Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-04-29

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

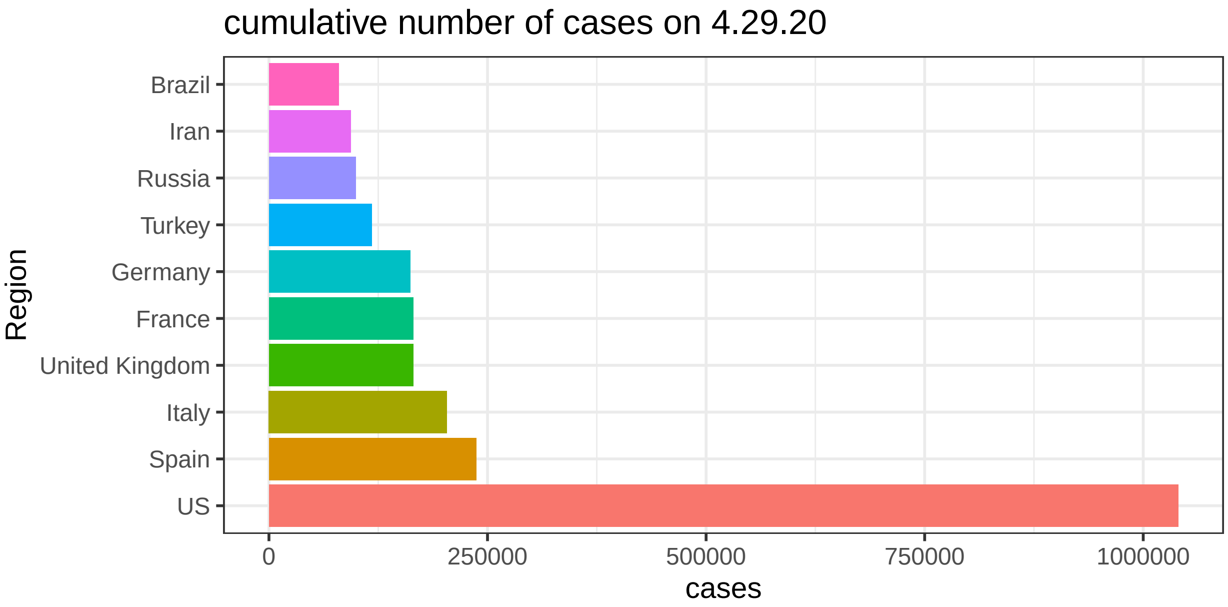
JHU

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

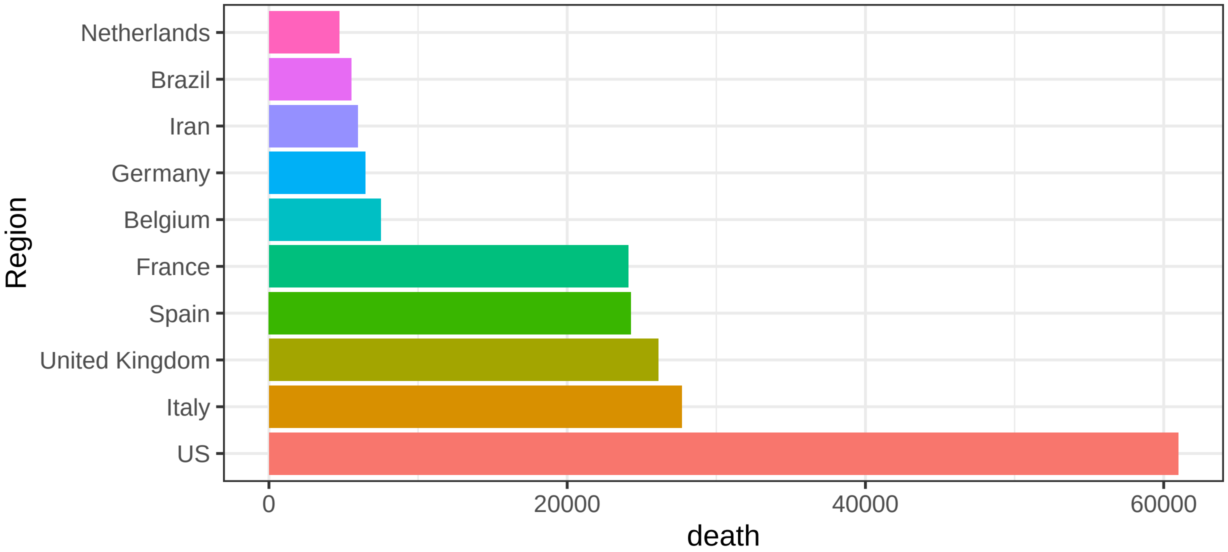
time series data

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.

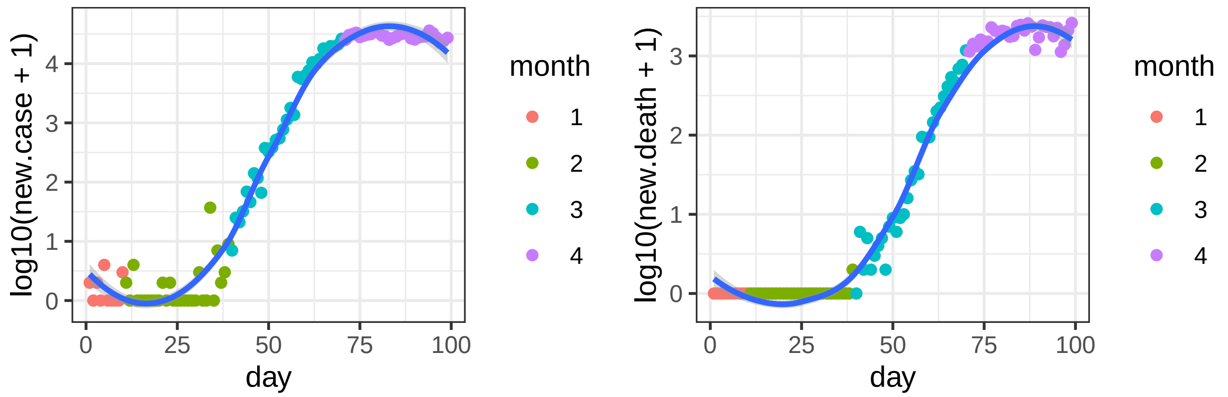


cumulative number of deaths on 4.29.20



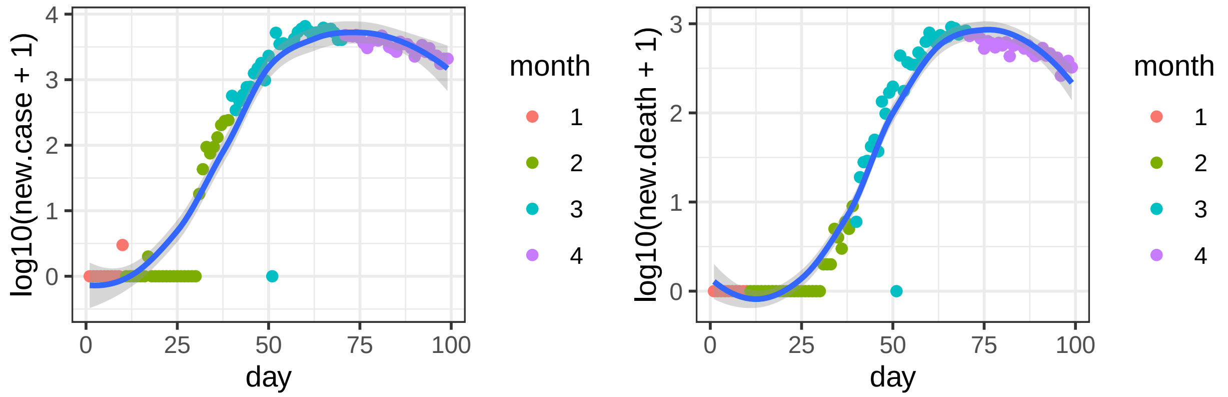
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



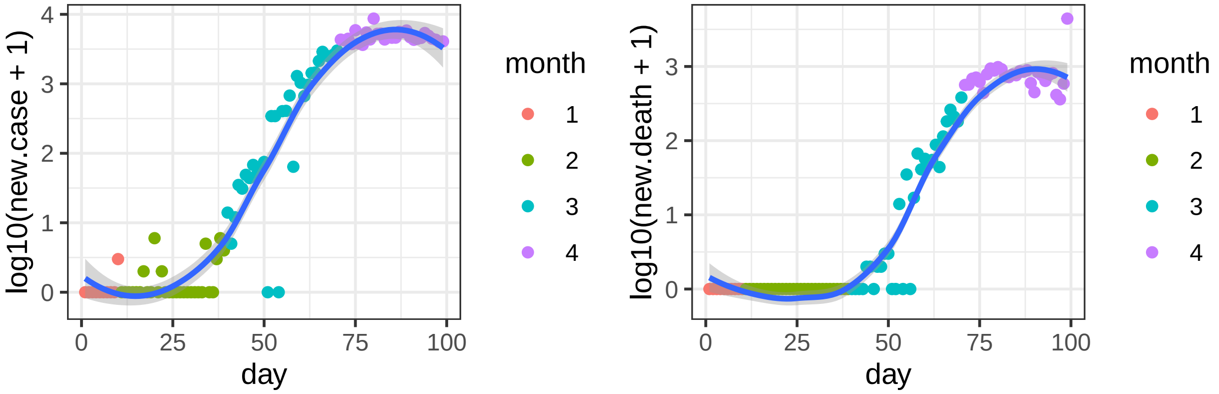
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



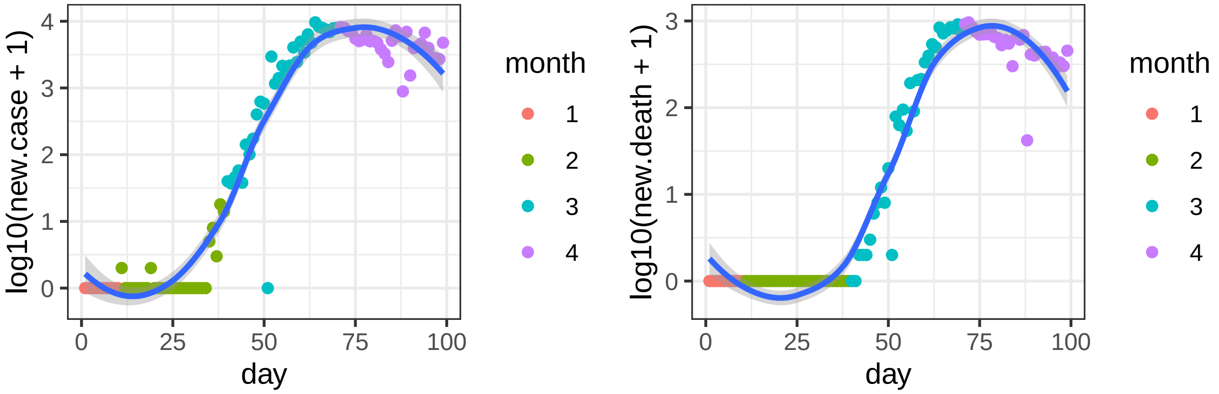
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



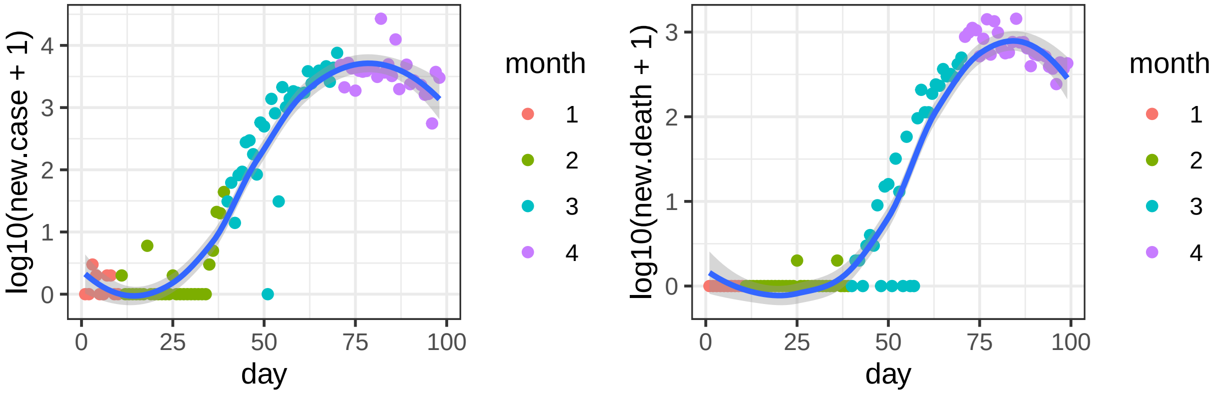
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



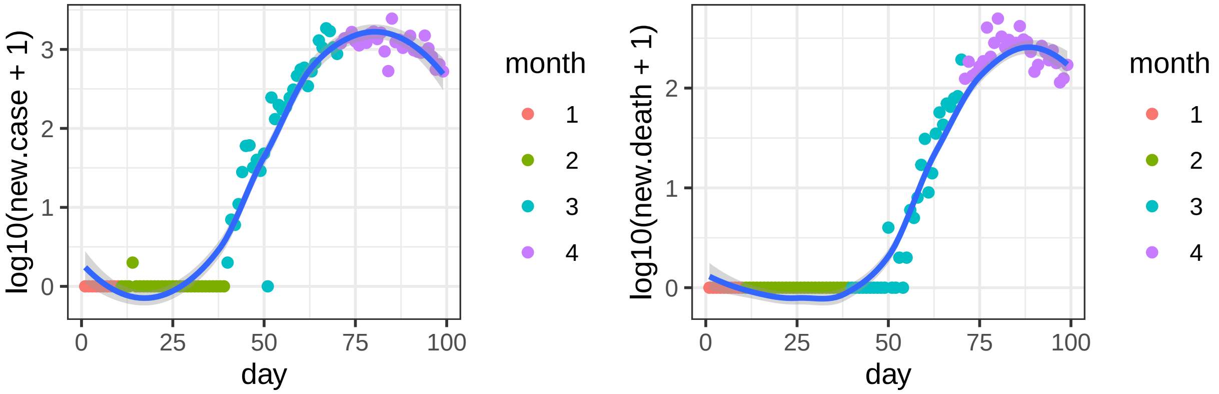
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



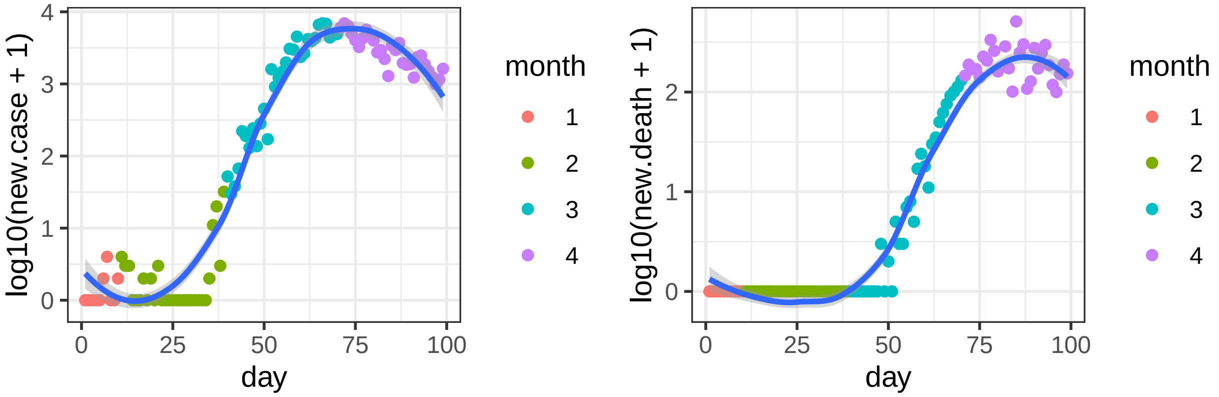
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium



