Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-24

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.
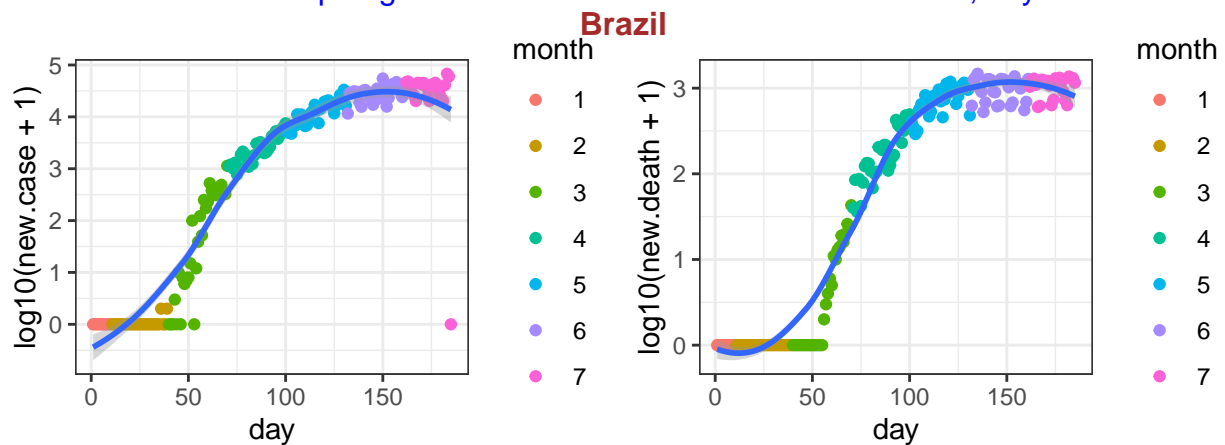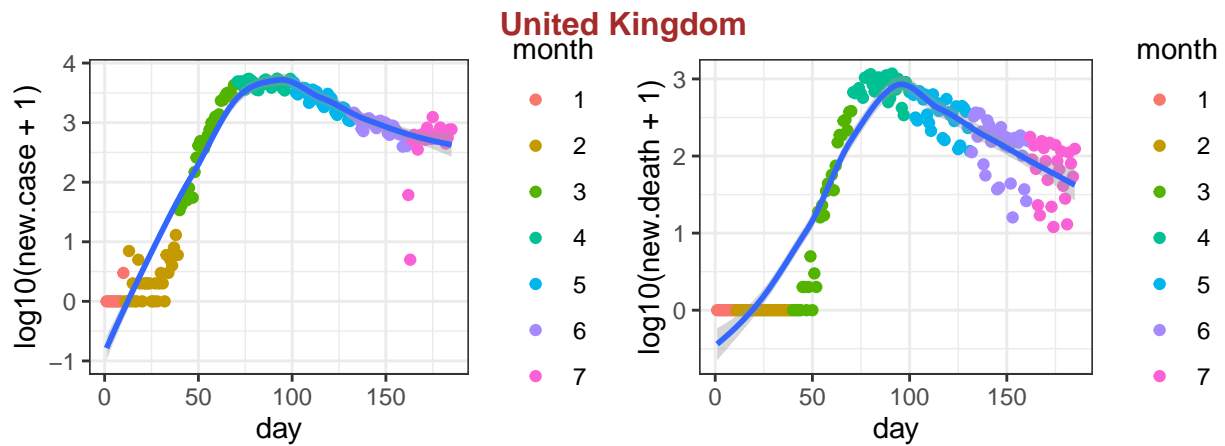
cumulative number of deaths on 7.24.20
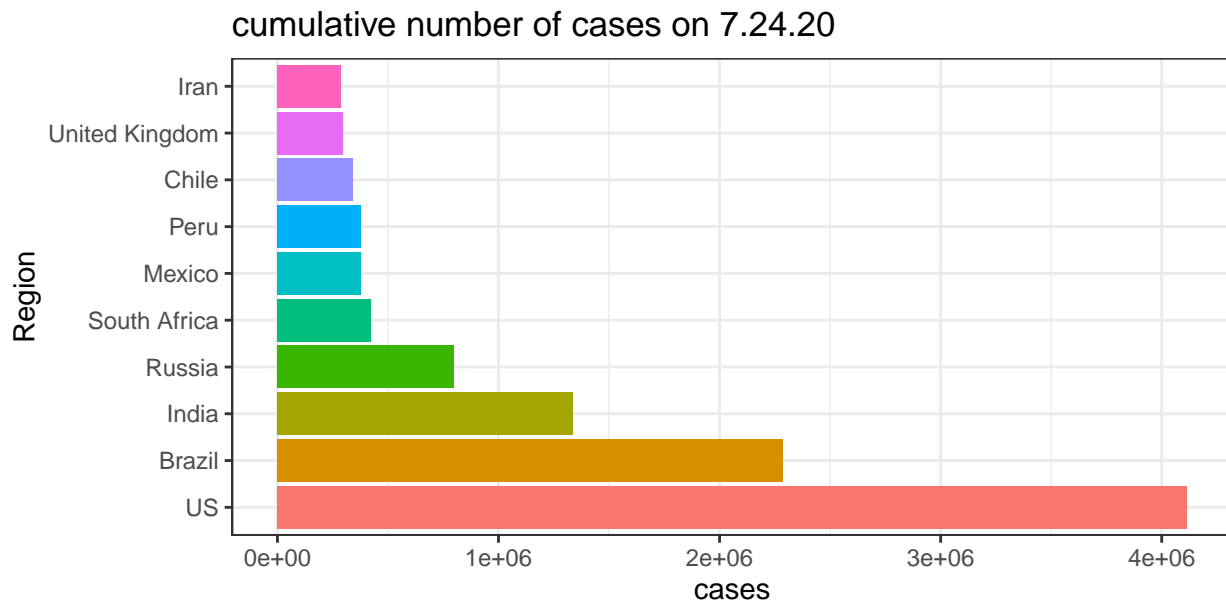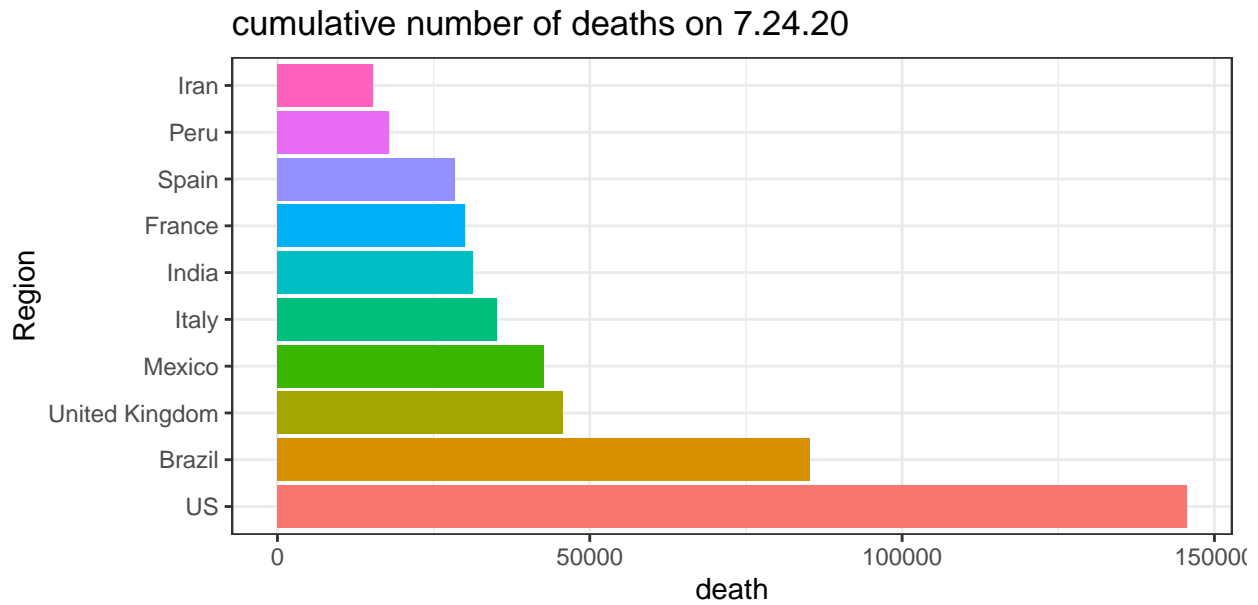
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.
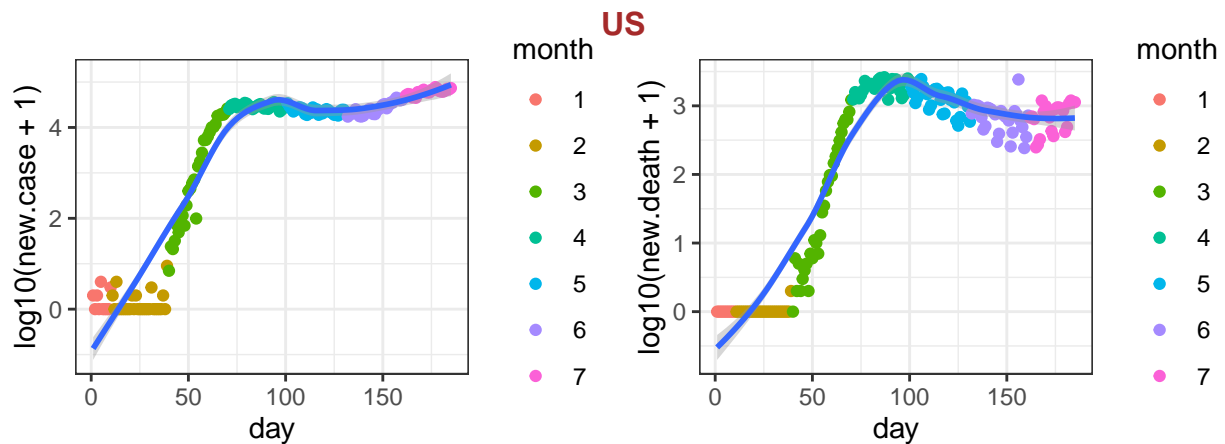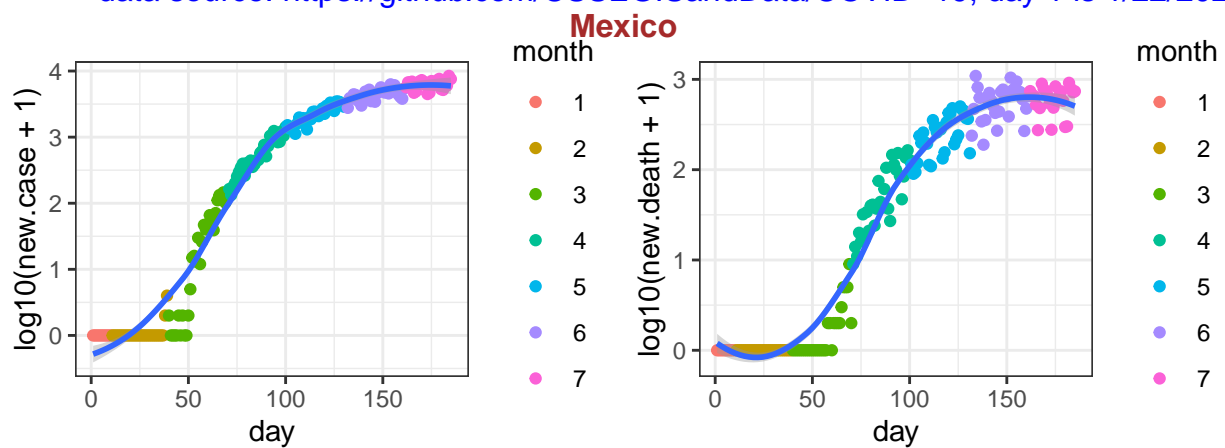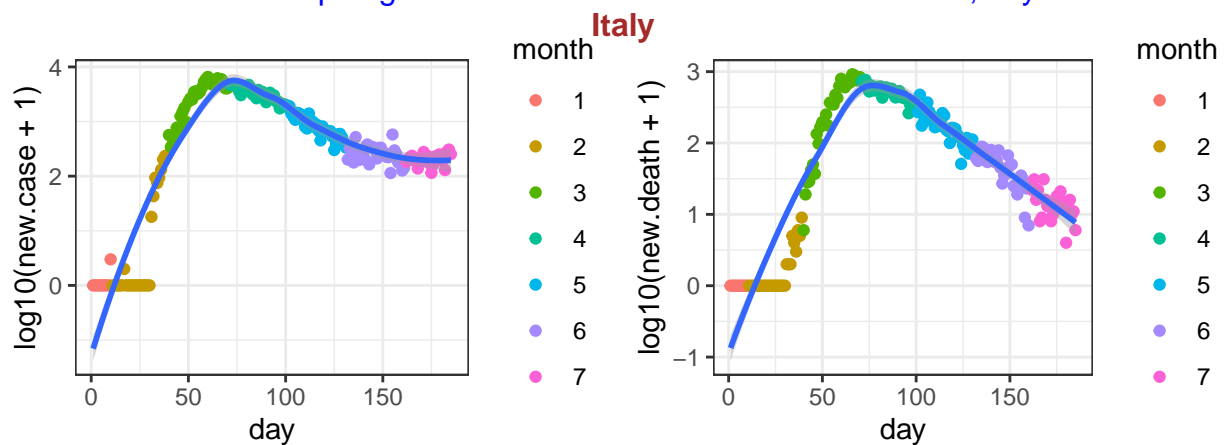

