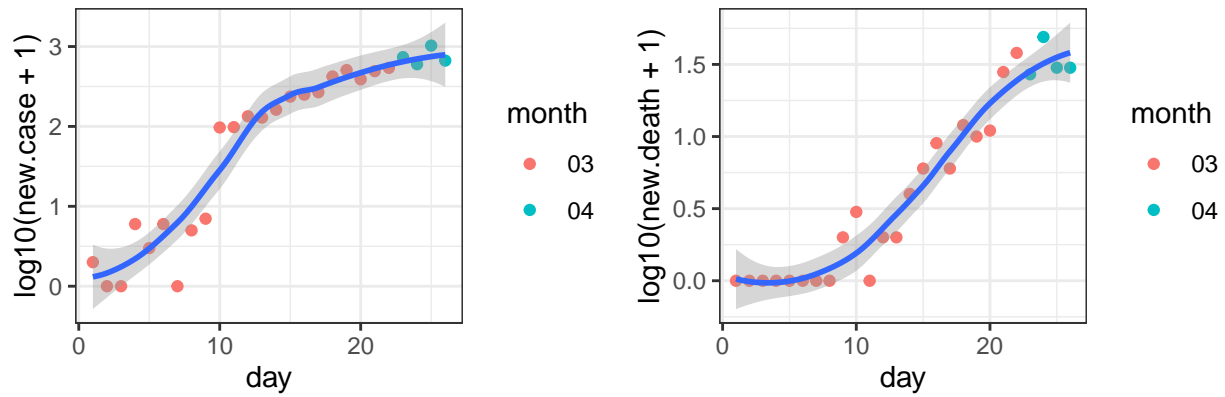
Nassau_New York

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

Unknown_New York

data source: https://github.com/nytimes/covid-19-data, day 1 is 04-03

Wayne_Michigan

data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

King_Washington
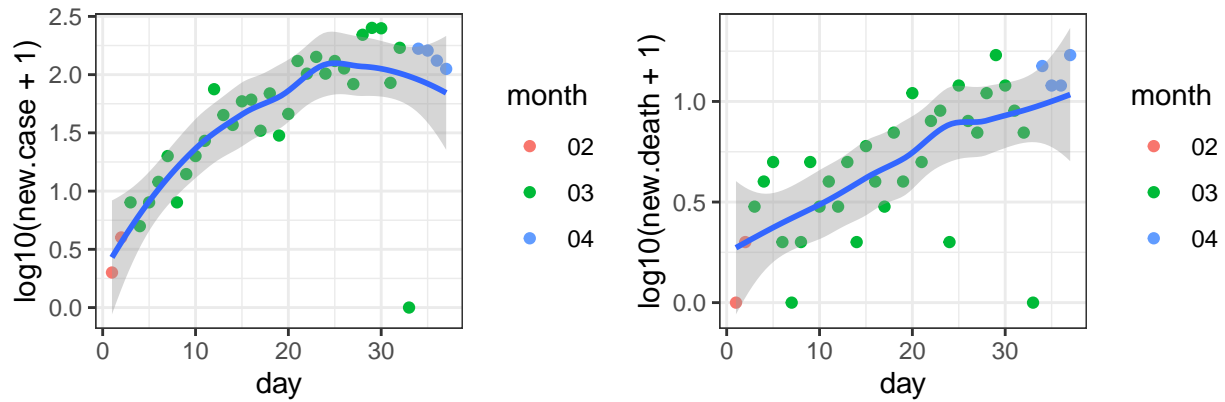
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-28

Westchester_New York
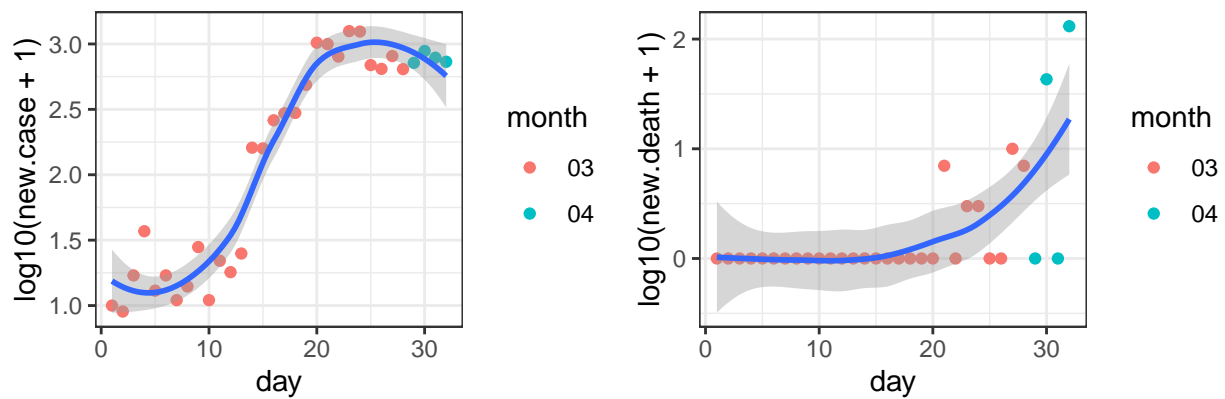
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04