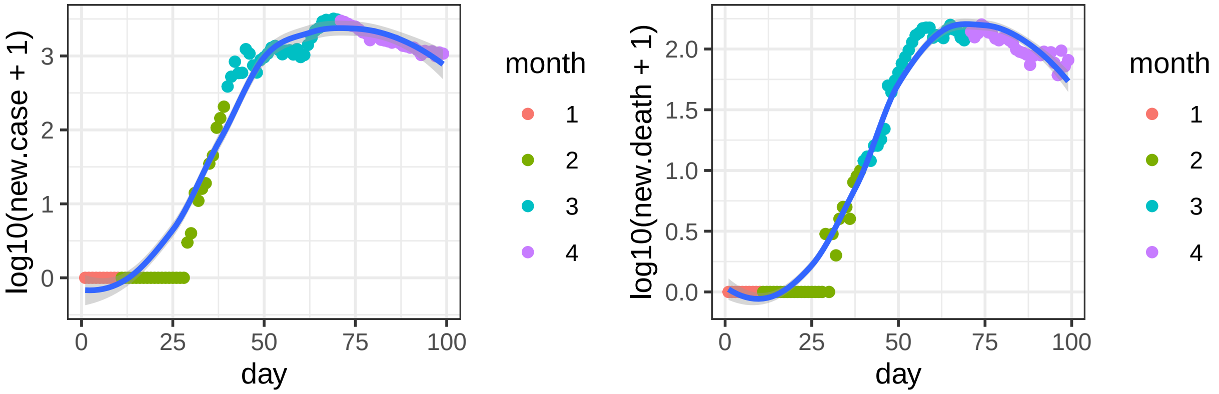
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Germany



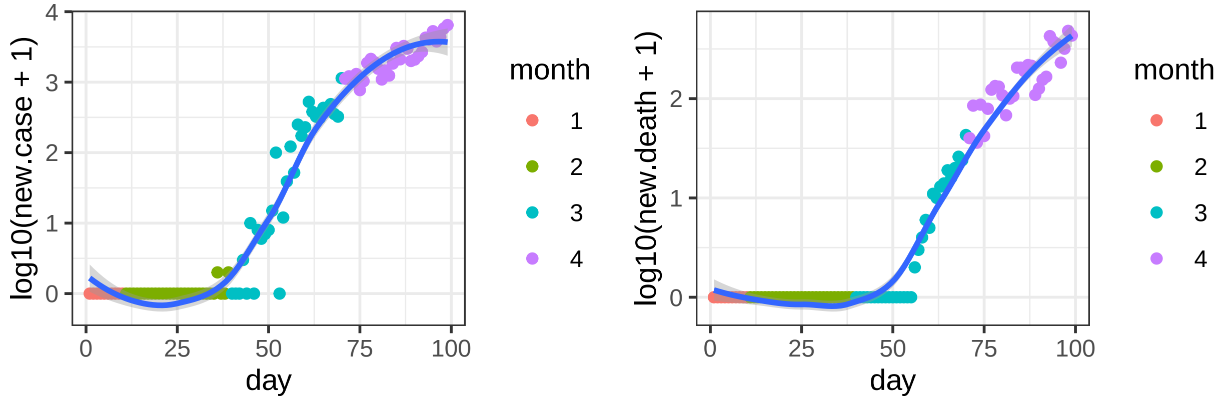
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Iran



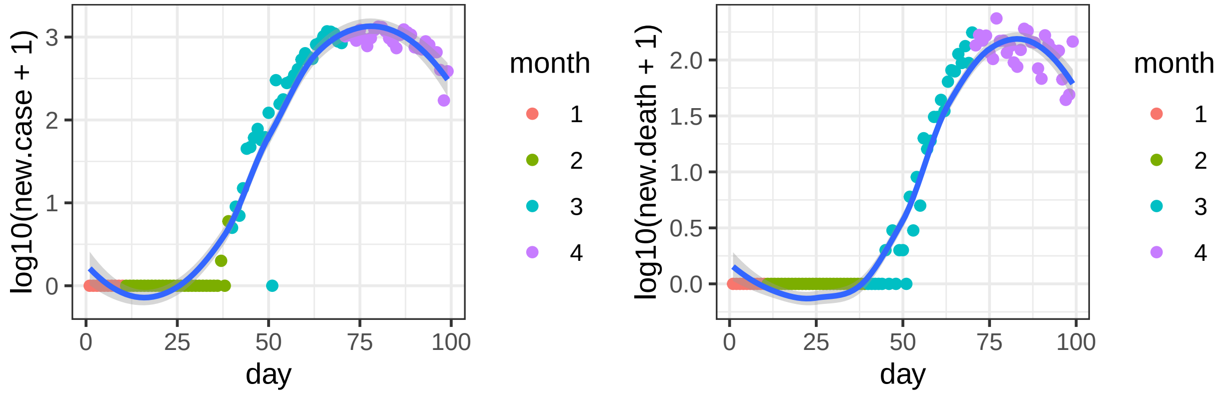
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



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

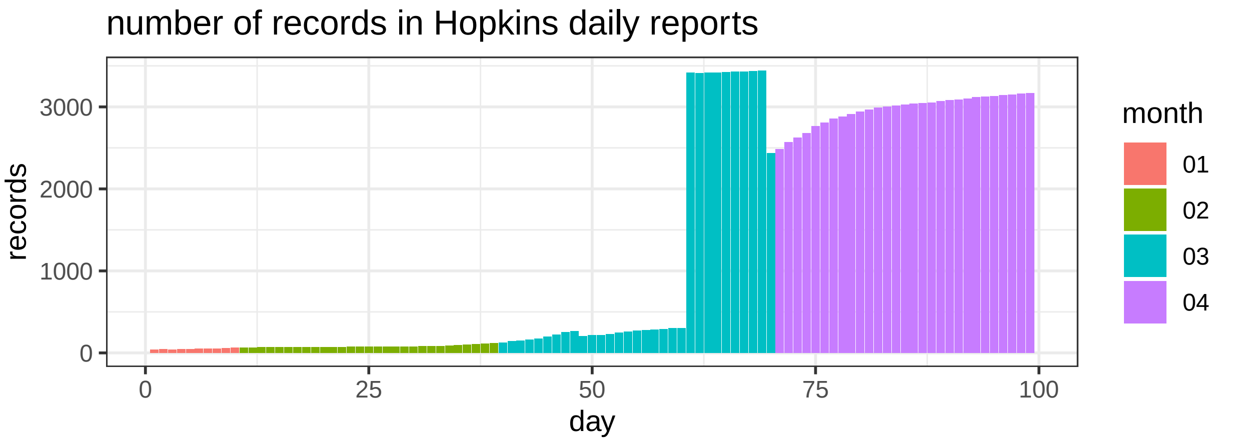
Netherlands



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

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.



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] "2020-04-28"
```

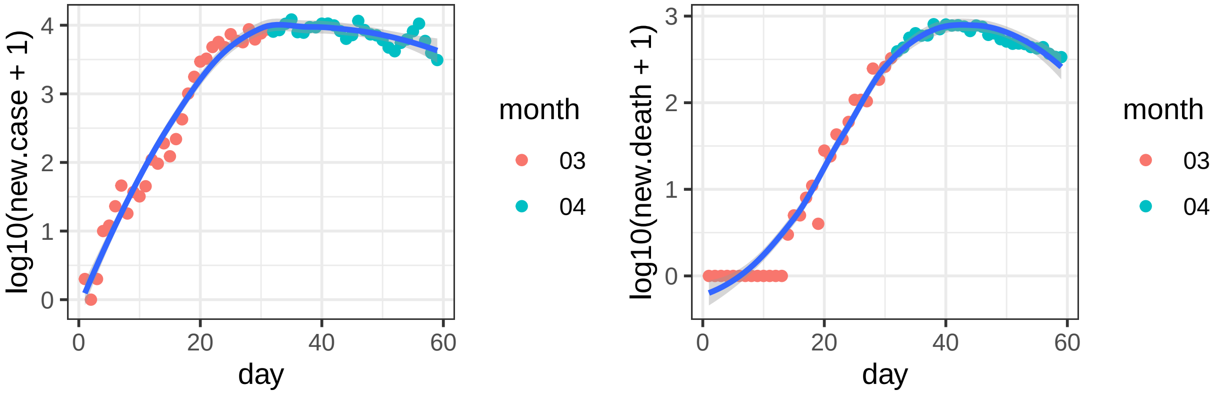
state level data

First check the 30 states with the largest number of deaths.