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

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

Mexico

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

India

France

Spain

Peru

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

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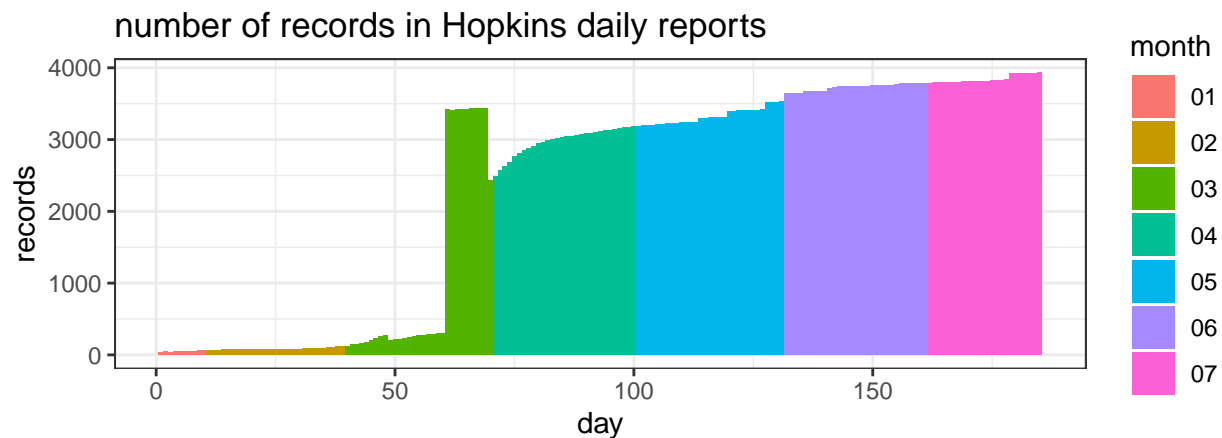
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

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.
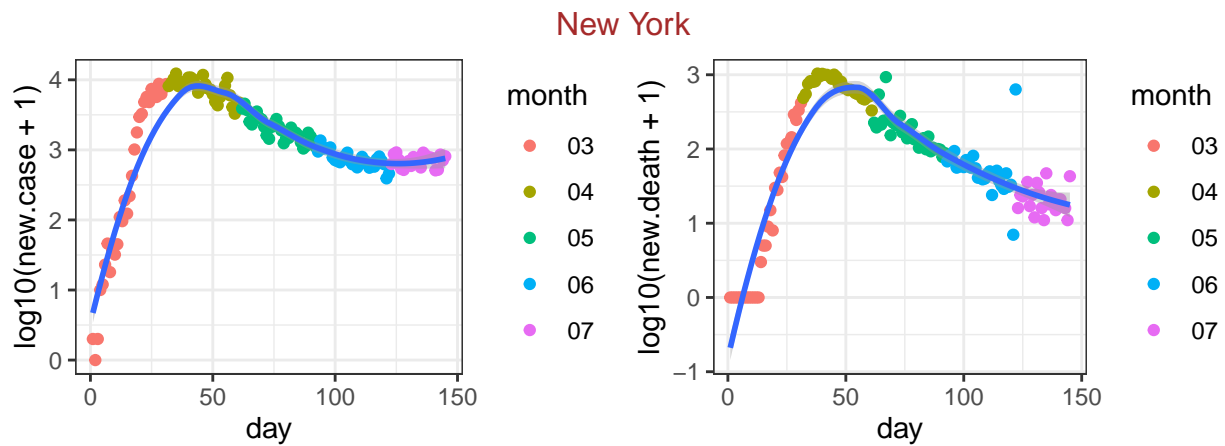
The currente date is

```
## [1] "2020-07-23"
```
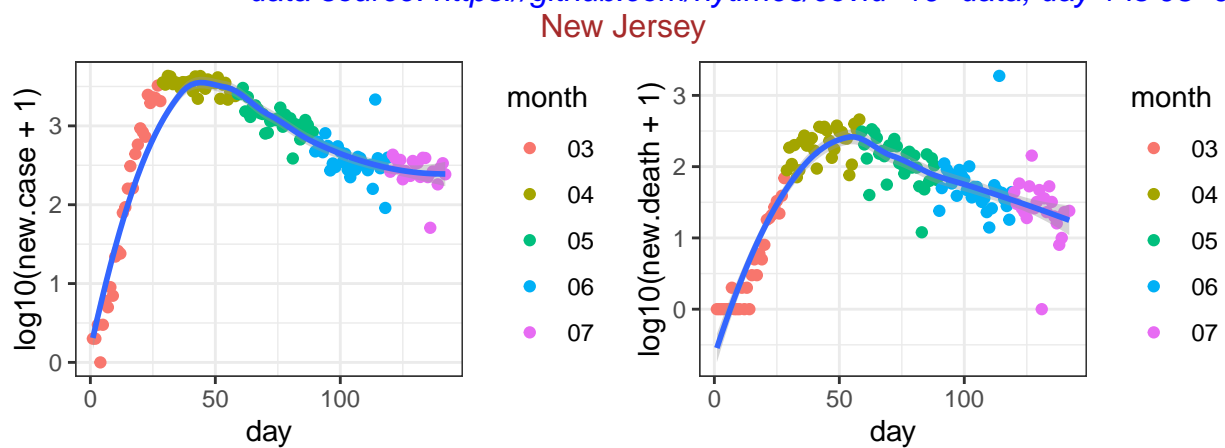
state level data