Bergen_New Jersey
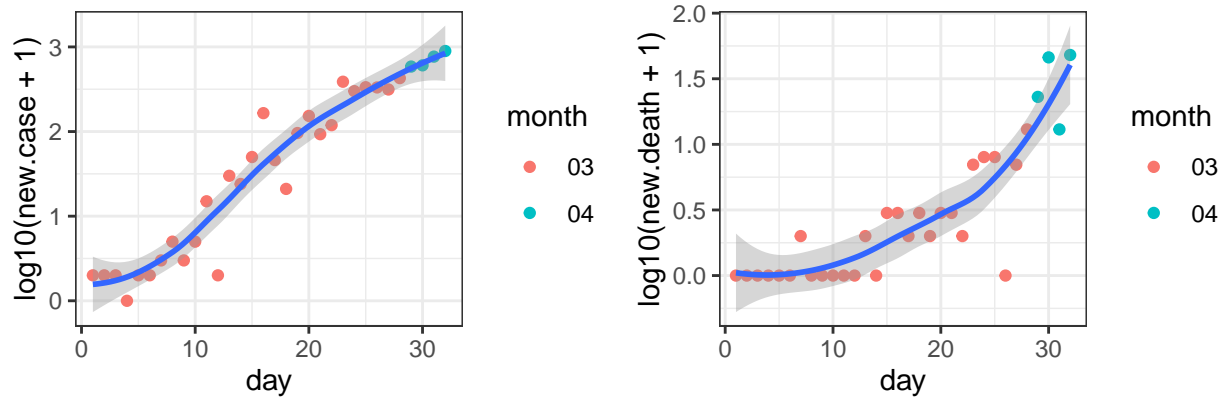
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Suffolk_New York
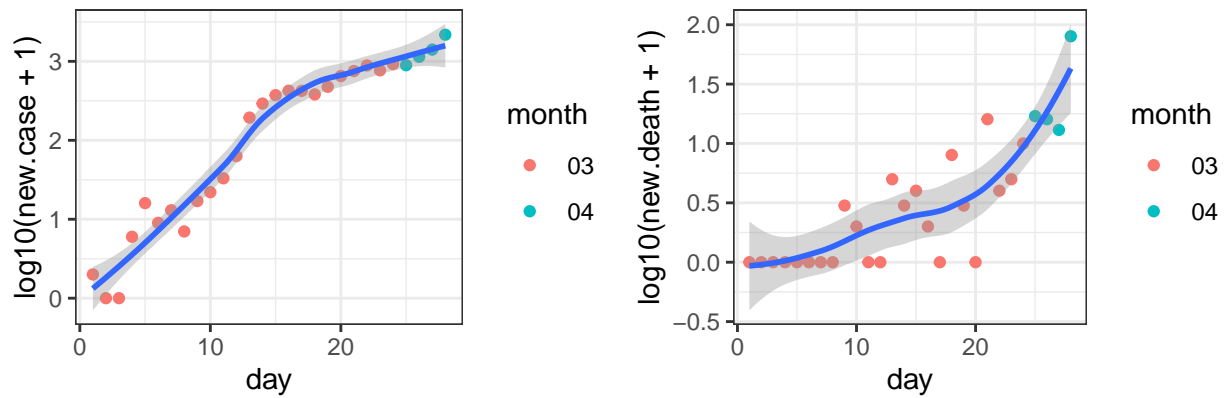
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Cook_Illinois
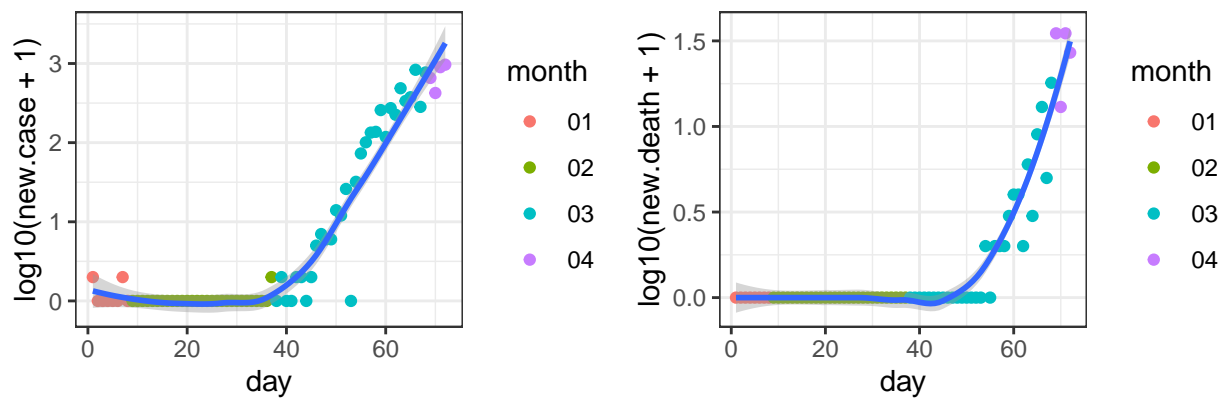
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Essex_New Jersey
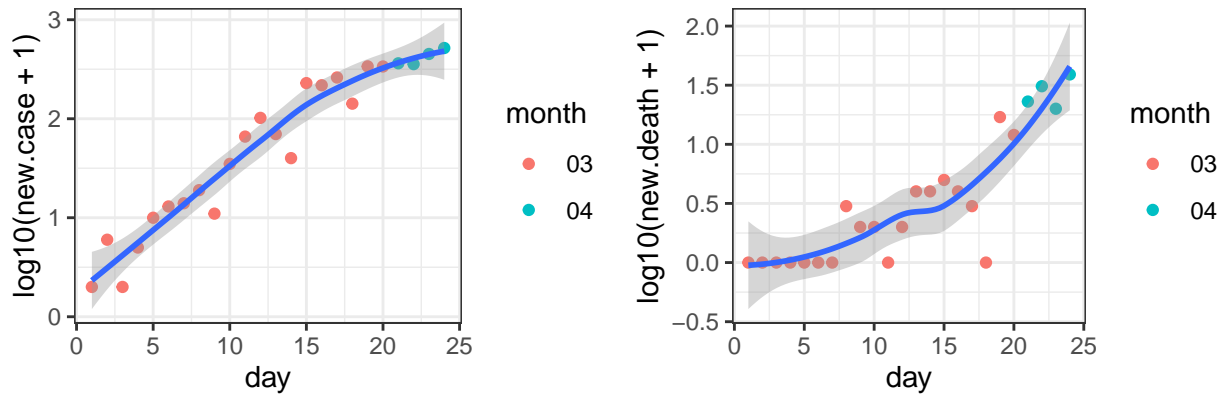
COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they are more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project proides a grade per state: "If you are calculating positive rates, it should only be with states that have an A grade. And be careful going back in time because almost all the states have changed their level of reporting at different times." (https://covidtracking.com/about-tracker/). The data are also availalbe for both counties and states, here I only look at state level data.