```
##            date          state fips  cases deaths
## 3128 2020-04-28       New York   36 295137  17638
## 3126 2020-04-28     New Jersey   34 113856   6442
## 3118 2020-04-28       Michigan   26  39234   3566
## 3117 2020-04-28  Massachusetts   25  58302   3153
## 3109 2020-04-28       Illinois   17  48102   2132
## 3135 2020-04-28   Pennsylvania   42  45323   2092
## 3101 2020-04-28    Connecticut    9  26312   2089
## 3099 2020-04-28     California    6  46570   1884
## 3114 2020-04-28      Louisiana   22  27286   1758
## 3104 2020-04-28        Florida   12  32838   1170
## 3105 2020-04-28        Georgia   13  23607   1022
## 3116 2020-04-28       Maryland   24  20113    929
## 3110 2020-04-28        Indiana   18  16588    901
## 3132 2020-04-28           Ohio   39  16769    799
## 3146 2020-04-28     Washington   53  14059    792
## 3141 2020-04-28          Texas   48  26865    738
## 3100 2020-04-28       Colorado    8  14239    734
## 3145 2020-04-28       Virginia   51  14339    492
## 3129 2020-04-28 North Carolina   37   9568    353
## 3121 2020-04-28       Missouri   29   7408    324
## 3119 2020-04-28      Minnesota   27   4181    301
## 3148 2020-04-28      Wisconsin   55   6289    300
## 3097 2020-04-28        Arizona    4   6948    297
## 3095 2020-04-28        Alabama    1   6750    242
## 3120 2020-04-28    Mississippi   28   6342    239
## 3137 2020-04-28   Rhode Island   44   7926    239
## 3113 2020-04-28       Kentucky   21   4375    231
## 3124 2020-04-28         Nevada   32   4812    225
## 3133 2020-04-28       Oklahoma   40   3410    207
## 3140 2020-04-28      Tennessee   47  10031    198
```

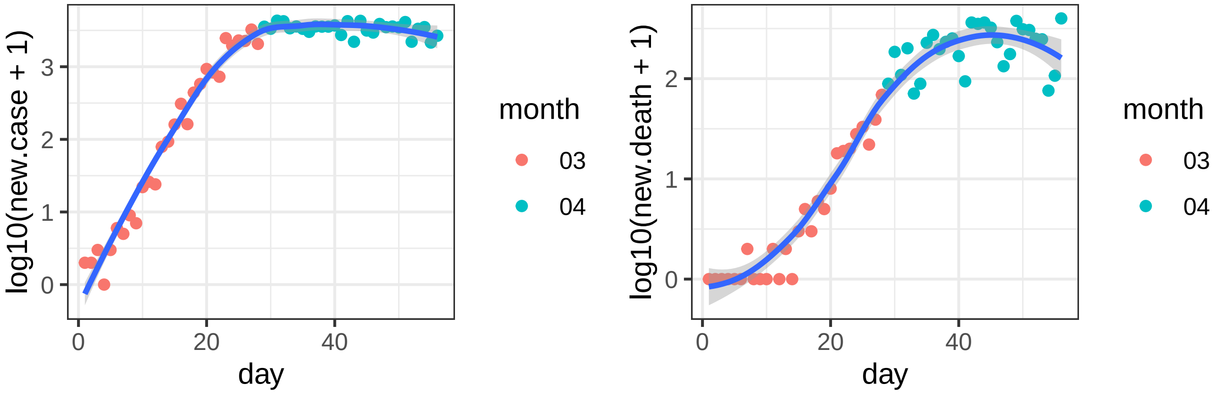