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

```
##                date          state fips   cases deaths
## 7858 2020-07-23       New York   36 414405  32270
## 7856 2020-07-23     New Jersey   34 179807  15730
## 7847 2020-07-23  Massachusetts   25 114647   8484
## 7829 2020-07-23     California    6 433175   8190
## 7839 2020-07-23       Illinois   17 168766   7563
## 7865 2020-07-23   Pennsylvania   42 109153   7142
## 7848 2020-07-23       Michigan   26  84585   6397
## 7834 2020-07-23        Florida   12 389860   5517
## 7871 2020-07-23          Texas   48 376104   4714
## 7831 2020-07-23    Connecticut    9  48232   4410
## 7844 2020-07-23      Louisiana   22 101762   3686
## 7846 2020-07-23       Maryland   24  81396   3409
## 7835 2020-07-23        Georgia   13 143857   3288
## 7862 2020-07-23           Ohio   39  80186   3256
## 7827 2020-07-23        Arizona    4 153029   3067
## 7840 2020-07-23        Indiana   18  61079   2880
## 7875 2020-07-23       Virginia   51  81237   2054
## 7830 2020-07-23       Colorado    8  42409   1788
## 7859 2020-07-23 North Carolina   37 107140   1754
## 7849 2020-07-23      Minnesota   27  48761   1601
## 7876 2020-07-23     Washington   53  52255   1548
## 7850 2020-07-23    Mississippi   28  48053   1436
## 7825 2020-07-23        Alabama    1  74212   1397
## 7868 2020-07-23 South Carolina   45  76606   1334
## 7851 2020-07-23       Missouri   29  39923   1212
## 7867 2020-07-23   Rhode Island   44  18148   1001
## 7870 2020-07-23      Tennessee   47  84652    913
## 7878 2020-07-23      Wisconsin   55  49751    887
## 7841 2020-07-23           Iowa   19  40796    820
## 7854 2020-07-23         Nevada   32  40088    709
```
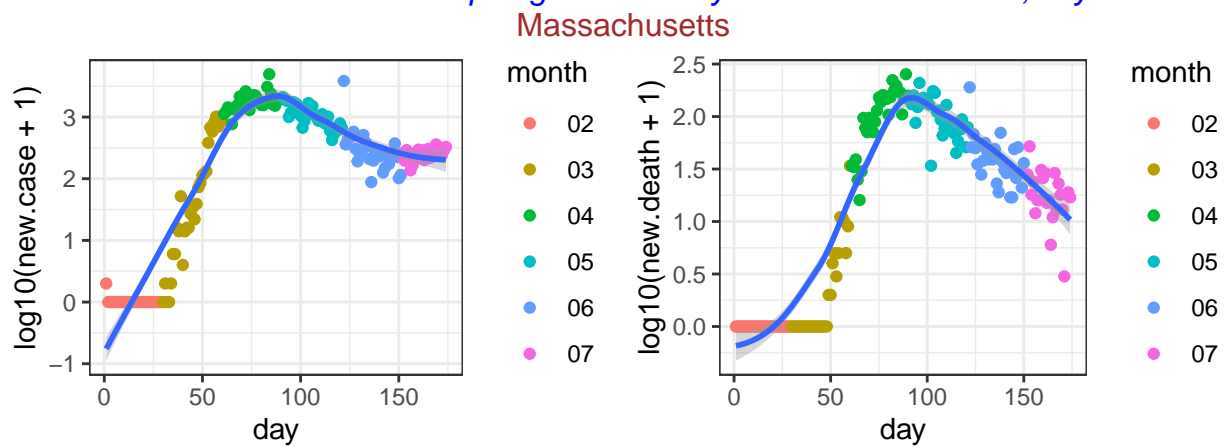
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.
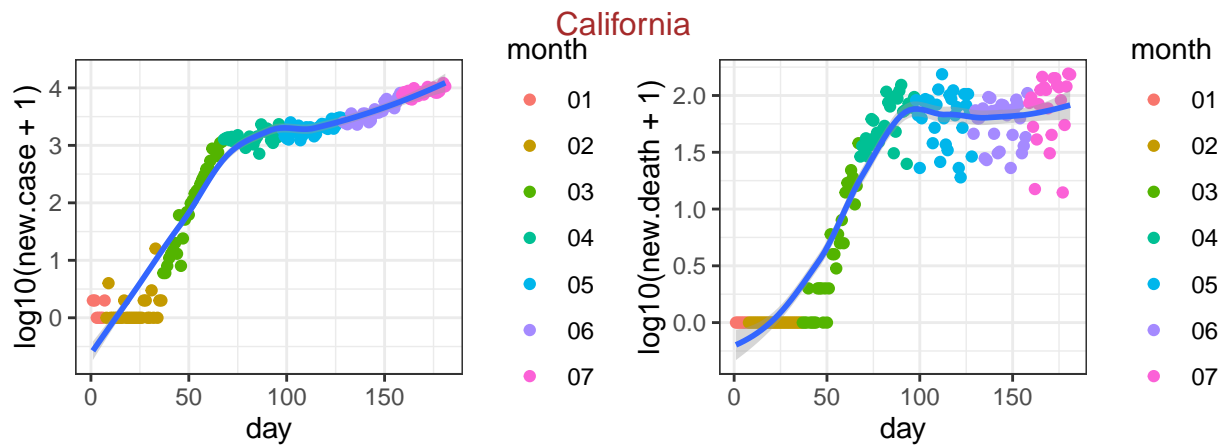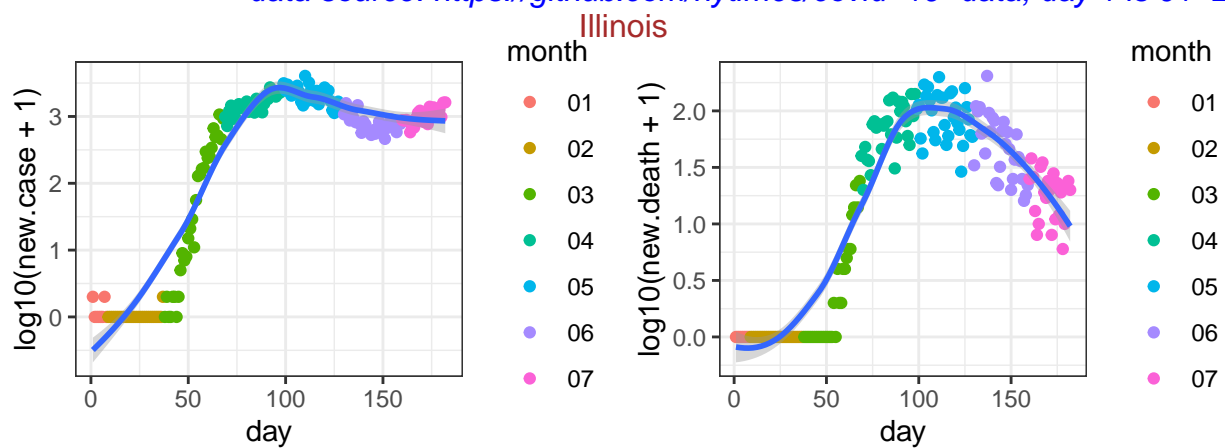
New York