Since the daily postive rate can fluctuate a lot, here I only illustrae the cumulative positave rate across time, for four states with grade A data. Of course since this is an R markdown file, you can modify the source code and check for other states.
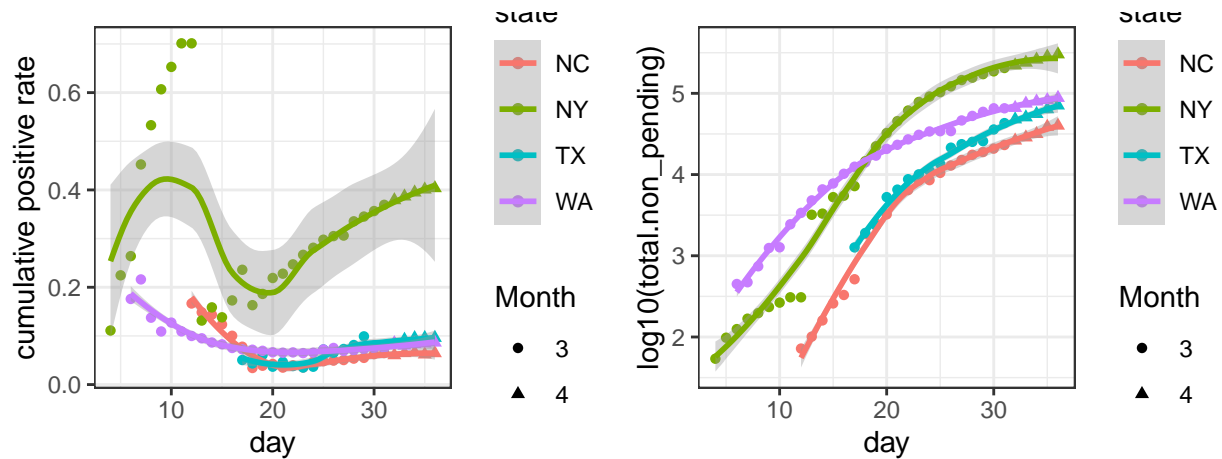
Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
```

15

```
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] httr_1.4.1    ggpubr_0.2.5  magrittr_1.5  ggplot2_3.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3        pillar_1.4.3      compiler_3.6.2    tools_3.6.2
##  [5] digest_0.6.23     evaluate_0.14     lifecycle_0.1.0   tibble_2.1.3
##  [9] gtable_0.3.0      pkgconfig_2.0.3   rlang_0.4.4       yaml_2.2.1
## [13] xfun_0.12         gridExtra_2.3     withr_2.1.2       dplyr_0.8.4
## [17] stringr_1.4.0     knitr_1.28        grid_3.6.2        tidyselect_1.0.0
## [21] cowplot_1.0.0     glue_1.3.1        R6_2.4.1          rmarkdown_2.1
## [25] purrr_0.3.3       farver_2.0.3      scales_1.1.0      htmltools_0.4.0
## [29] assertthat_0.2.1  colorspace_1.4-1  ggsignif_0.6.0    labeling_0.3
## [33] stringi_1.4.5     lazyeval_0.2.2    munsell_0.5.0     crayon_1.3.4
```