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



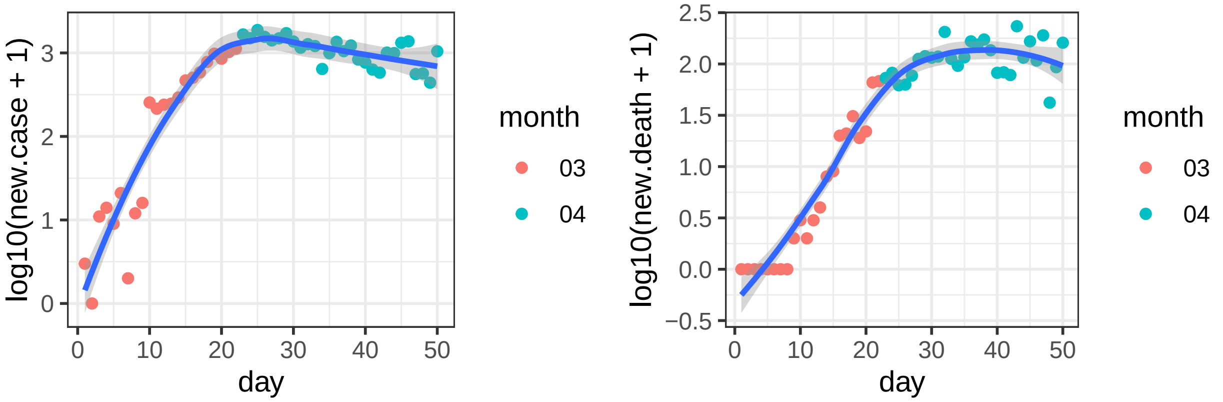
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



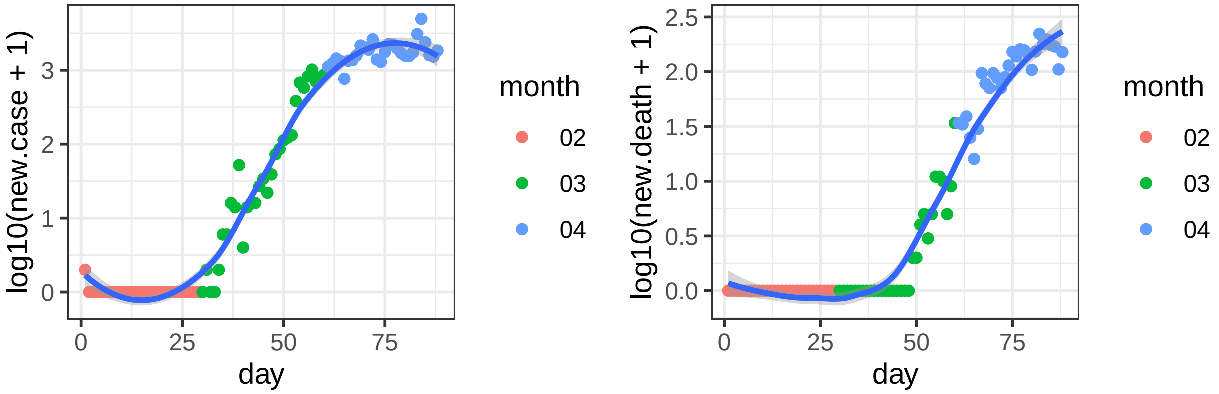
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Michigan



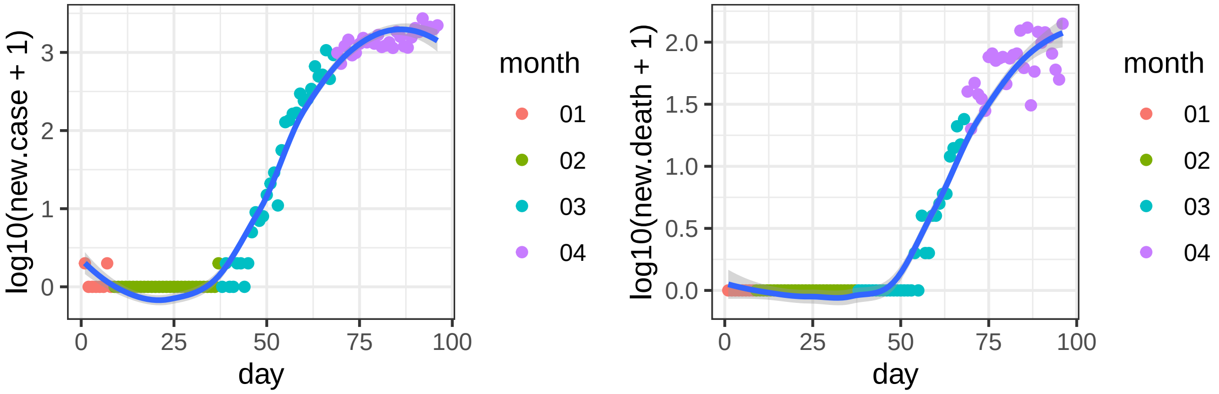
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Massachusetts



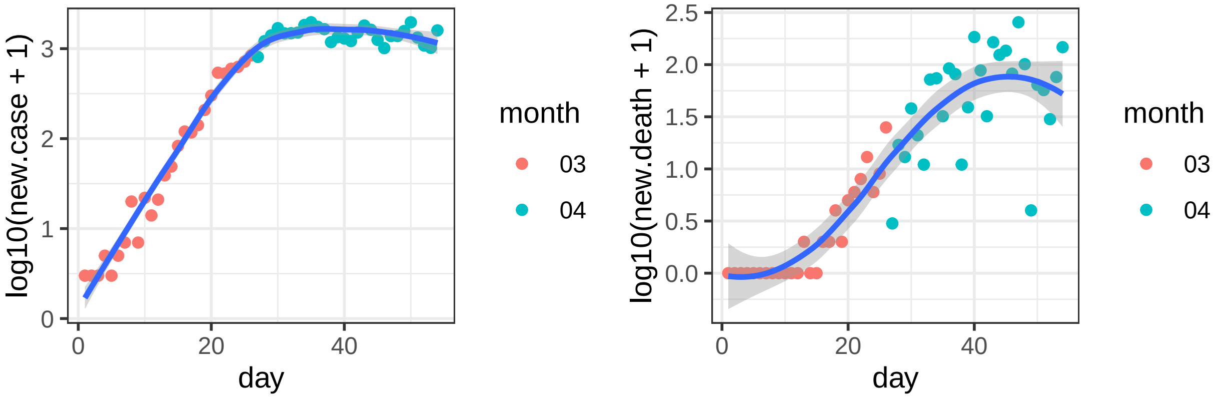
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois

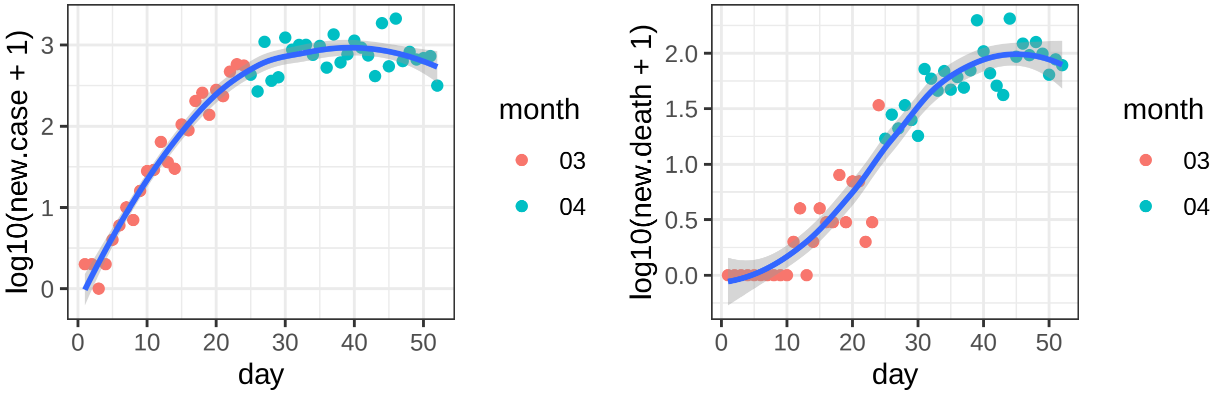


data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania

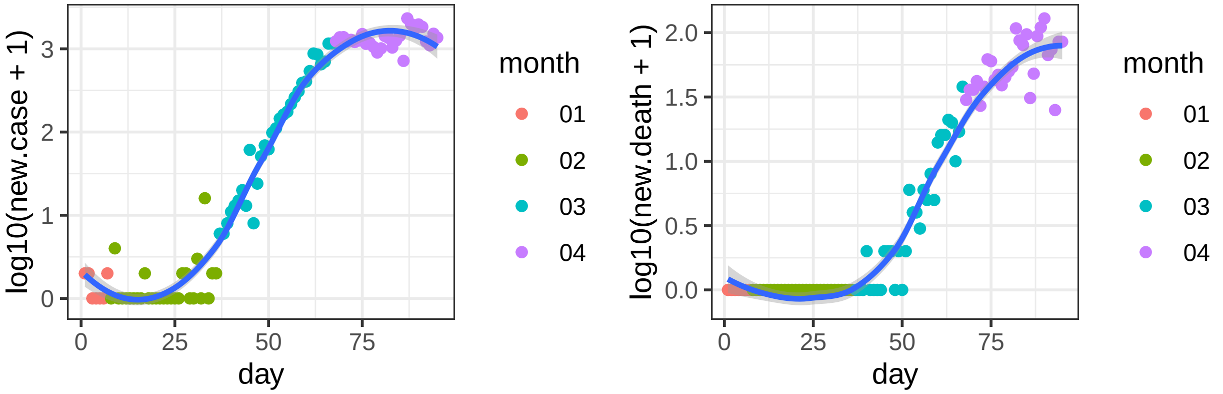


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



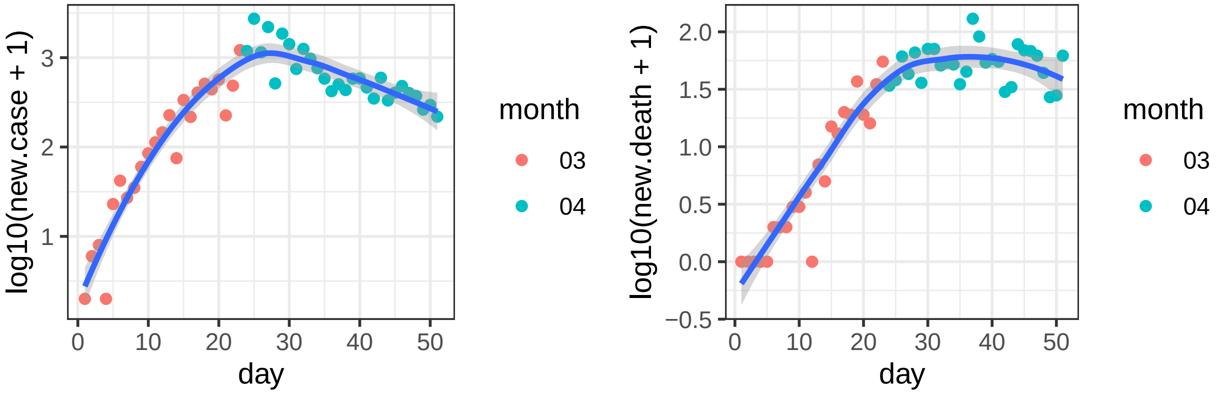
Connecticut

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



California

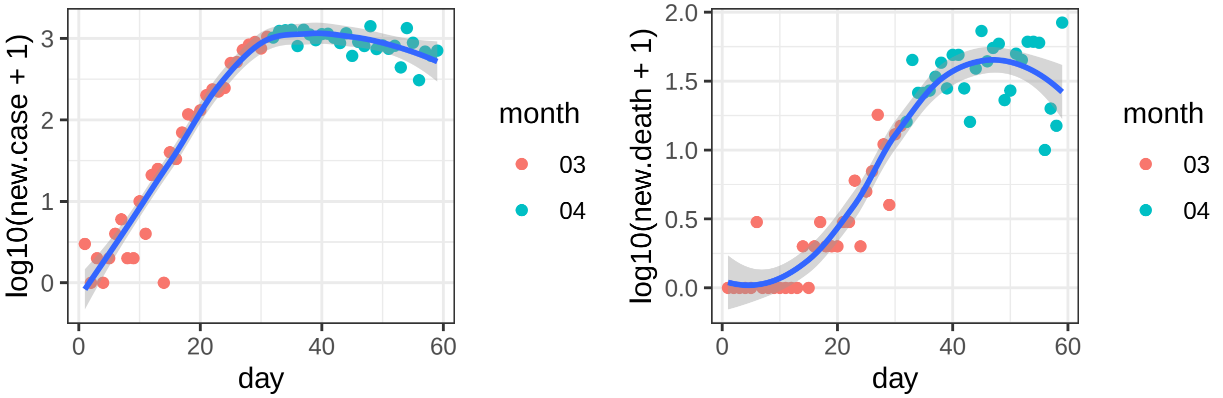
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25



Louisiana

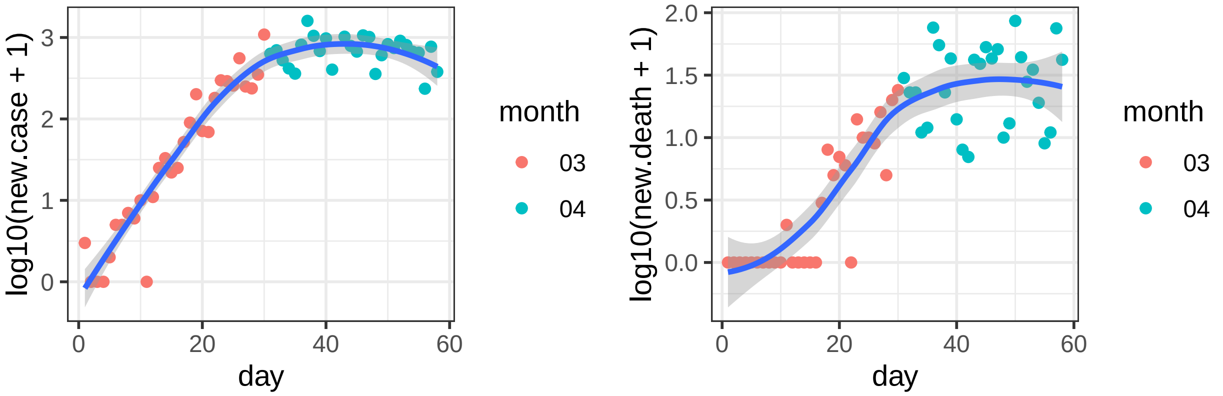
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Florida



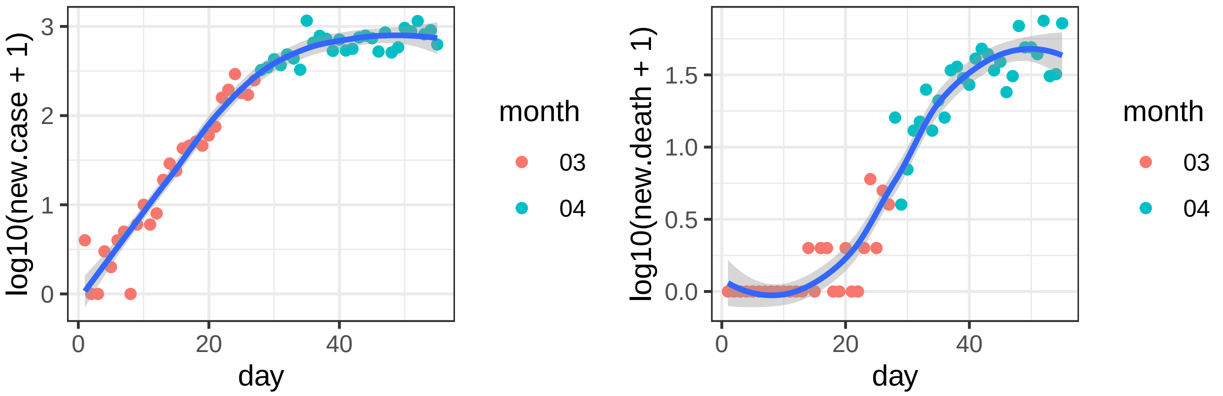
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Georgia



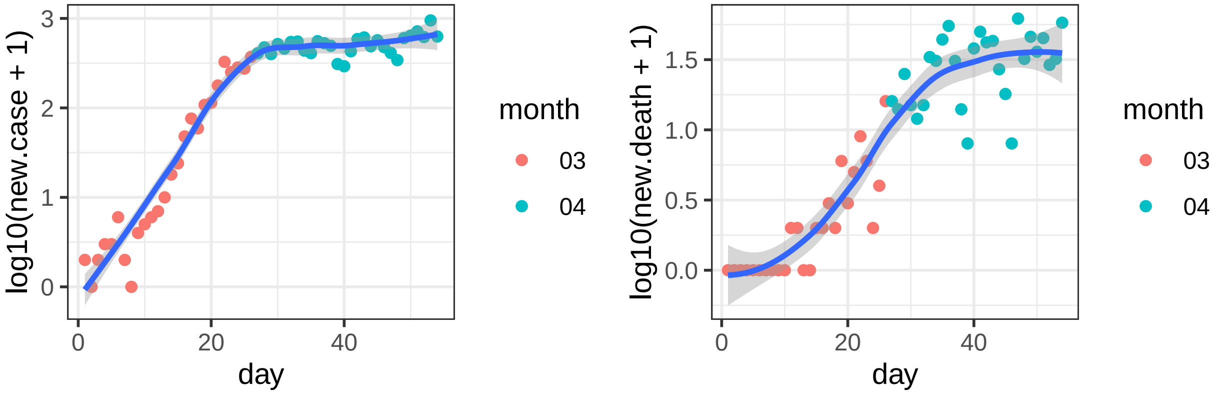
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Maryland



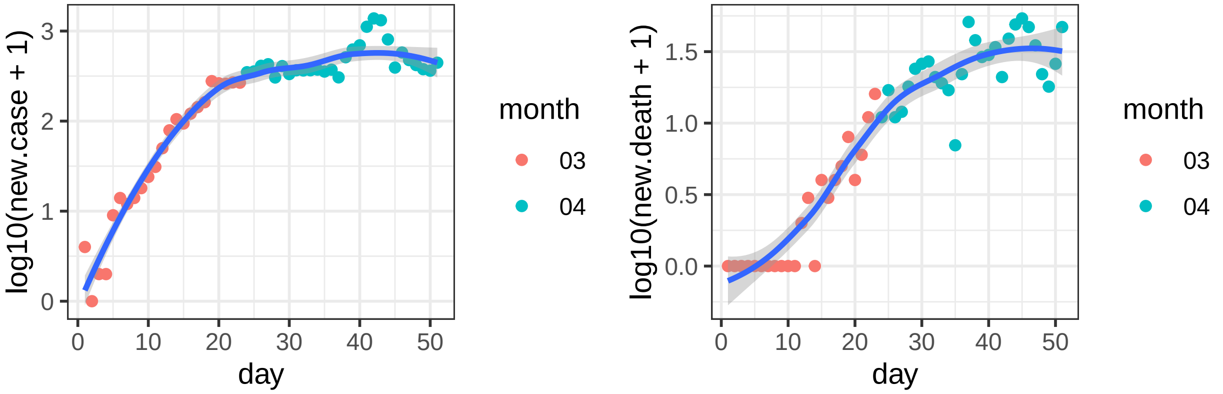
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



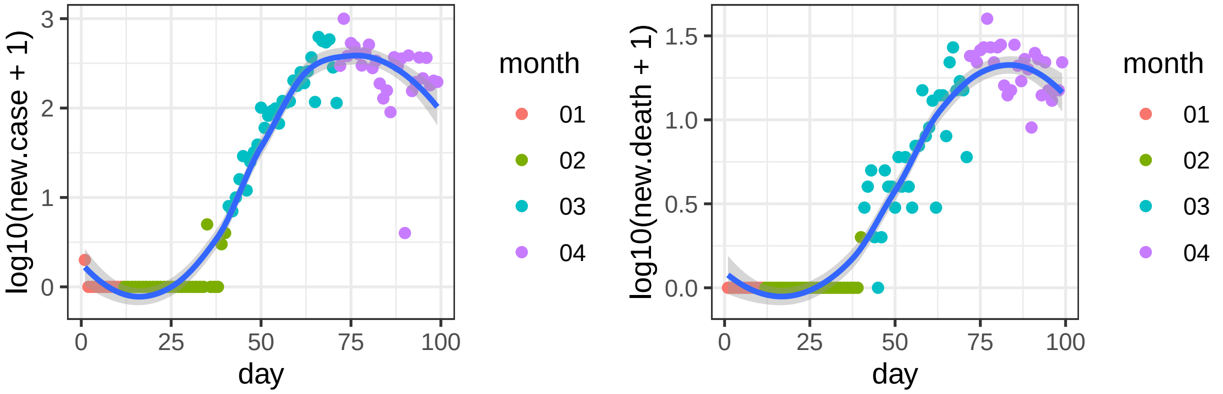
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Ohio



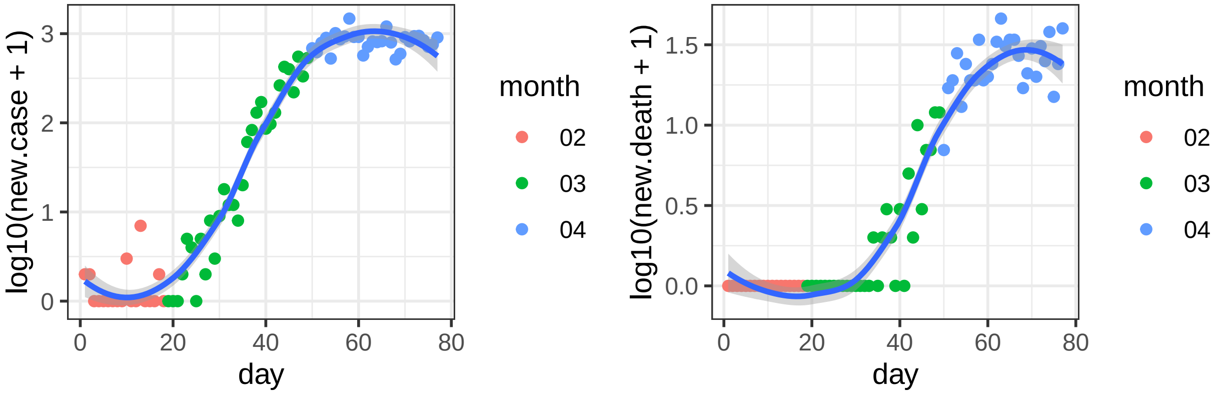
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Washington



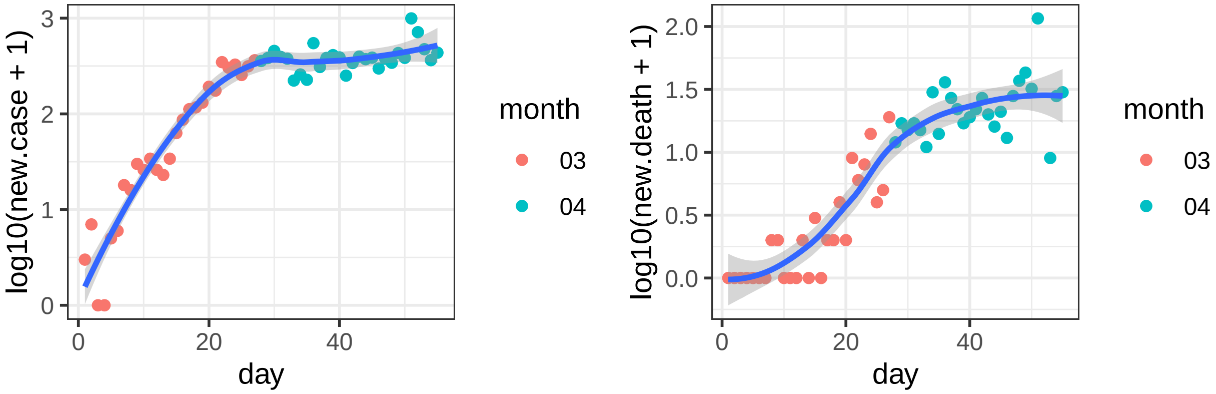
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Texas



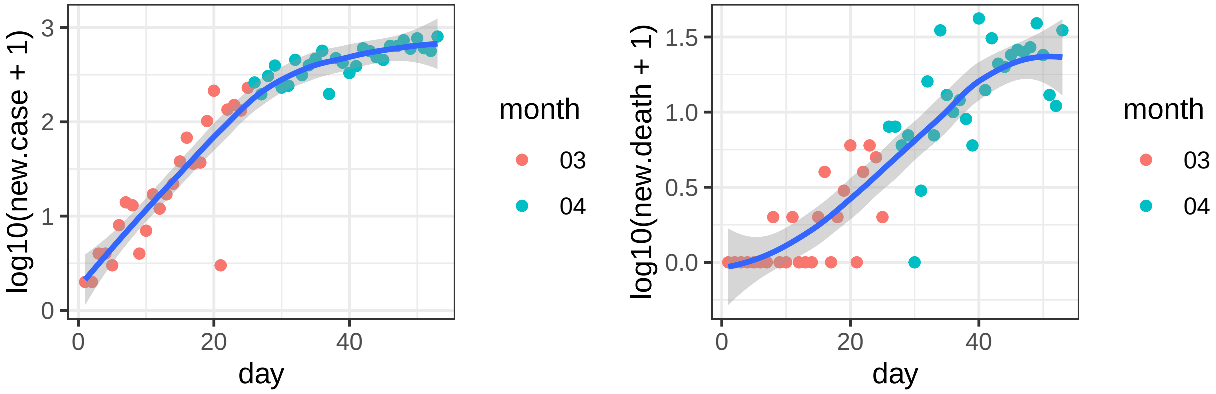