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

Massachusetts

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

8

California

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

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

Michigan

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

Florida

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

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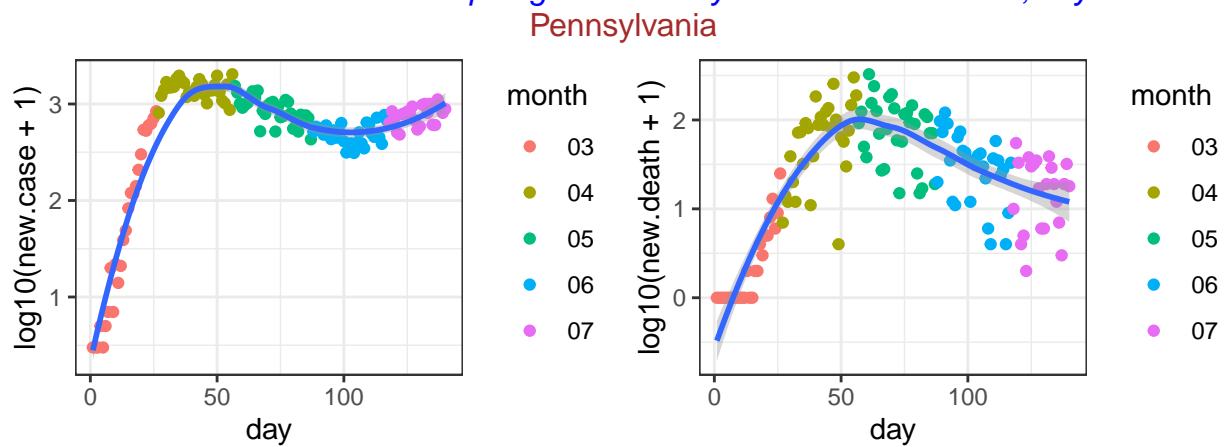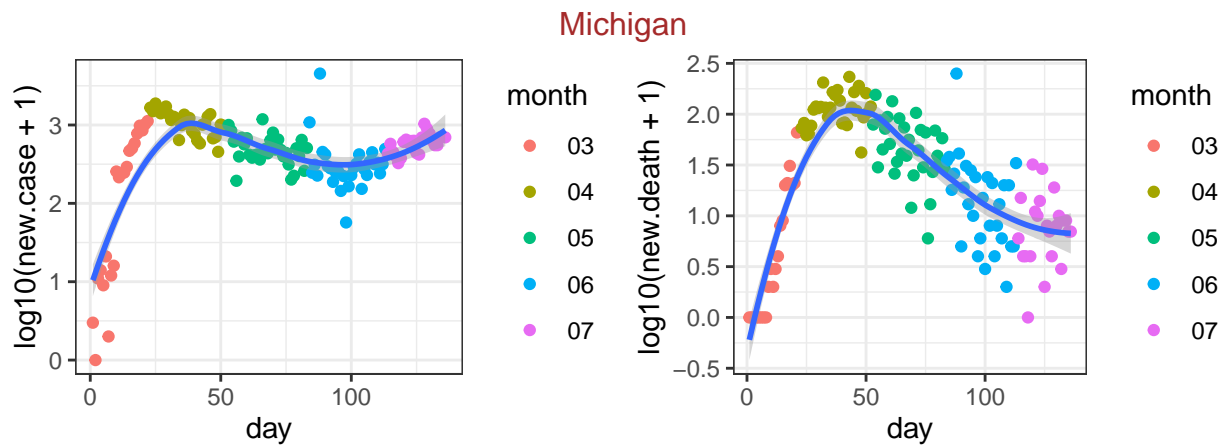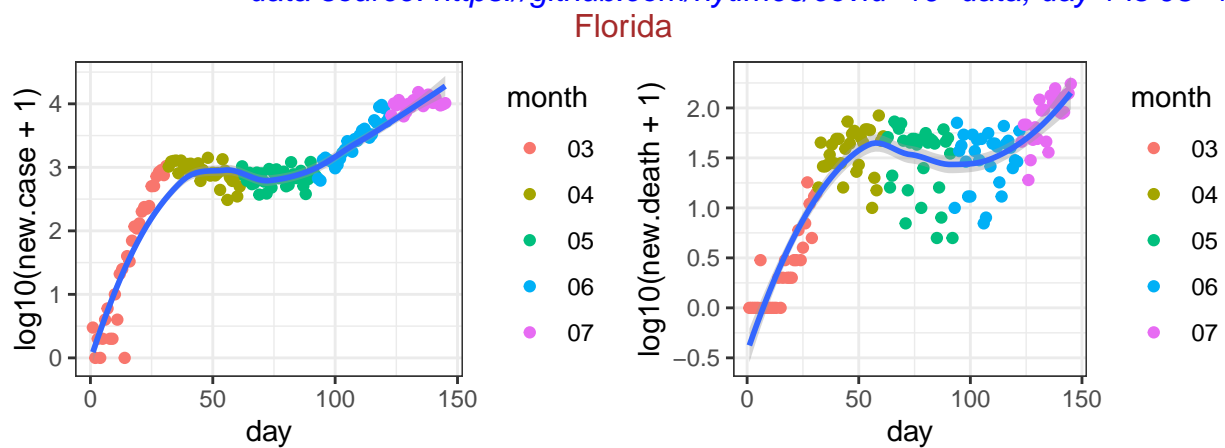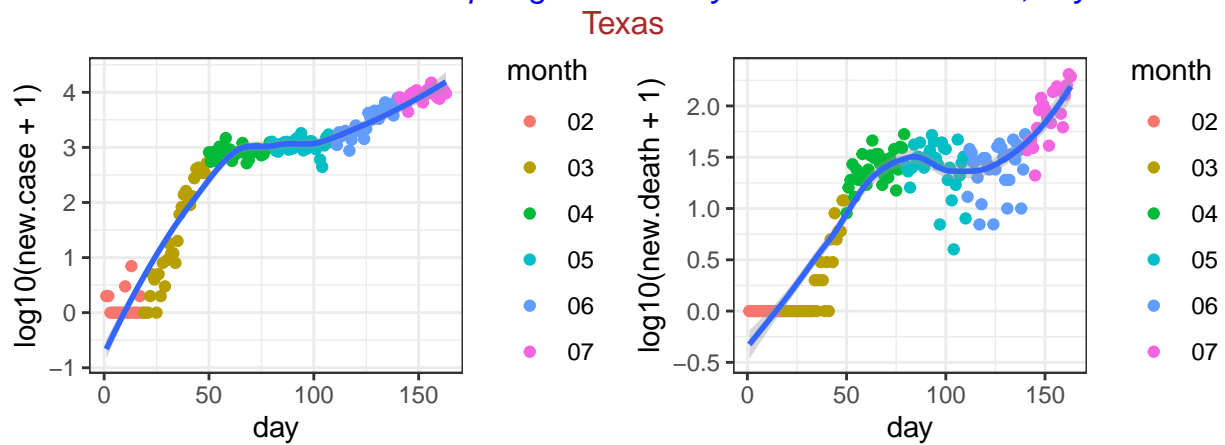
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

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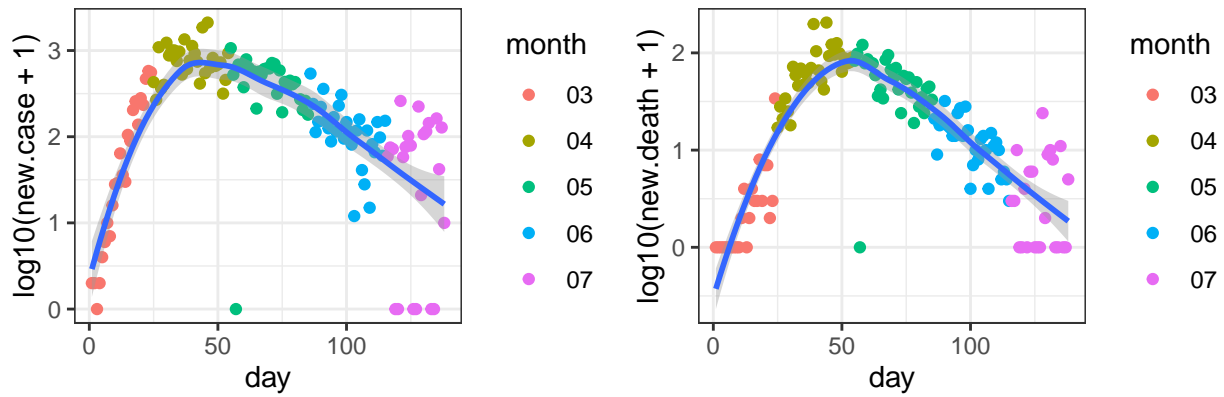
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