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



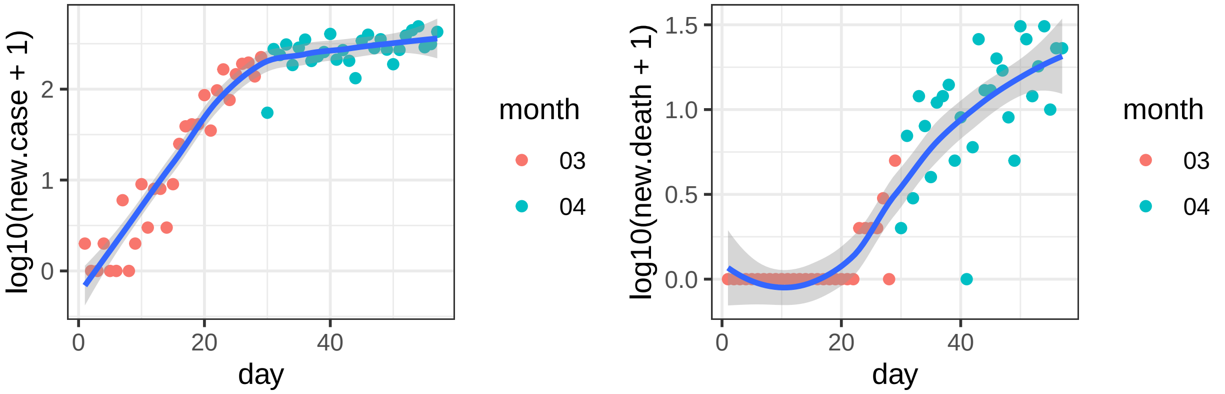
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



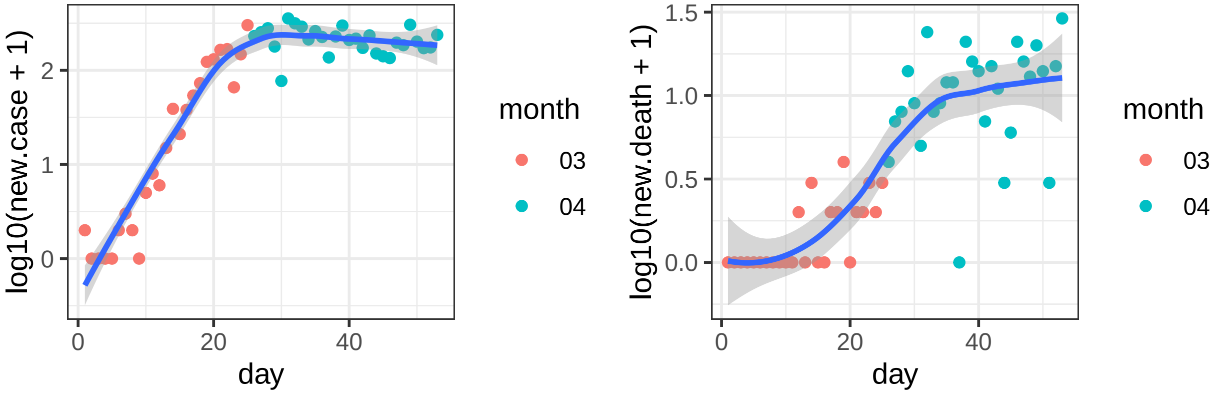
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



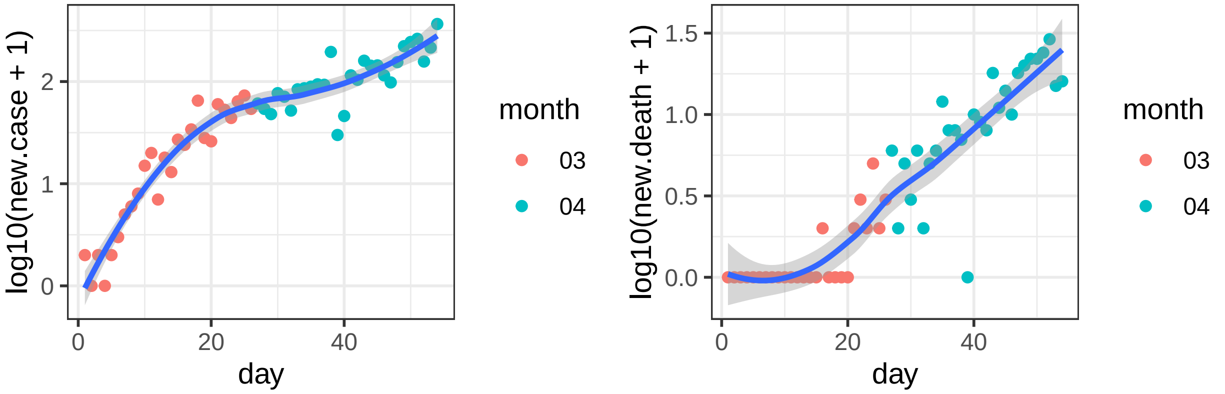
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Missouri



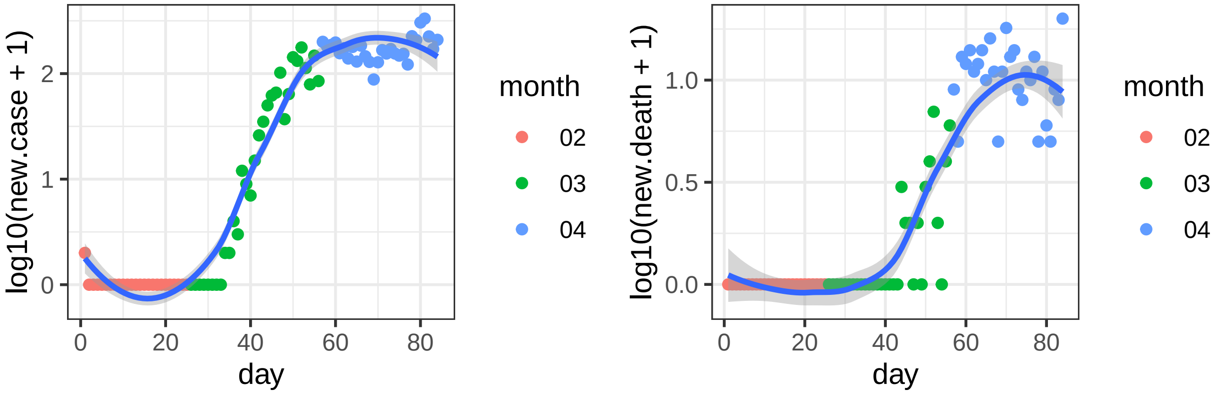
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



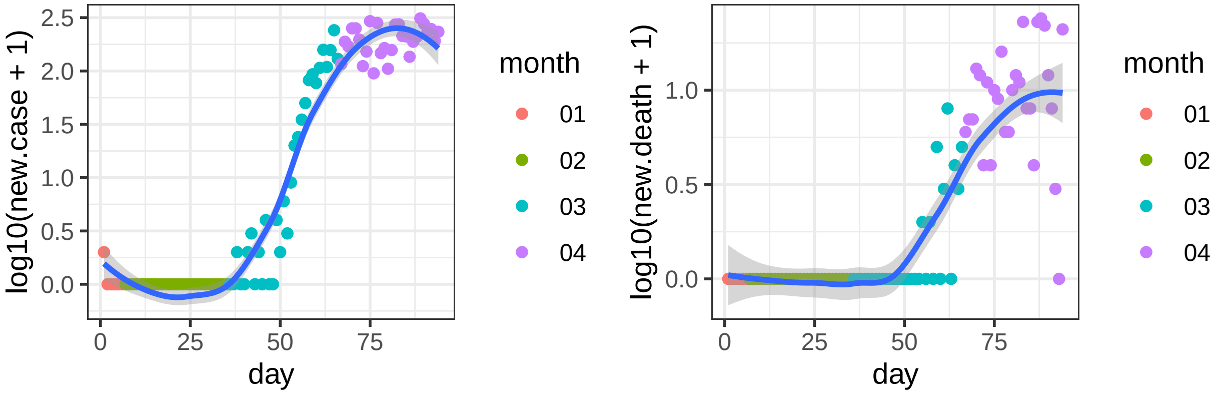
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Wisconsin



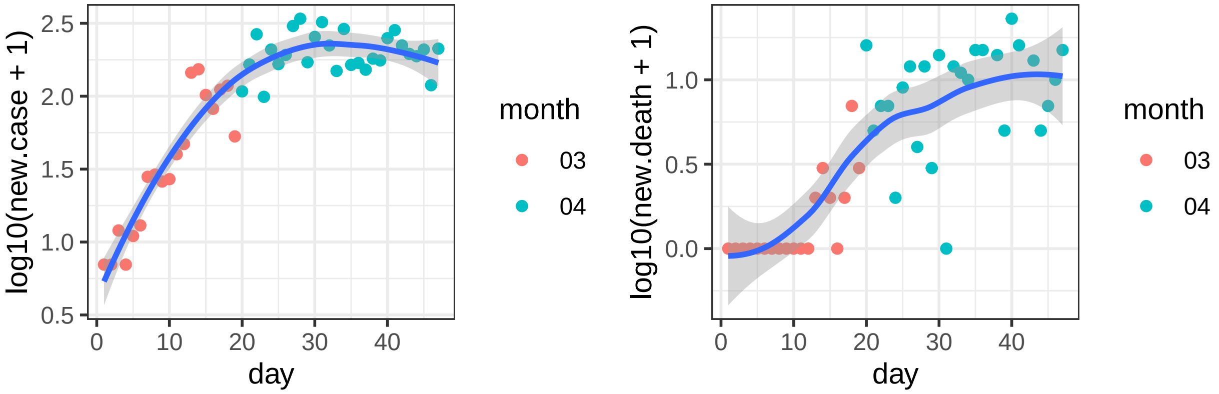
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Arizona



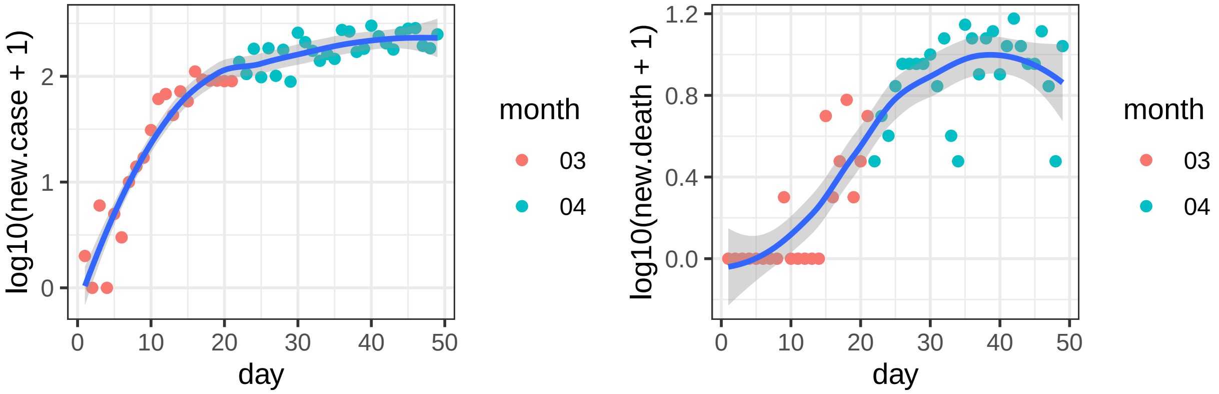
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Alabama



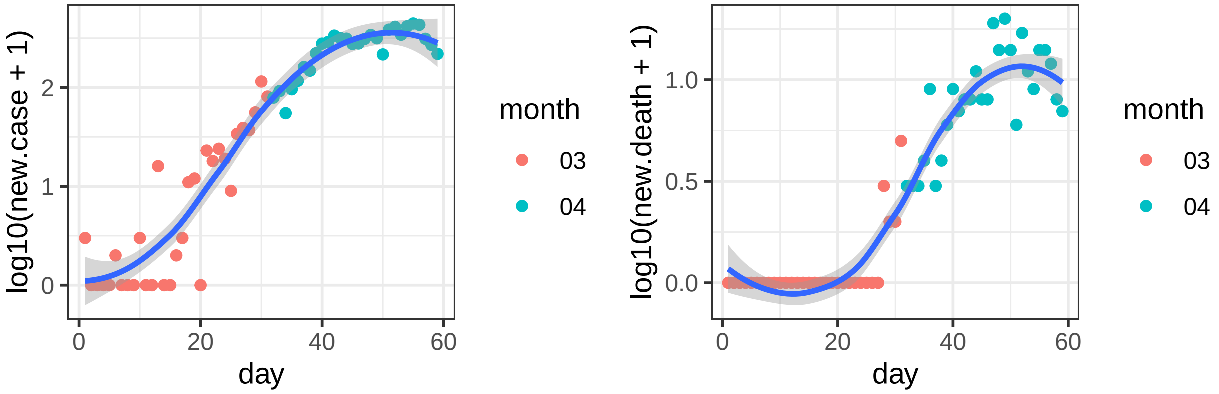
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Mississippi



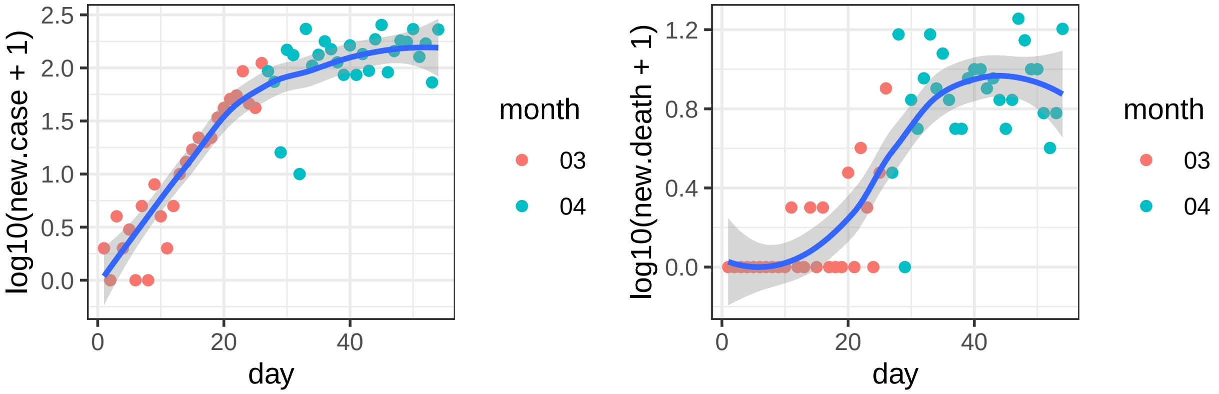
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



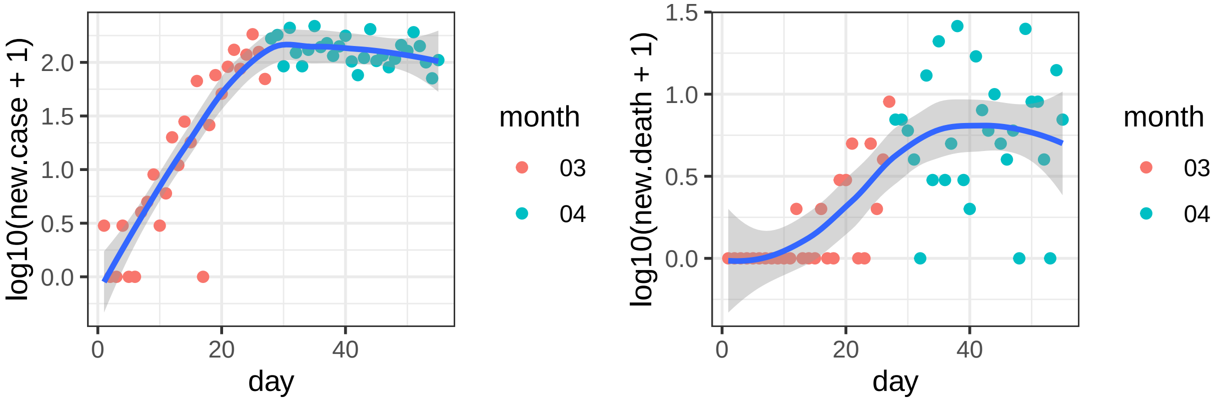
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Kentucky



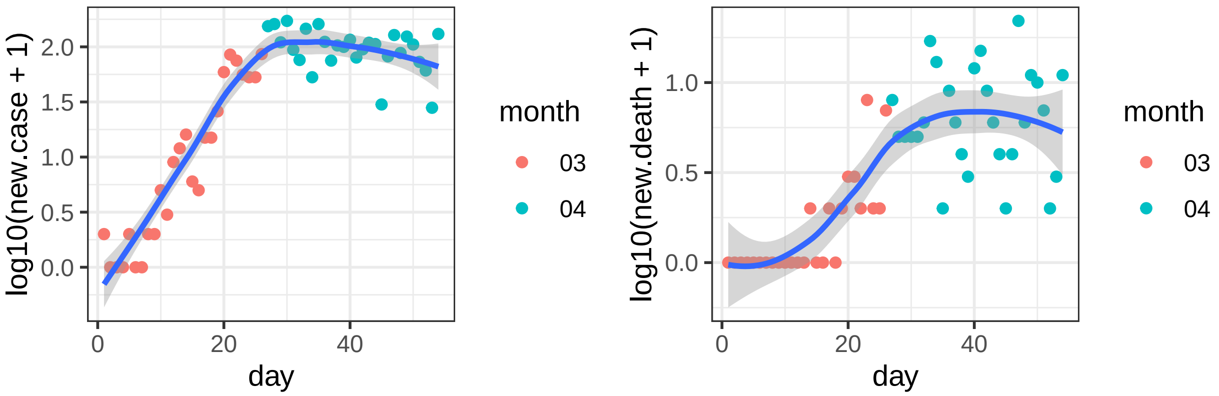
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Nevada



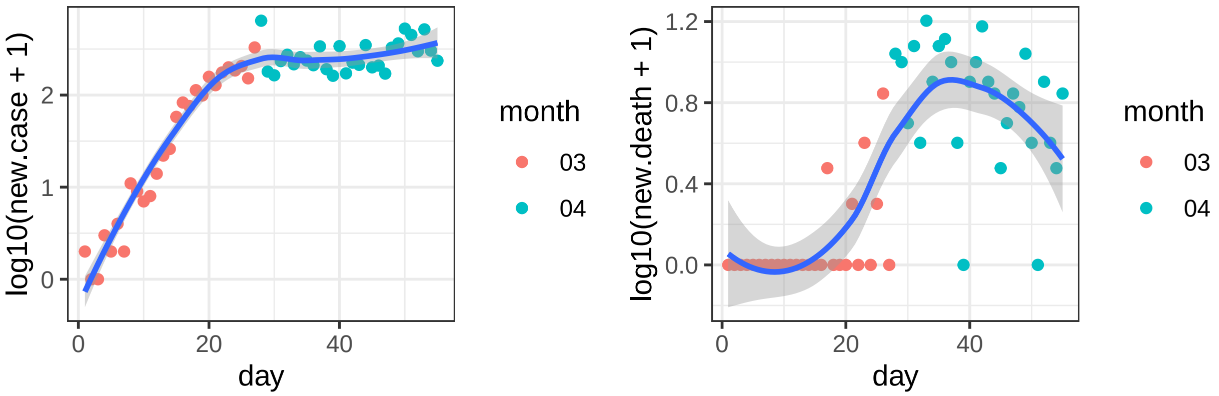
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Oklahoma



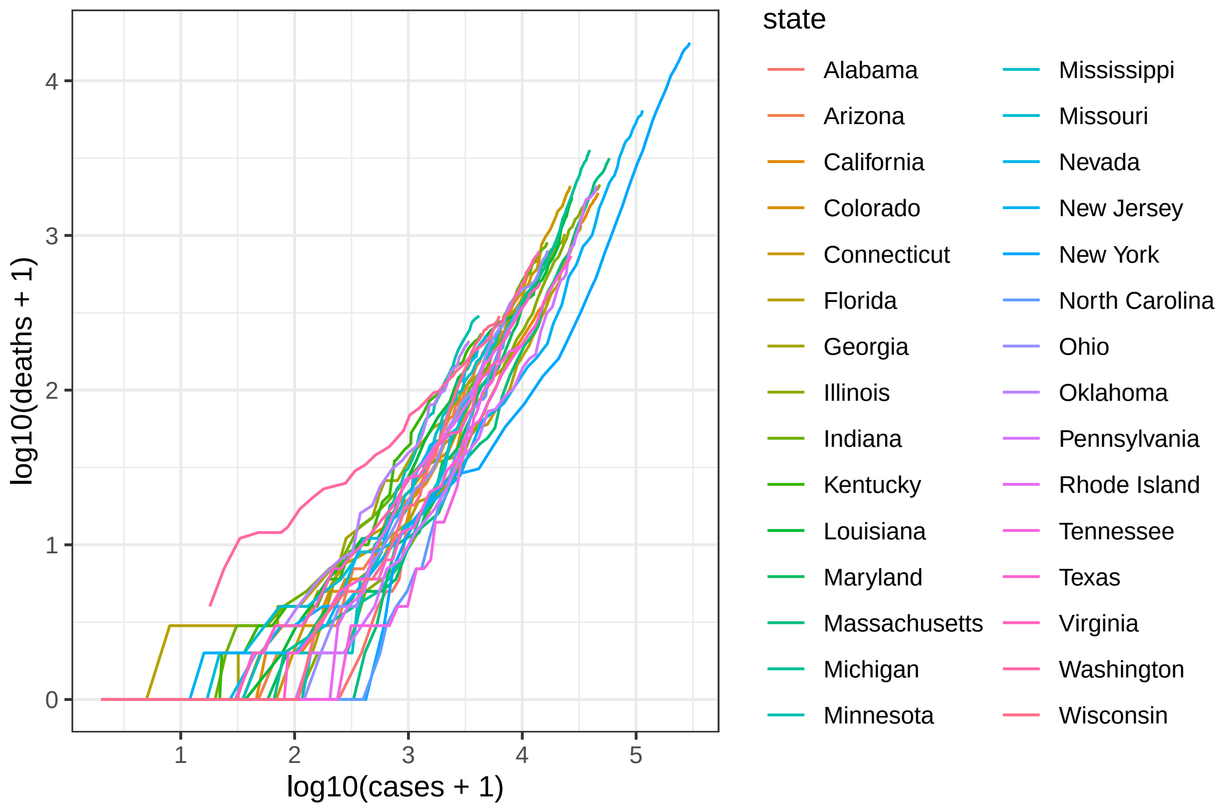
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Tennessee



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

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



data source: https://github.com/nytimes/covid−19−data

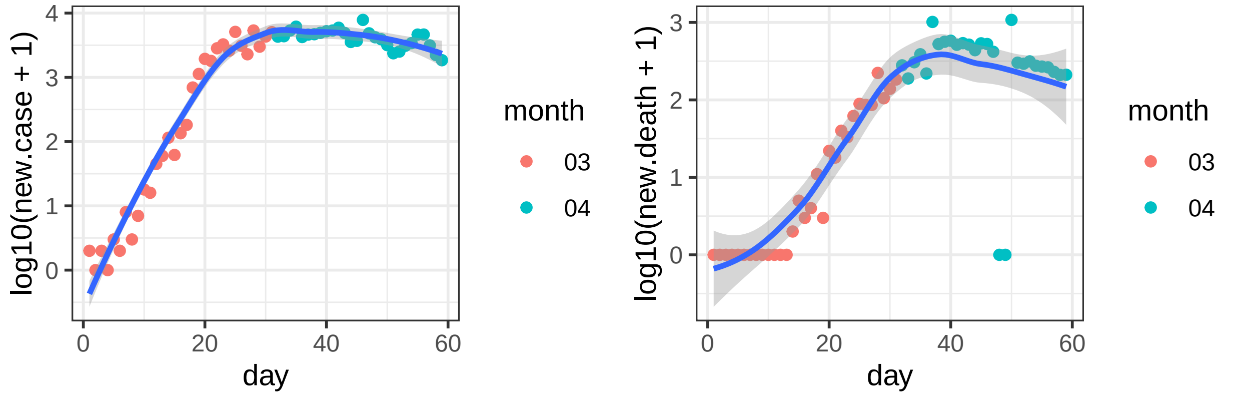
county level data

First check the 30 counties with the largest number of deaths.