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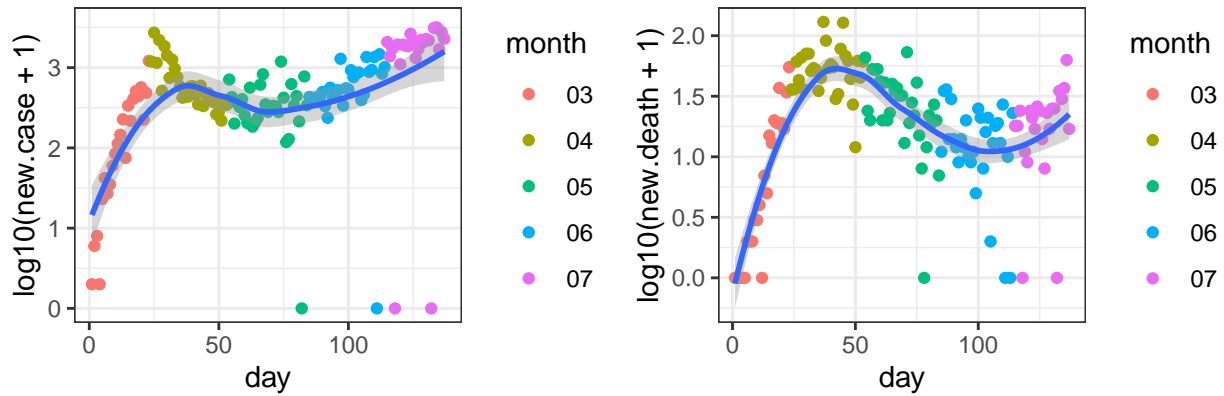
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09
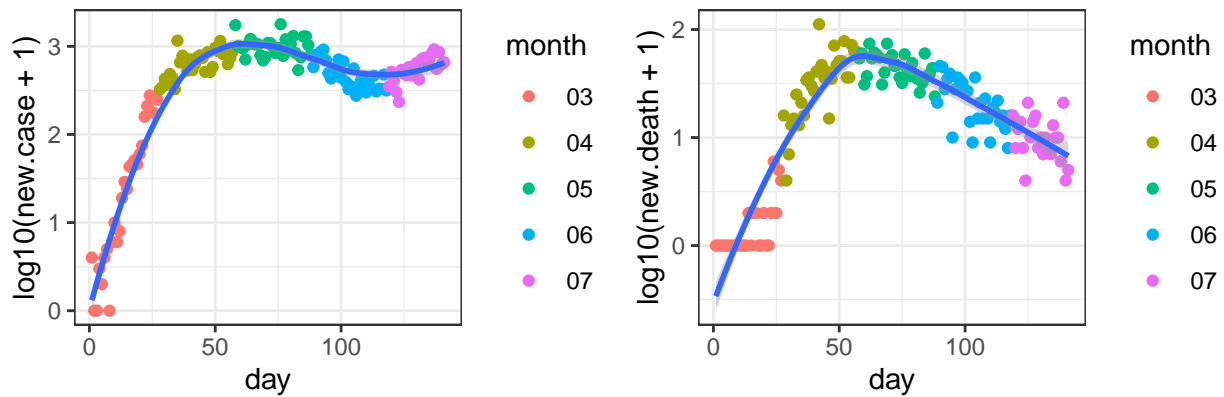
Maryland

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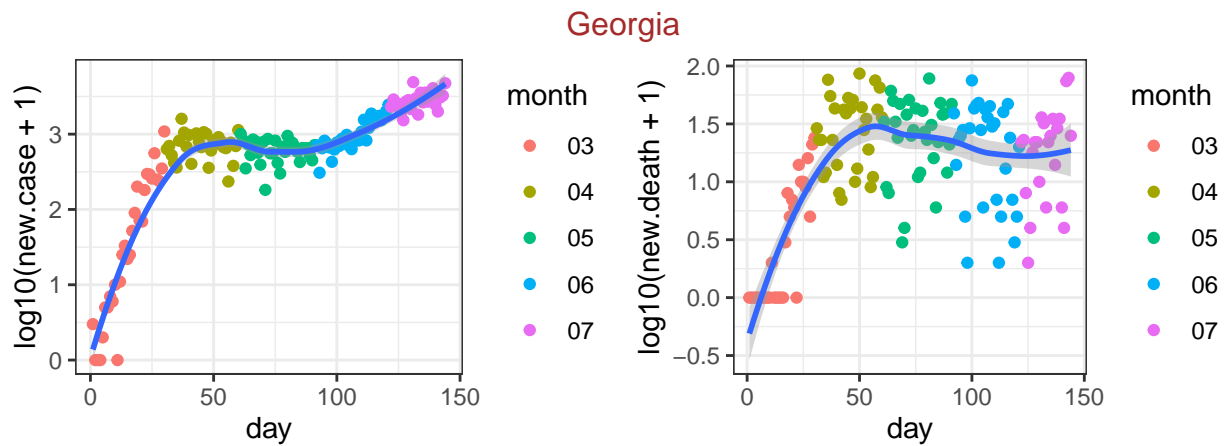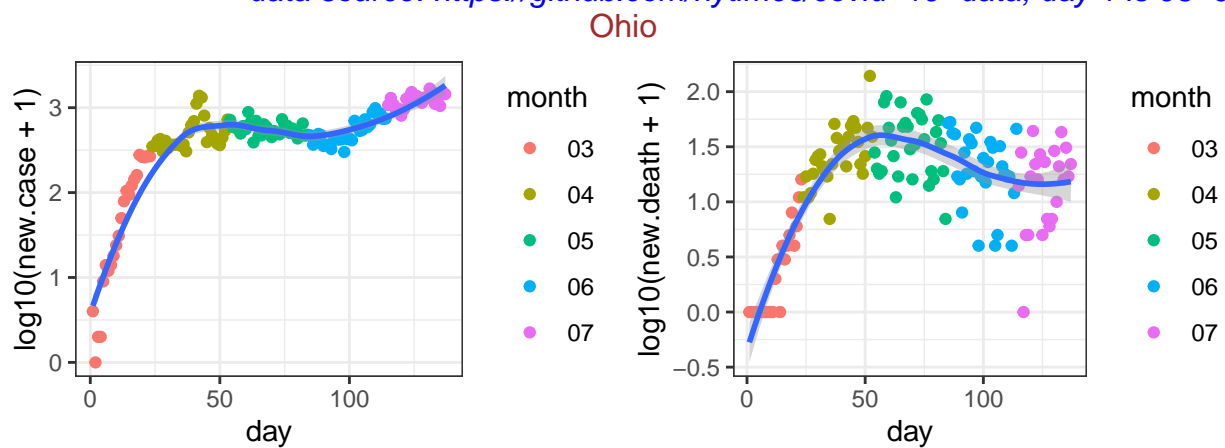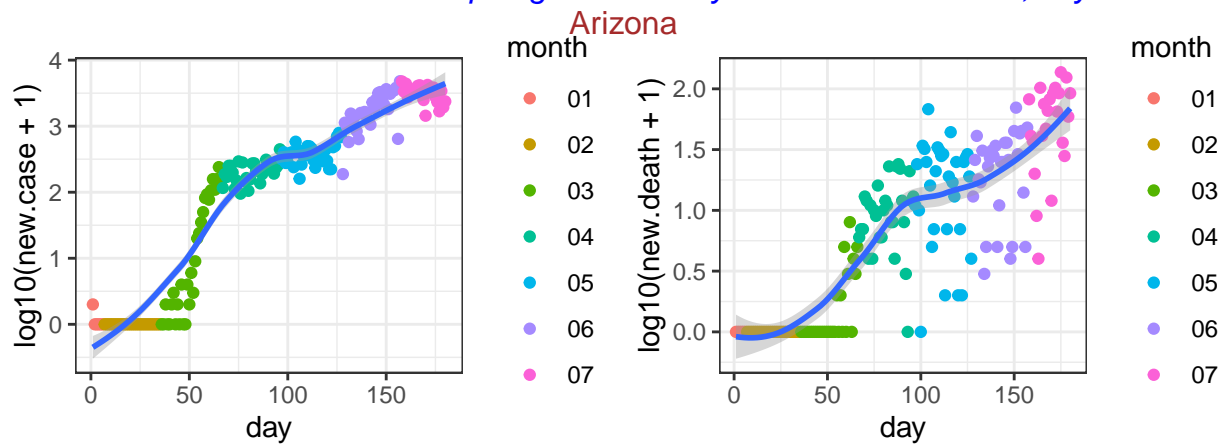
Georgia

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

Ohio

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

Arizona

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

Indiana

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

Virginia

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

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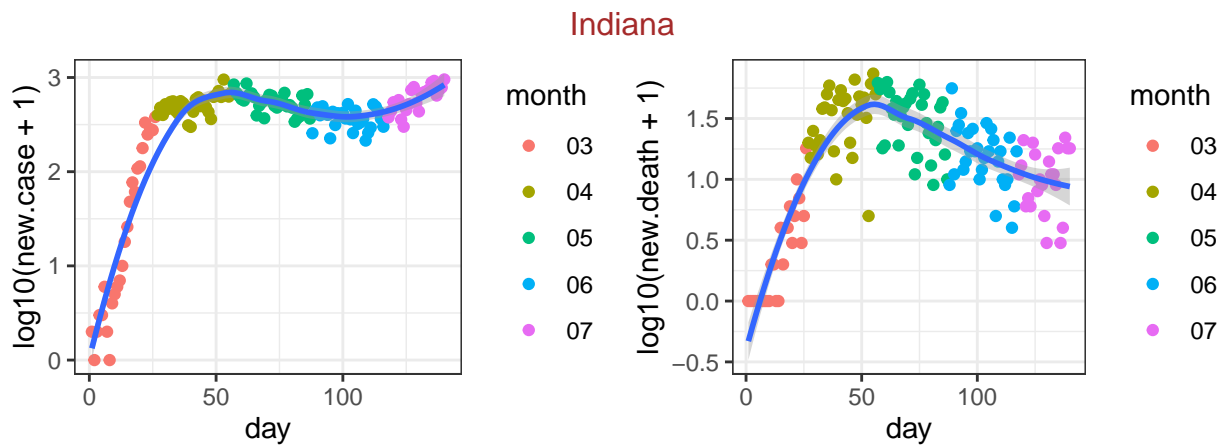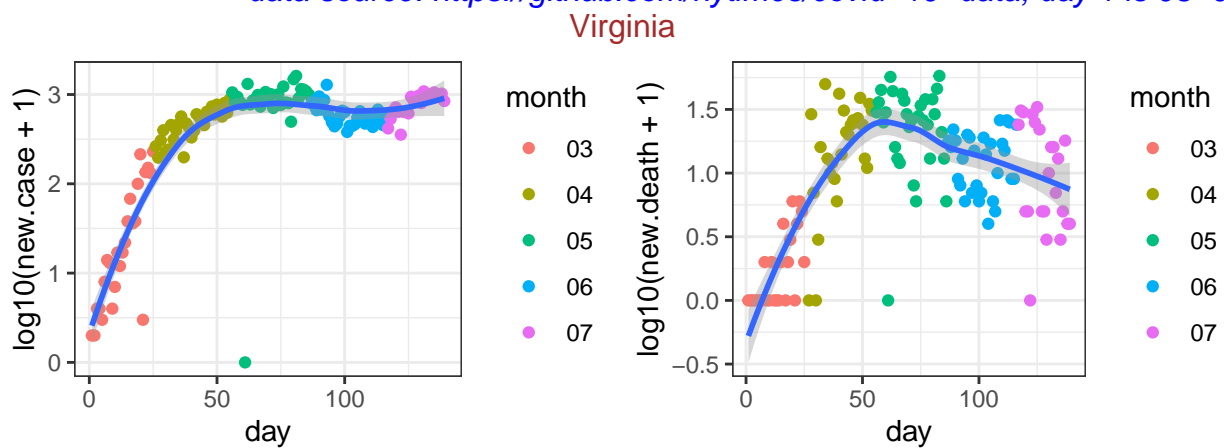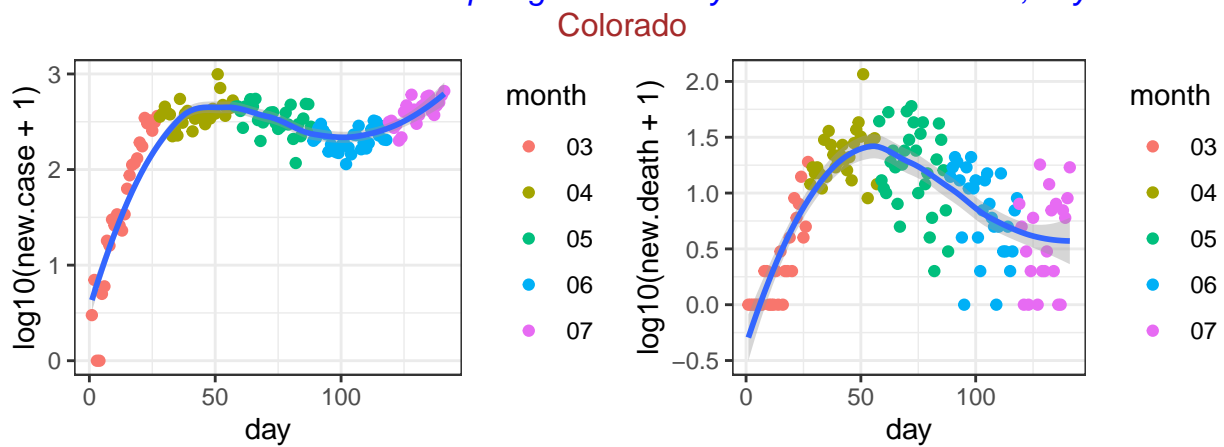
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

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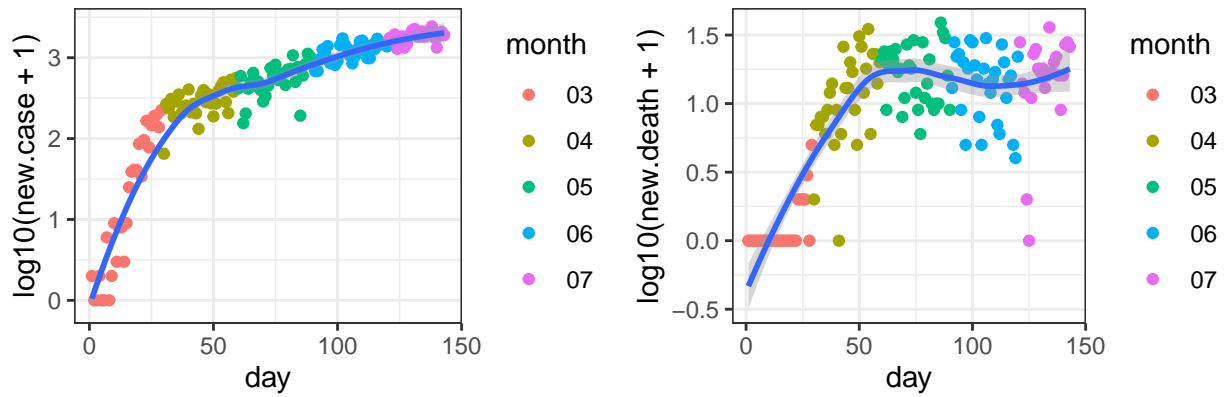
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

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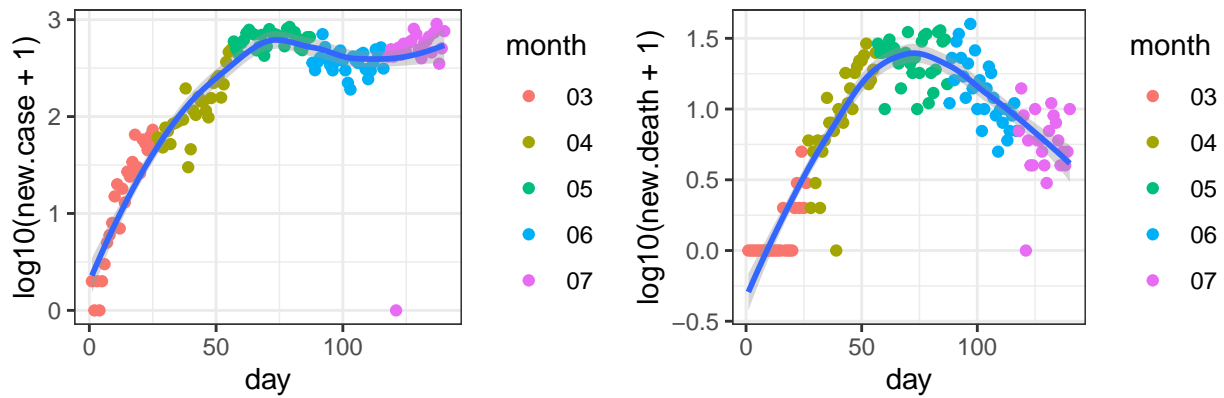
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington

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

14
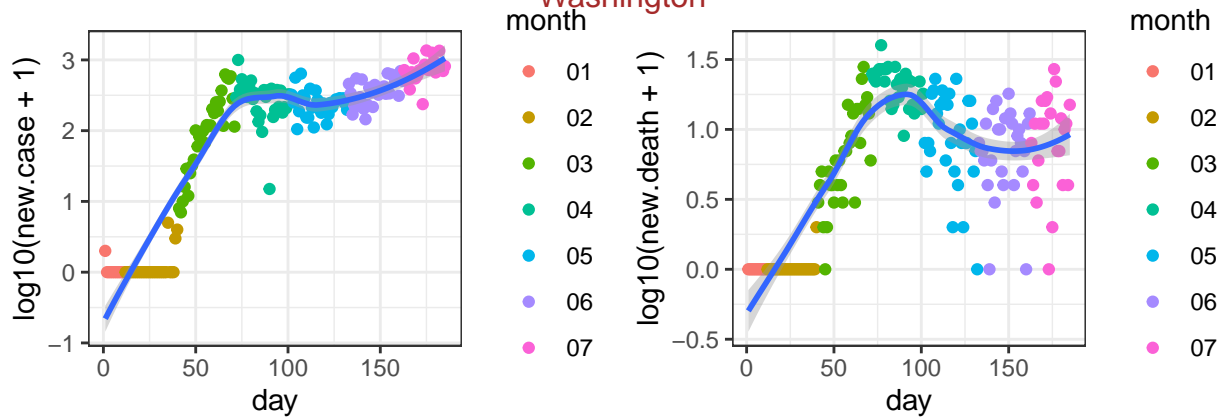
Mississippi
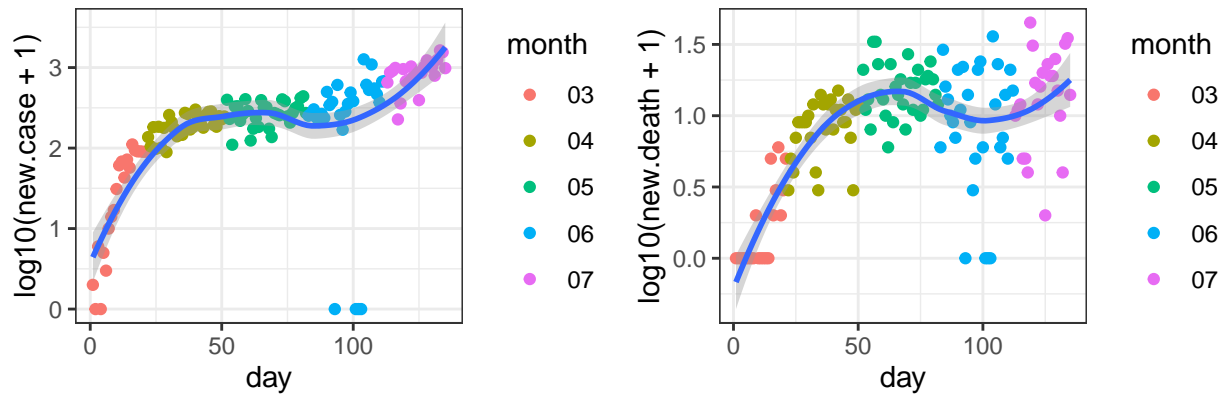
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11