```
##             date        county         state  fips  cases deaths
## 97099 2020-04-28 New York City      New York    NA 162348  12067
## 97098 2020-04-28        Nassau      New York 36059  35085   2039
## 96645 2020-04-28         Wayne      Michigan 26163  16173   1682
## 95994 2020-04-28          Cook      Illinois 17031  33449   1457
## 97118 2020-04-28       Suffolk      New York 36103  32724   1179
## 97126 2020-04-28   Westchester      New York 36119  28245   1096
## 97025 2020-04-28         Essex    New Jersey 34013  13190   1090
## 97020 2020-04-28        Bergen    New Jersey 34003  15251   1002
## 95608 2020-04-28   Los Angeles    California  6037  20976   1000
## 95702 2020-04-28     Fairfield   Connecticut  9001  10874    747
## 96560 2020-04-28     Middlesex Massachusetts 25017  13417    731
## 97027 2020-04-28        Hudson    New Jersey 34017  14309    722
## 96626 2020-04-28       Oakland      Michigan 26125   7012    654
## 95703 2020-04-28      Hartford   Connecticut  9003   5224    643
## 97038 2020-04-28         Union    New Jersey 34039  12188    627
## 96613 2020-04-28        Macomb      Michigan 26099   5339    572
## 97501 2020-04-28  Philadelphia  Pennsylvania 42101  13445    516
## 97030 2020-04-28     Middlesex    New Jersey 34023  11102    515
## 95706 2020-04-28     New Haven   Connecticut  9009   7089    478
## 97034 2020-04-28       Passaic    New Jersey 34031  11755    475
## 96564 2020-04-28       Suffolk Massachusetts 25025  12140    469
## 96562 2020-04-28       Norfolk Massachusetts 25021   5567    448
## 98102 2020-04-28          King    Washington 53033   6056     429
## 96480 2020-04-28       Orleans     Louisiana 22071   6380     410
## 96556 2020-04-28         Essex Massachusetts 25009   7972     383
## 97032 2020-04-28        Morris    New Jersey 34027   5128     377
## 97110 2020-04-28      Rockland      New York 36087  11453     359
## 96558 2020-04-28       Hampden Massachusetts 25013   3546     346
## 97033 2020-04-28         Ocean    New Jersey 34029   6151     342
## 96470 2020-04-28     Jefferson     Louisiana 22051   6135     340
```