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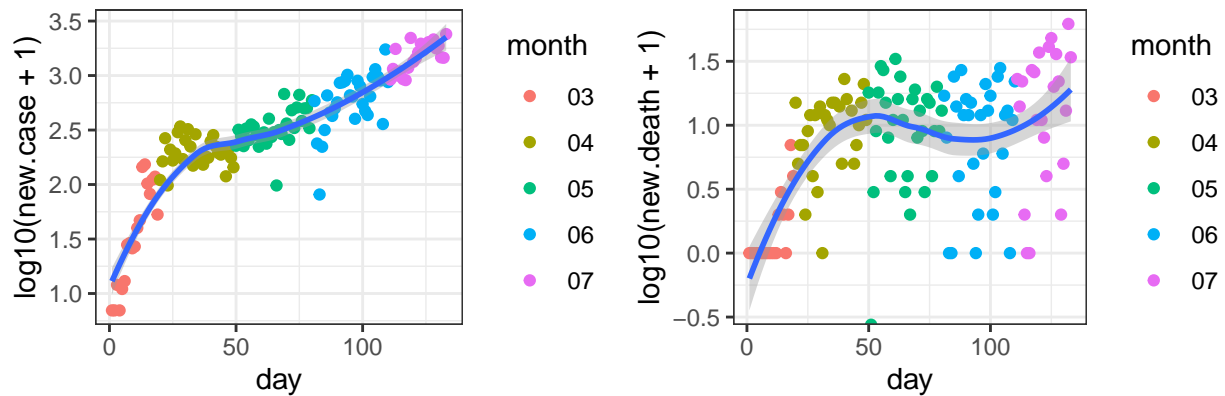
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13

South Carolina
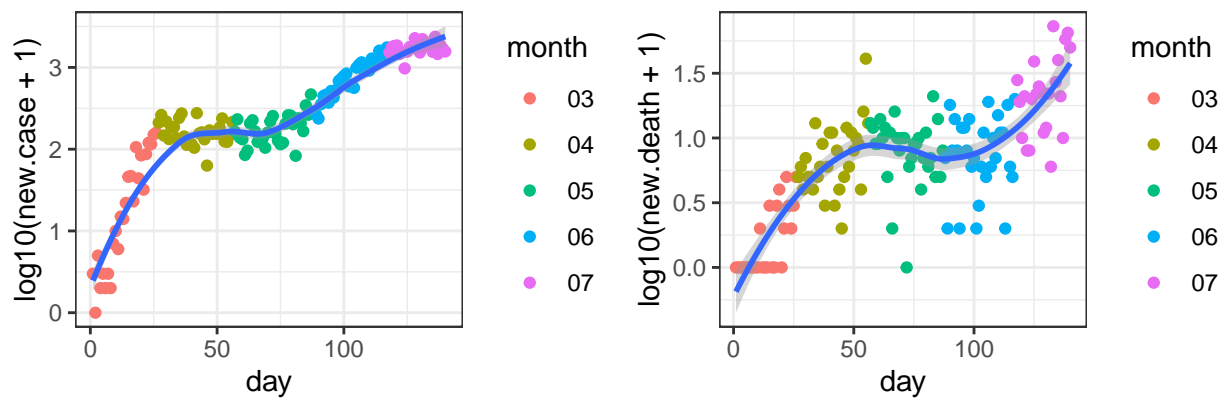
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06
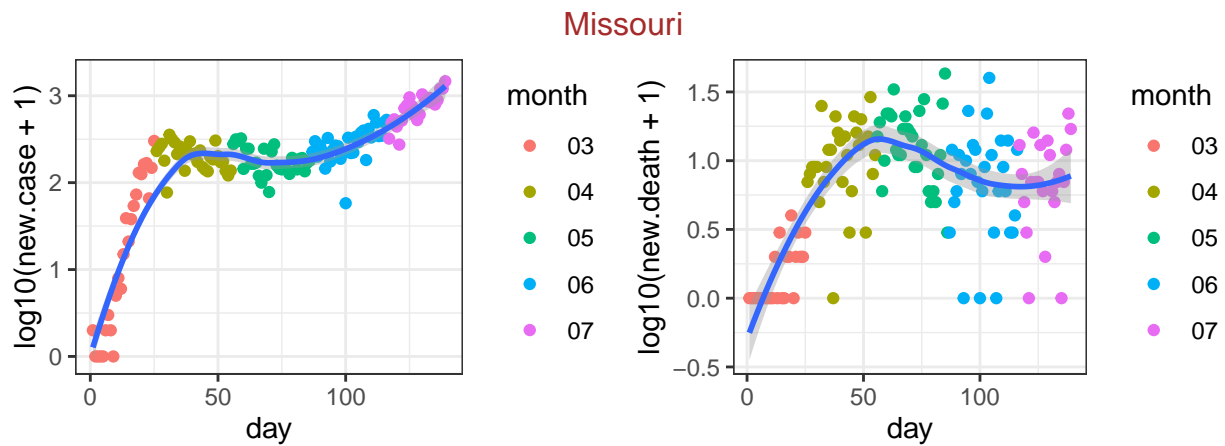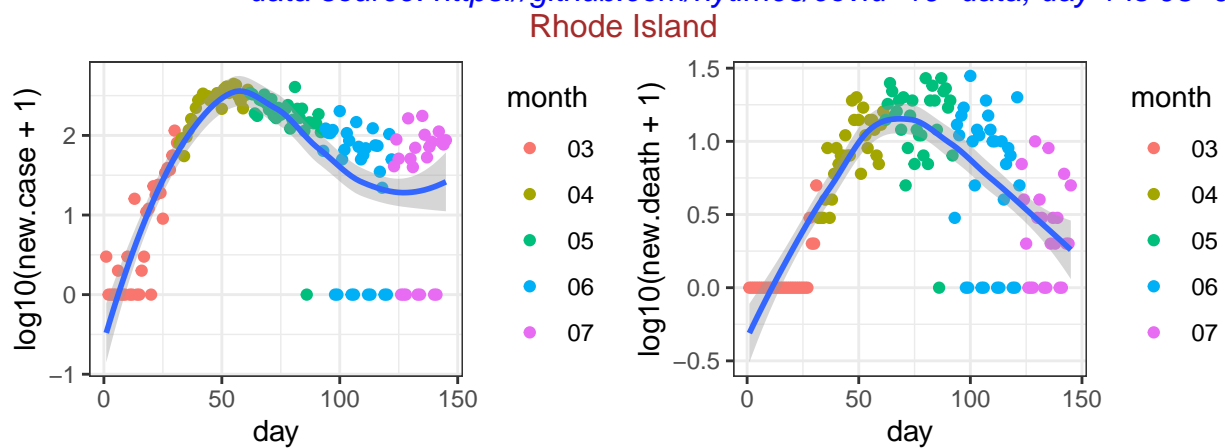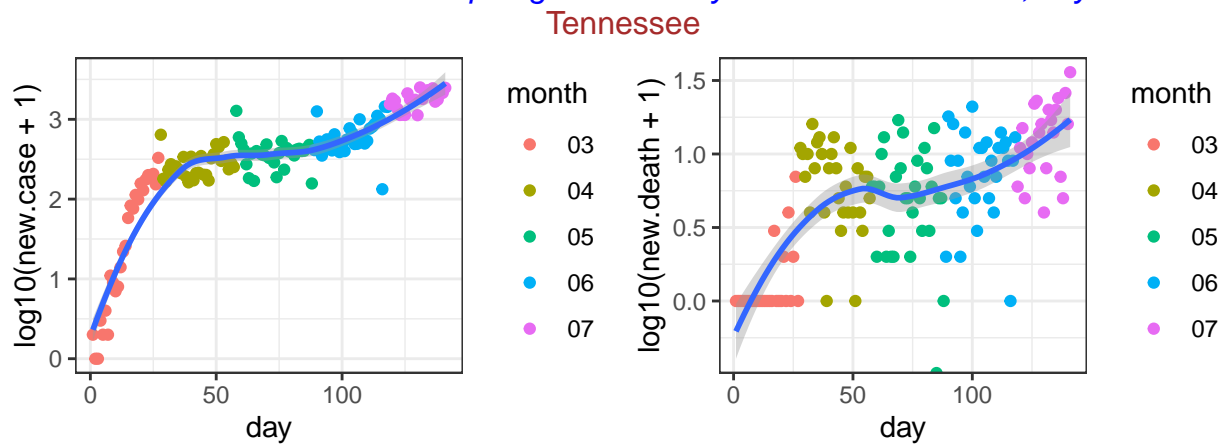
Missouri