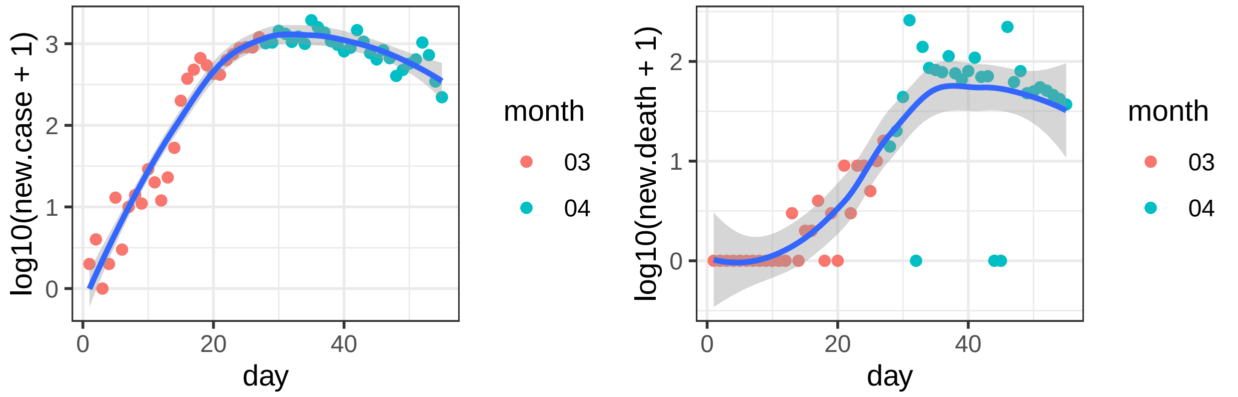
For these 30 counties, I check the number of new cases and the number of new deaths.

New York City_New York



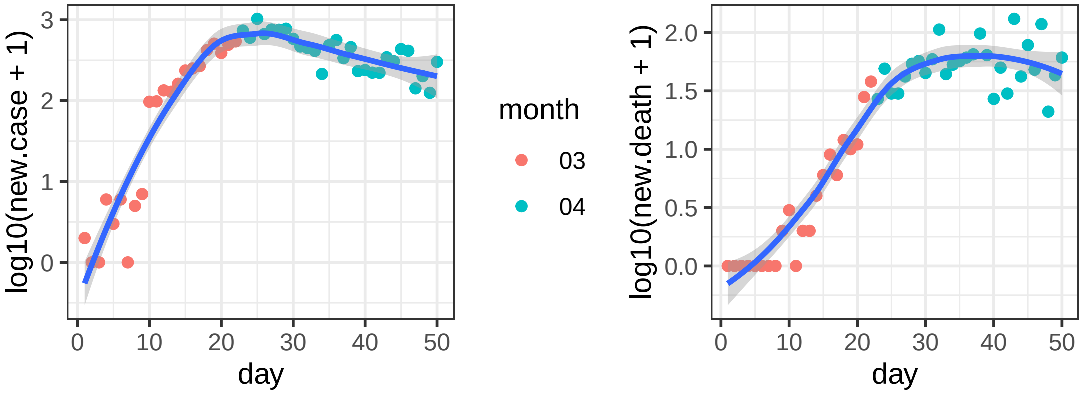
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nassau_New York



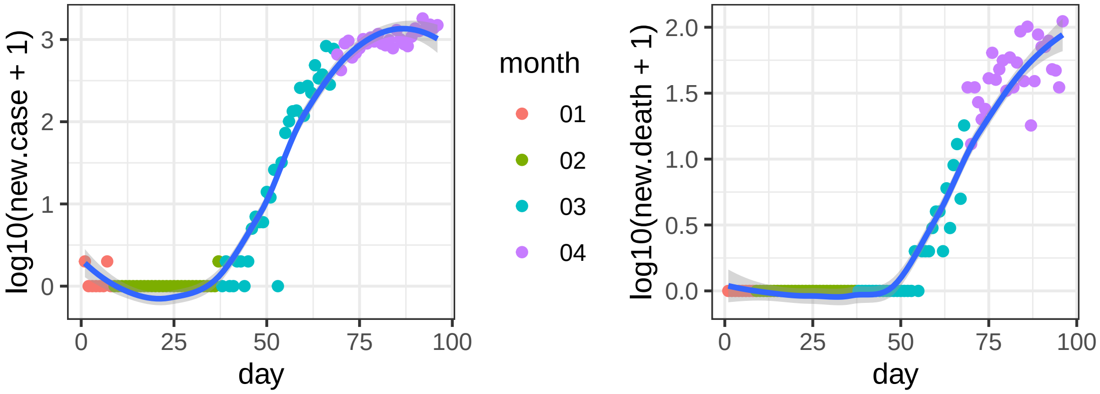
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



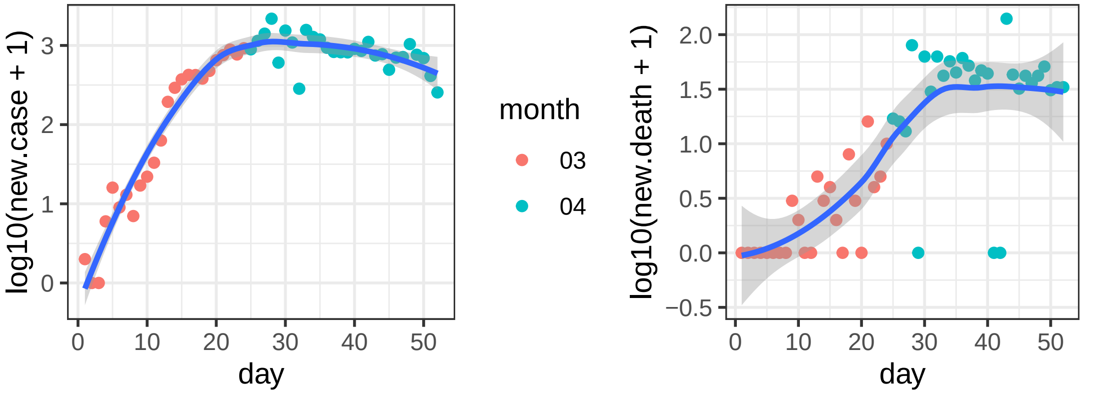
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Cook_Illinois



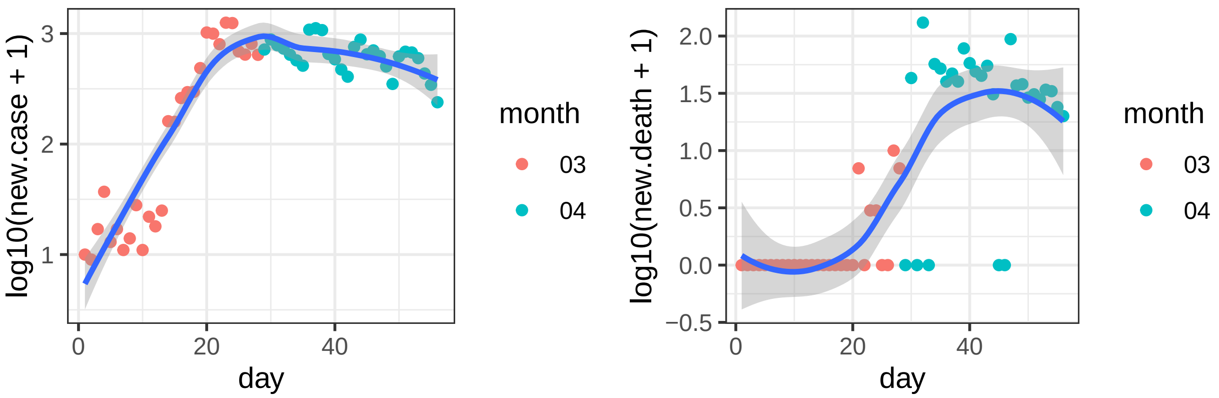
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Suffolk_New York



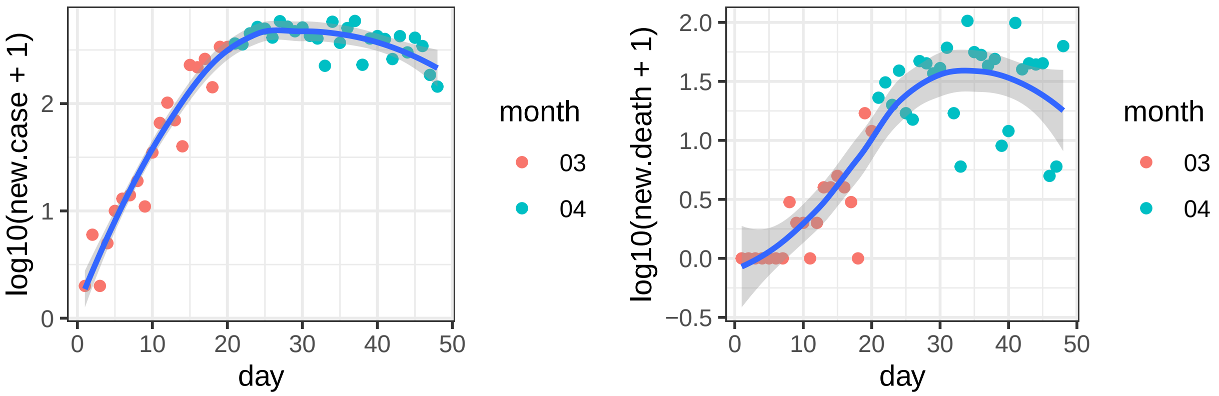
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Westchester_New York



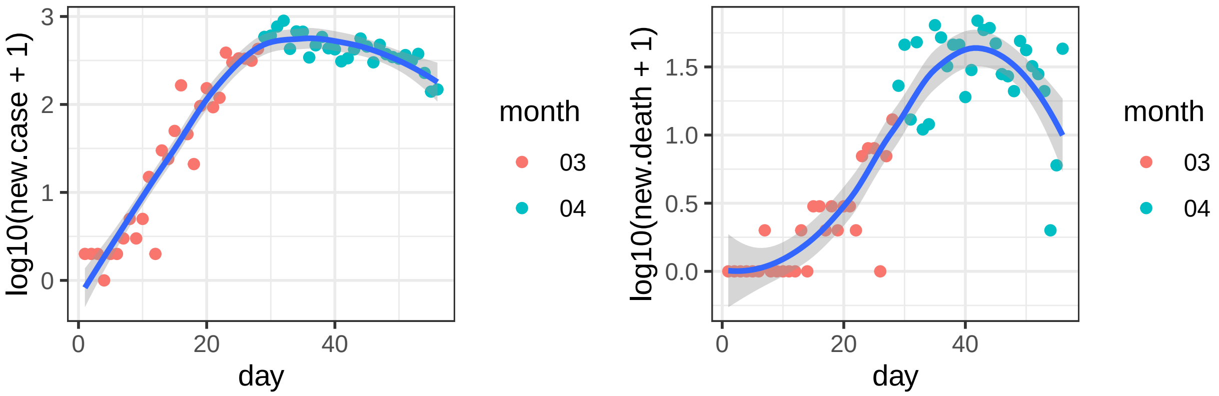
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Essex_New Jersey

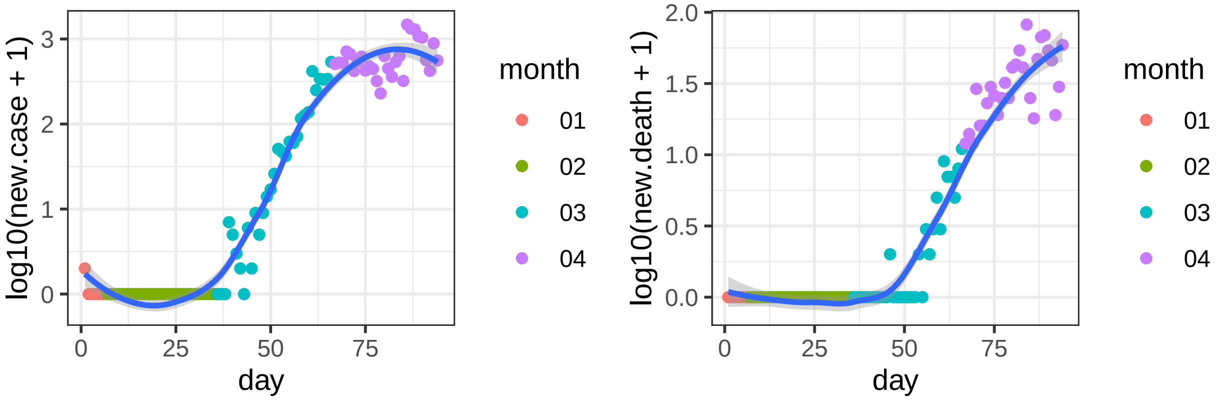


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

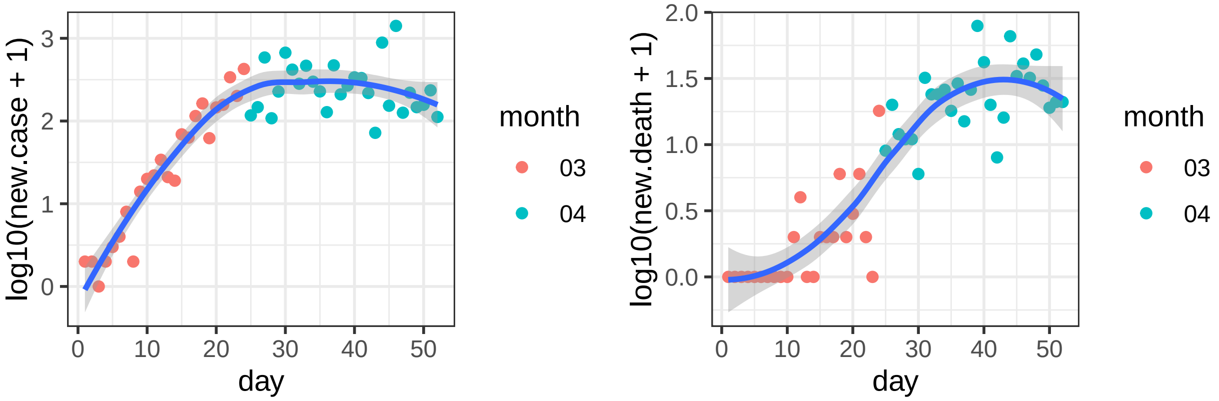
Bergen_New Jersey



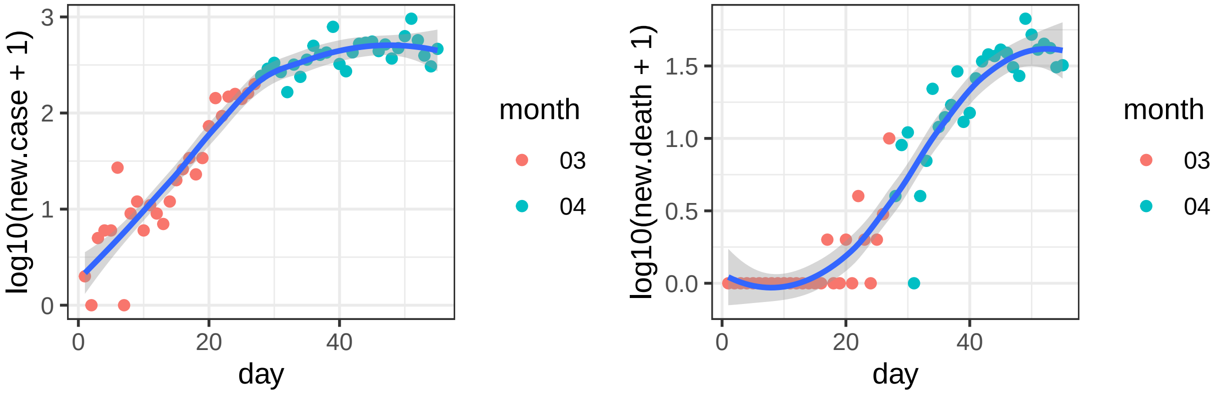
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04


Los Angeles_California

data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

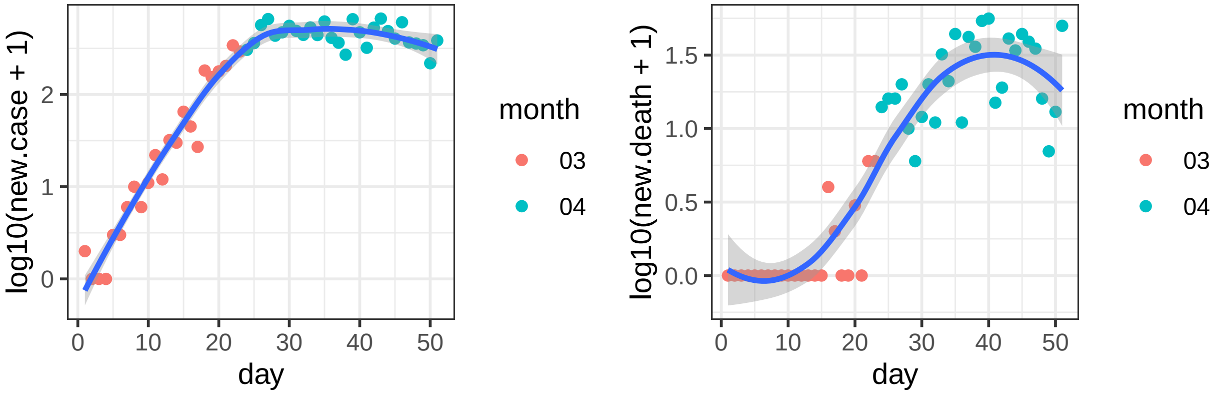

Fairfield_Connecticut

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


Middlesex_Massachusetts

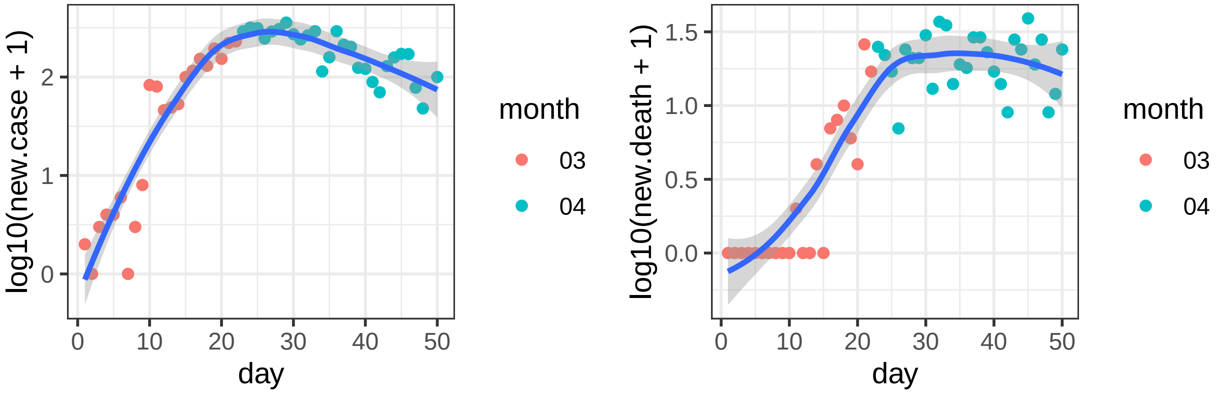
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Hudson_New Jersey



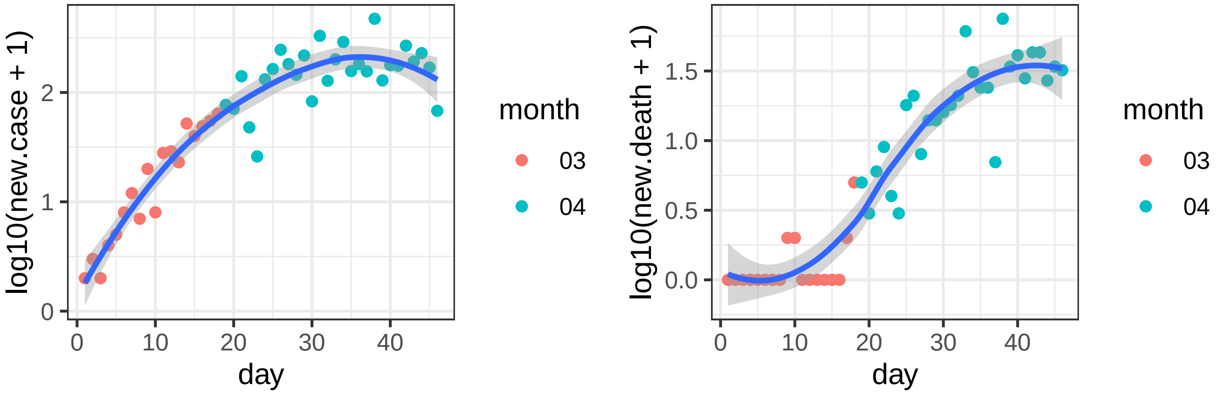
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



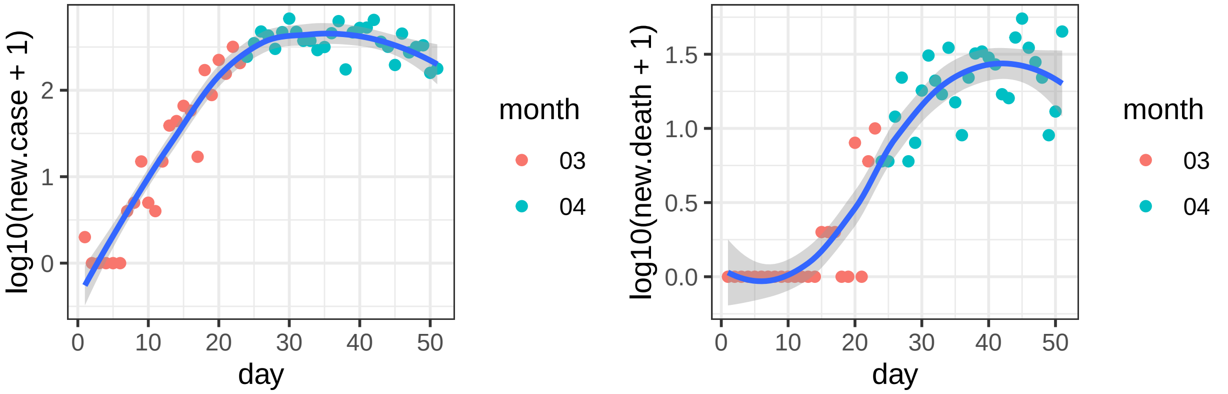
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hartford_Connecticut



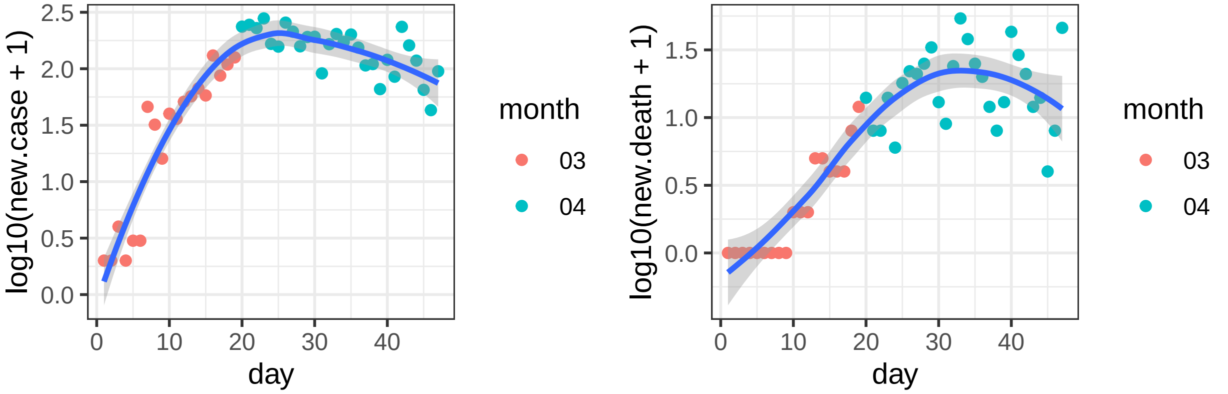
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Union_New Jersey



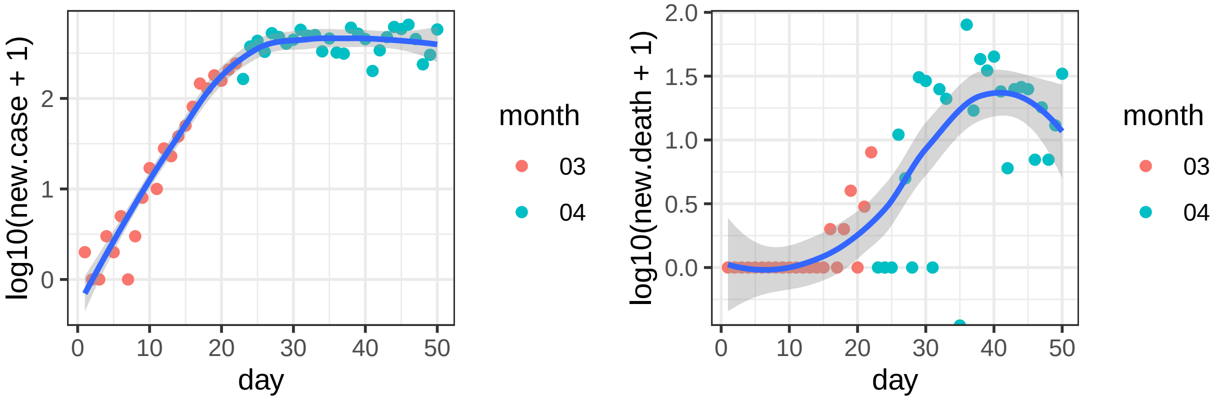
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Macomb_Michigan



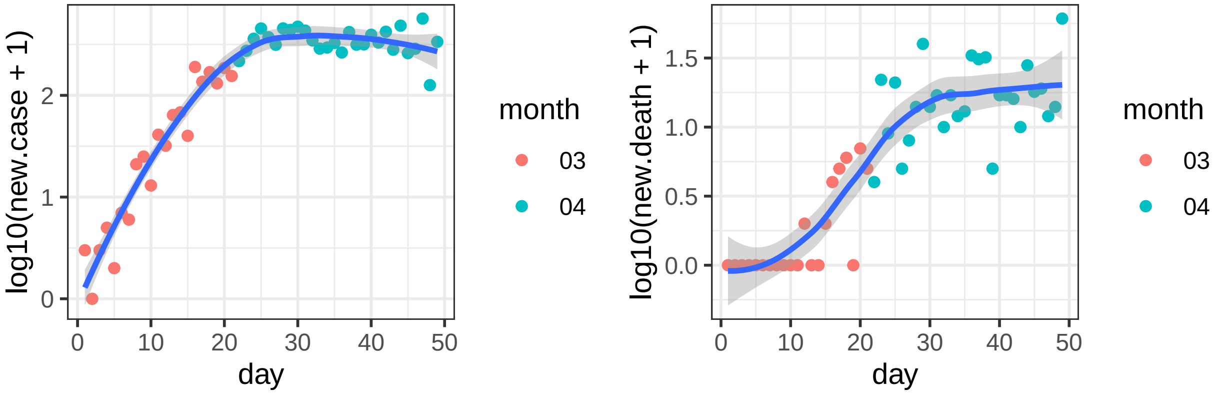
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Philadelphia_Pennsylvania



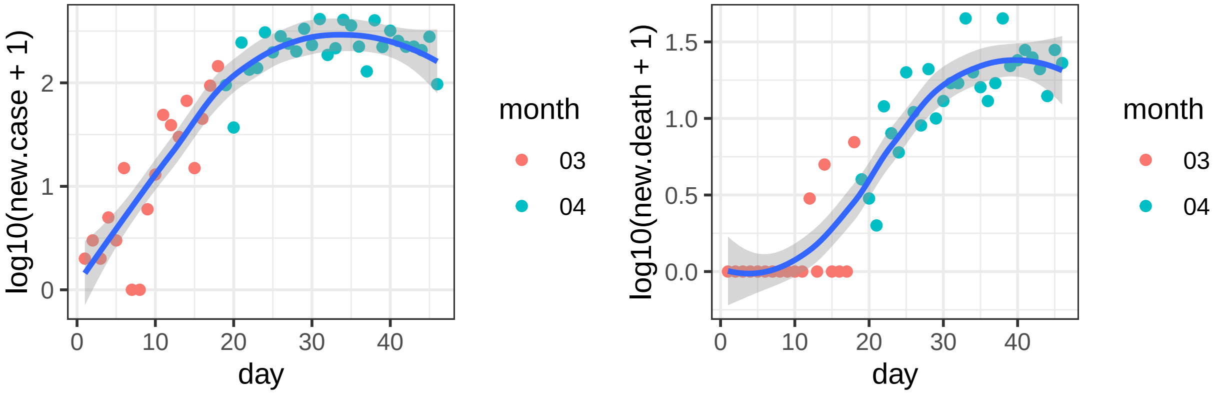
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



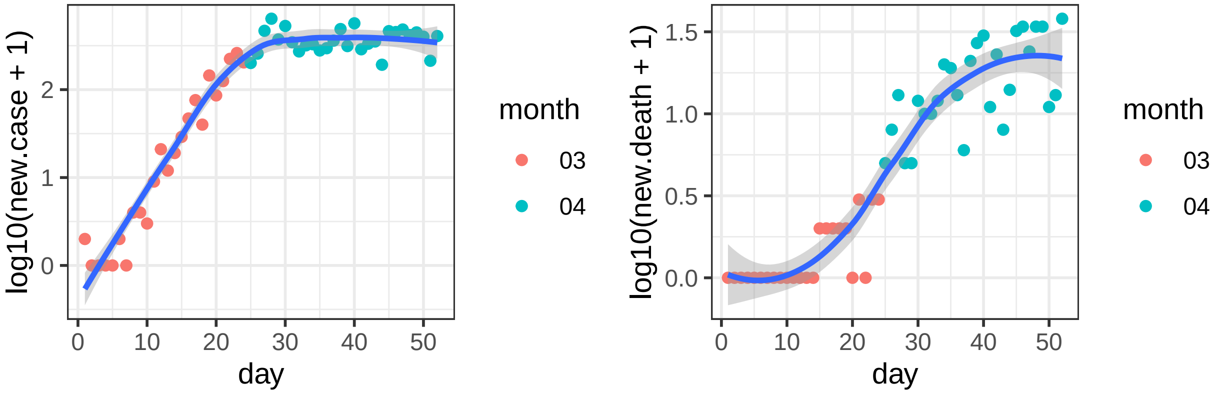
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

New Haven_Connecticut



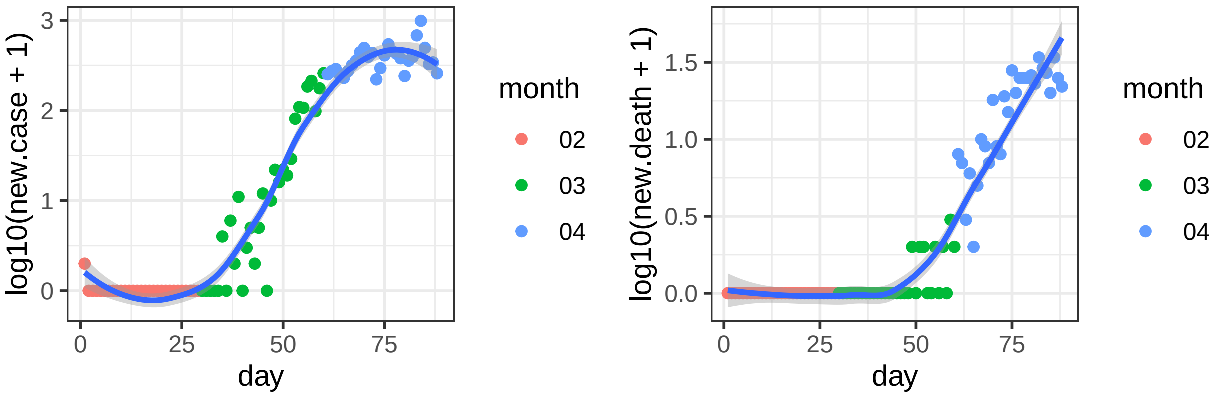
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Passaic_New Jersey



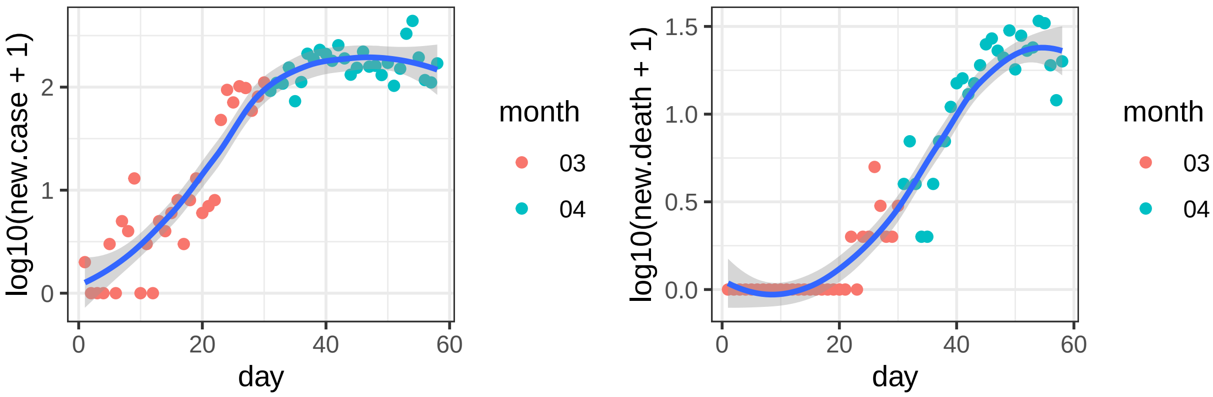
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Suffolk_Massachusetts



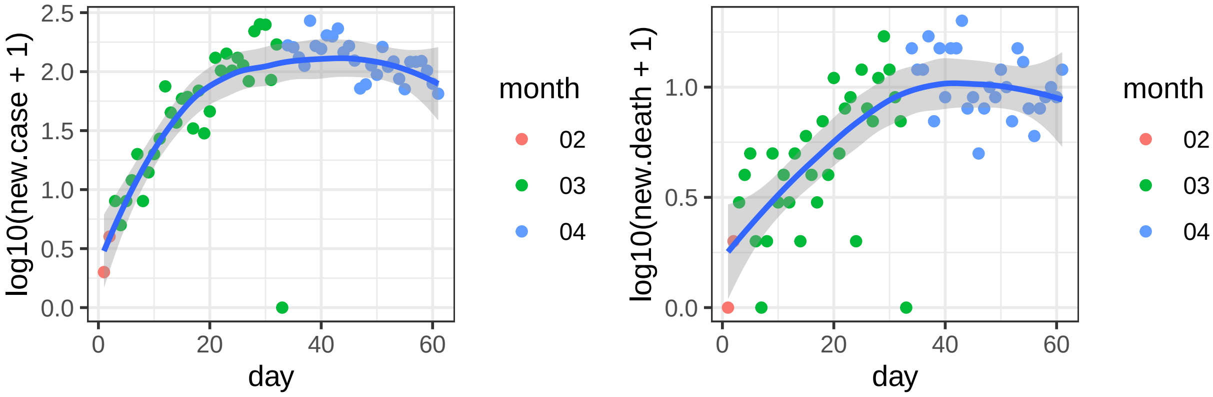
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Norfolk_Massachusetts



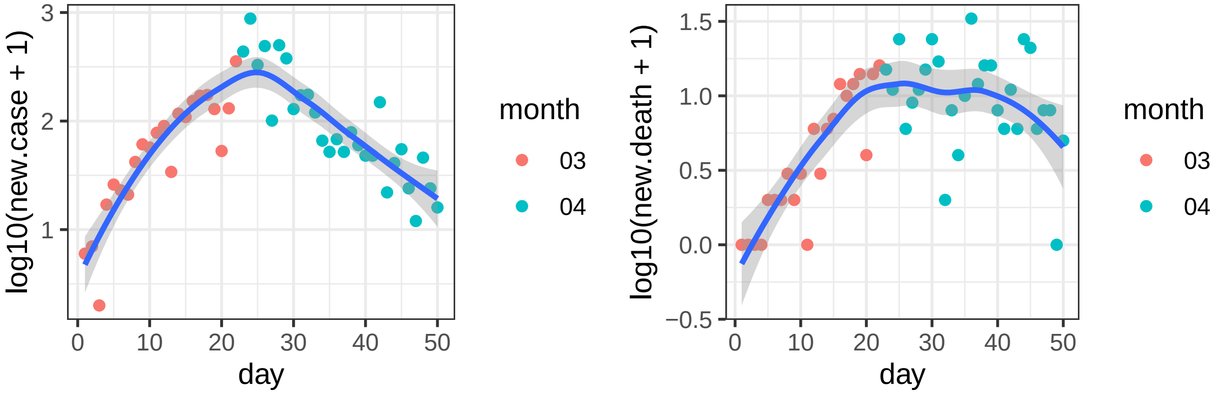
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

King_Washington



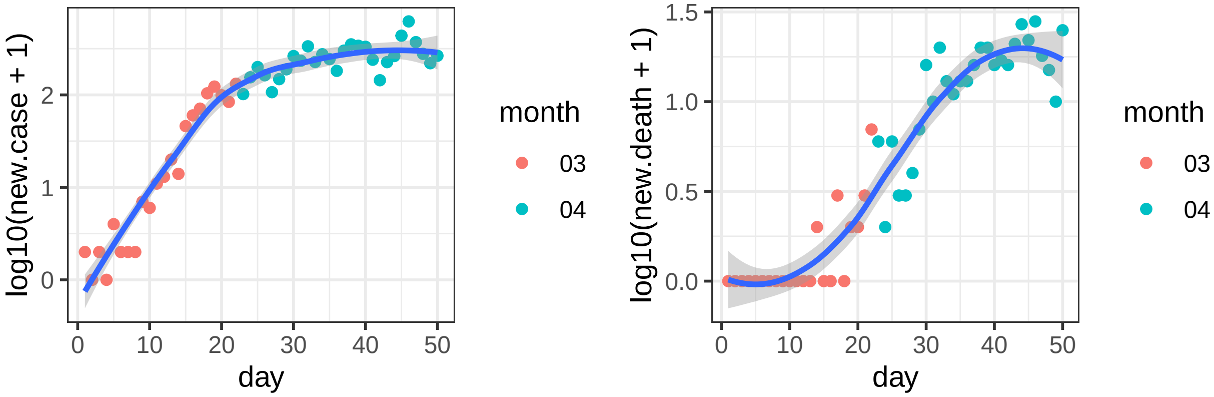
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Orleans_Louisiana



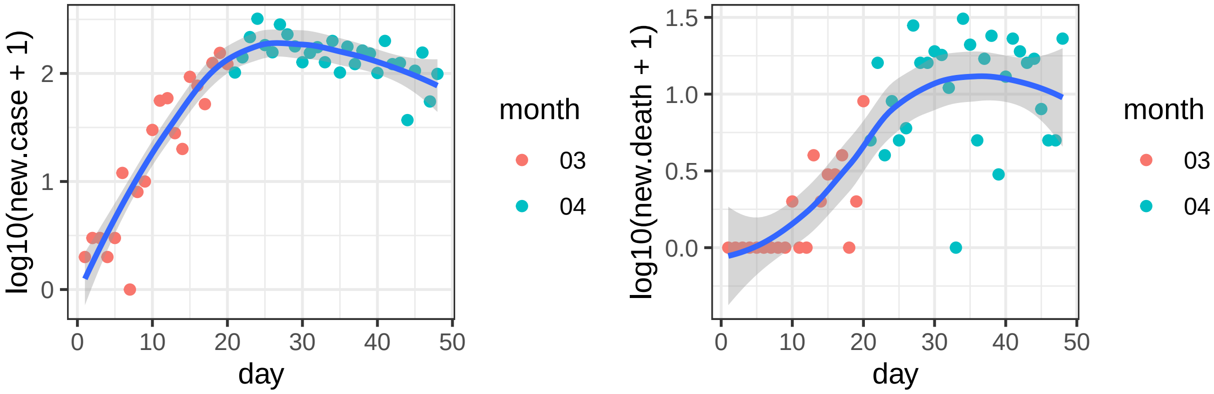
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Essex_Massachusetts



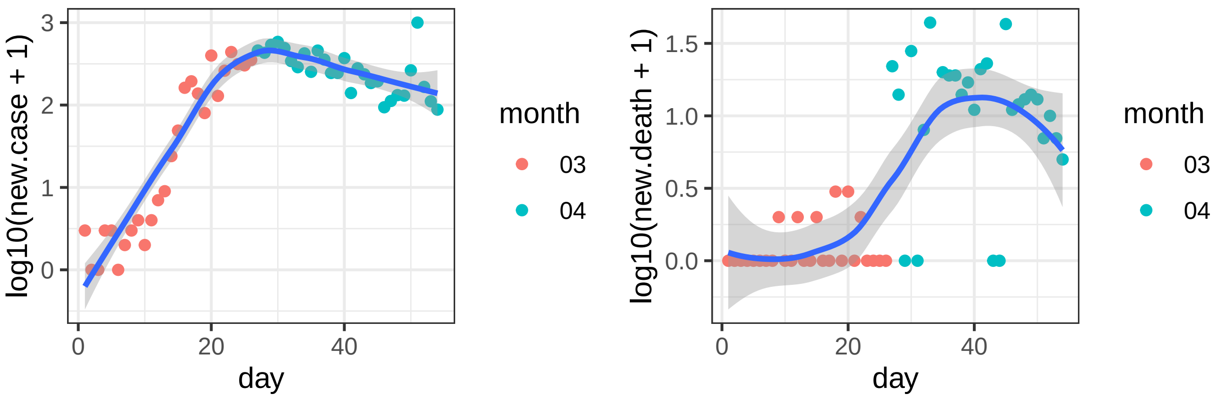
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Morris_New Jersey



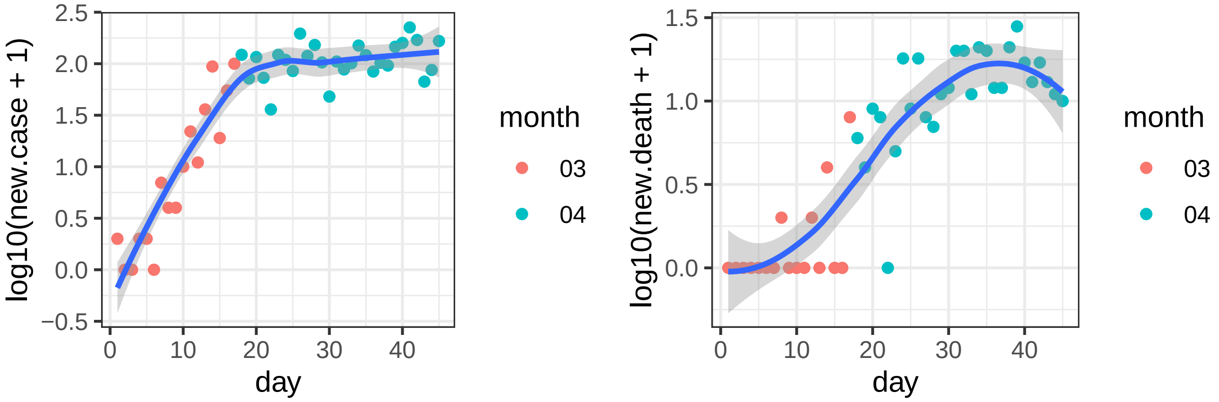
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Rockland_New York



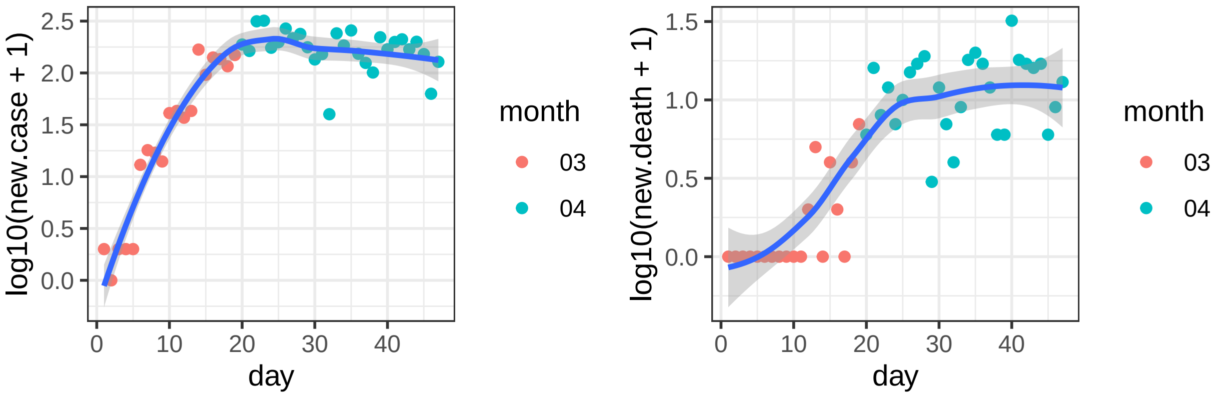
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Hampden_Massachusetts

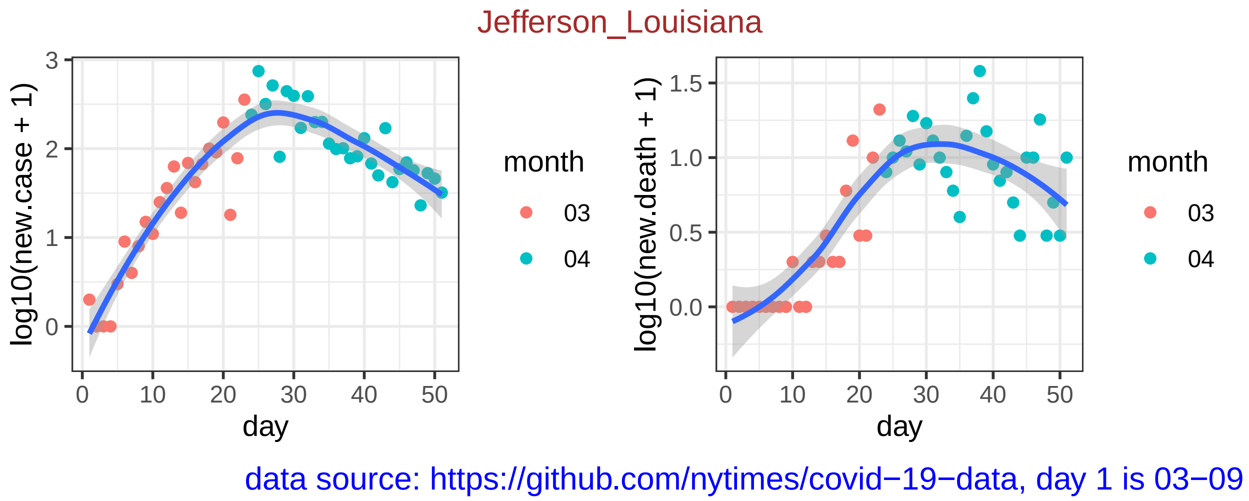


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Ocean_New Jersey



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



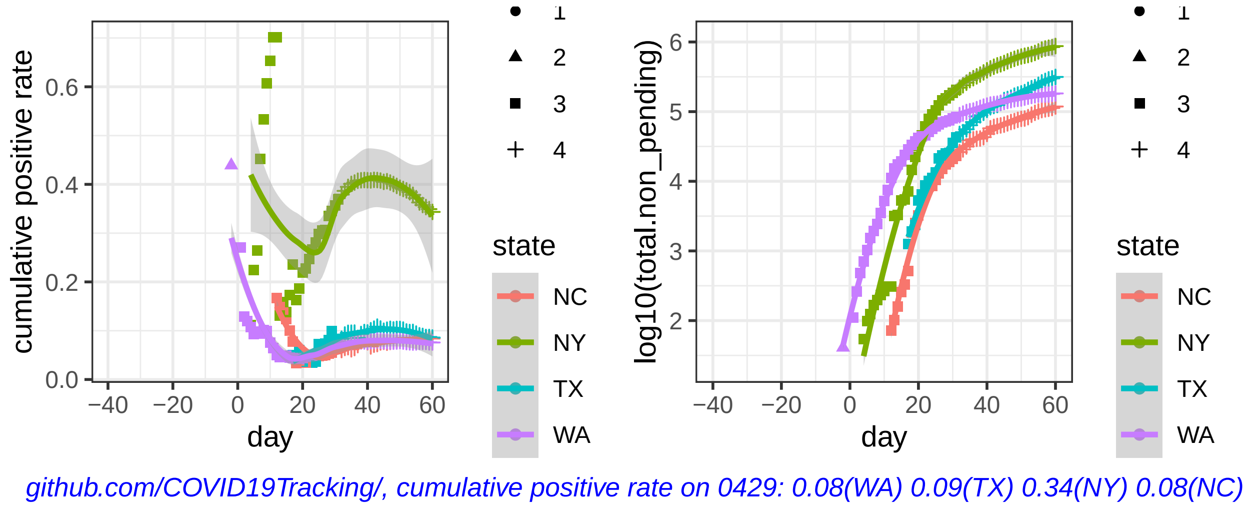
Jefferson_Louisiana

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

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.



github.com/COVID19Tracking/, cumulative positive rate on 0429: 0.08(WA) 0.09(TX) 0.34(NY) 0.08(NC)

Session information

```
sessionInfo()
```

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