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

Rhode Island

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

Tennessee

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

Wisconsin

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

Iowa

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

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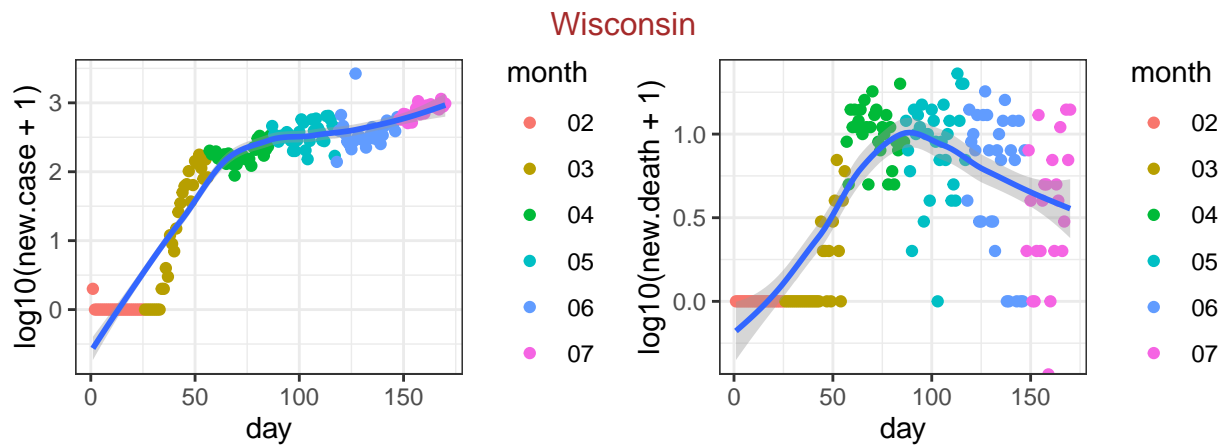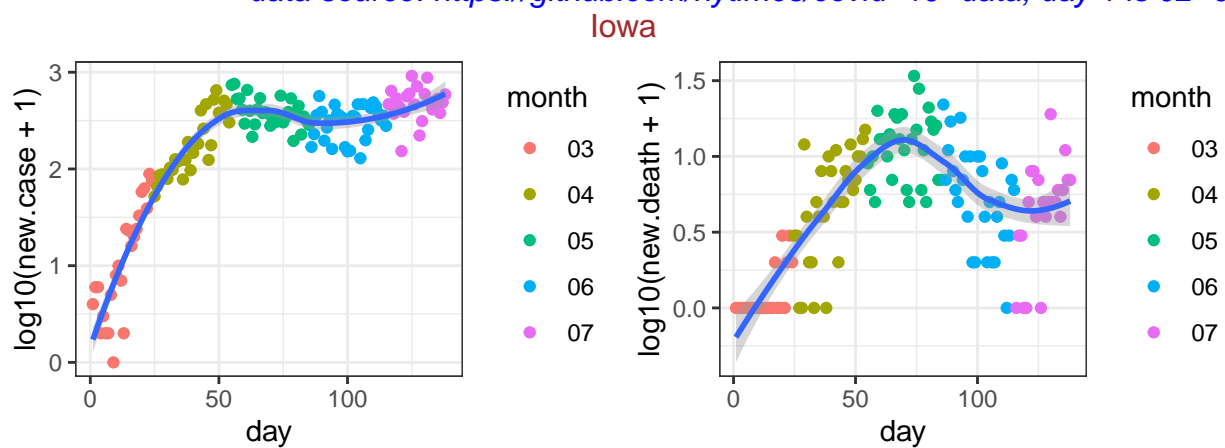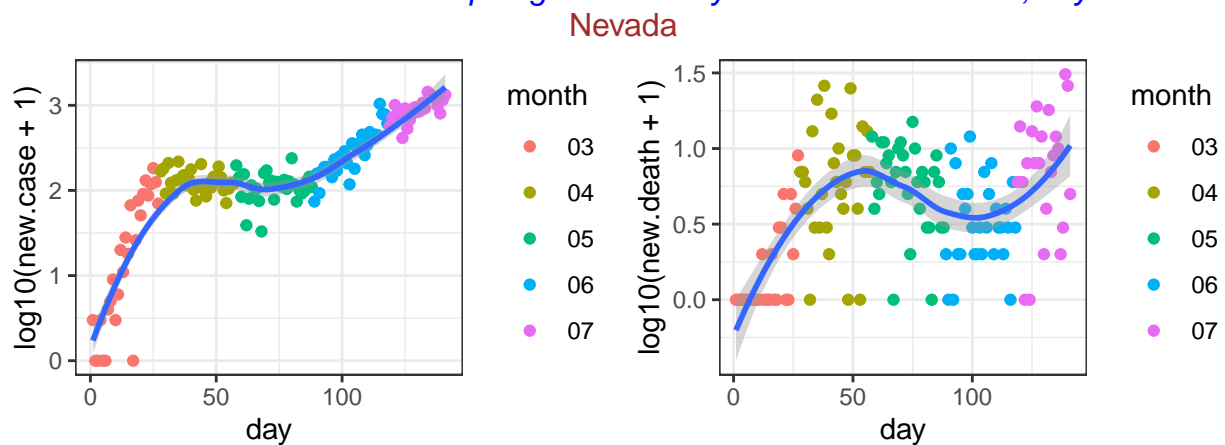
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March
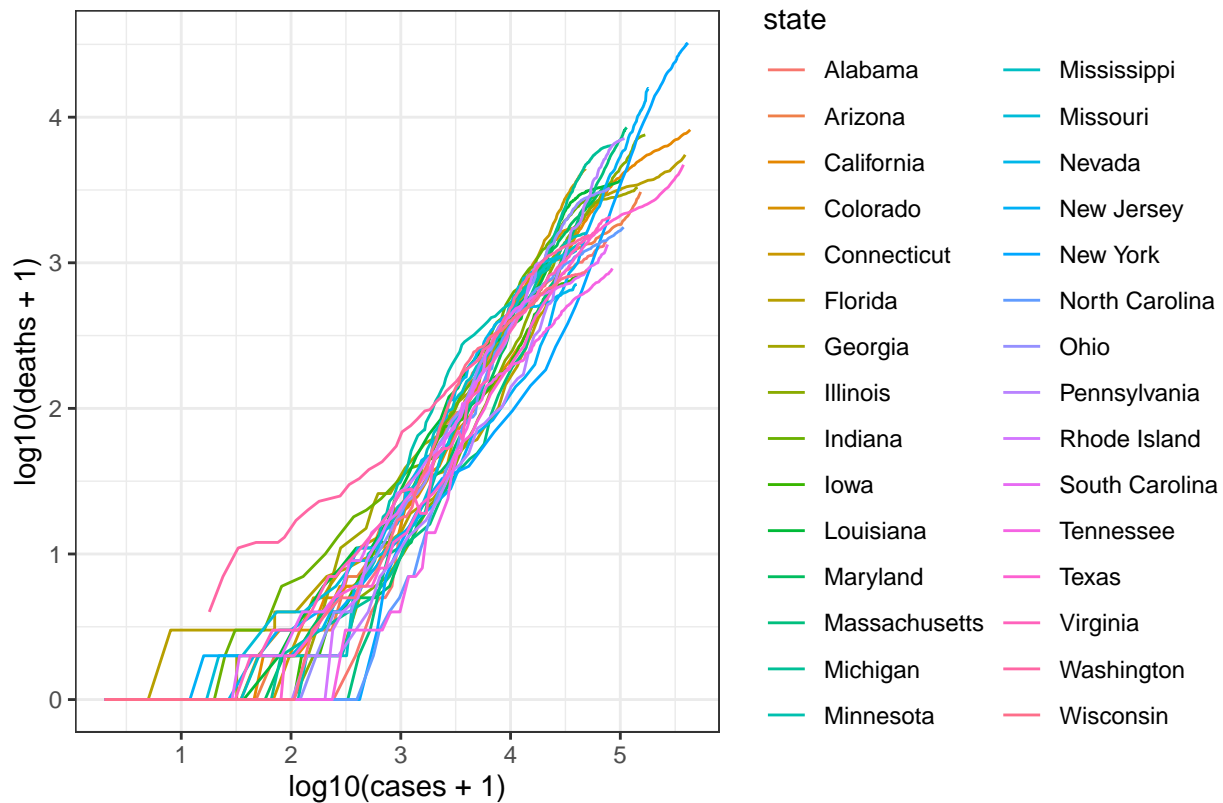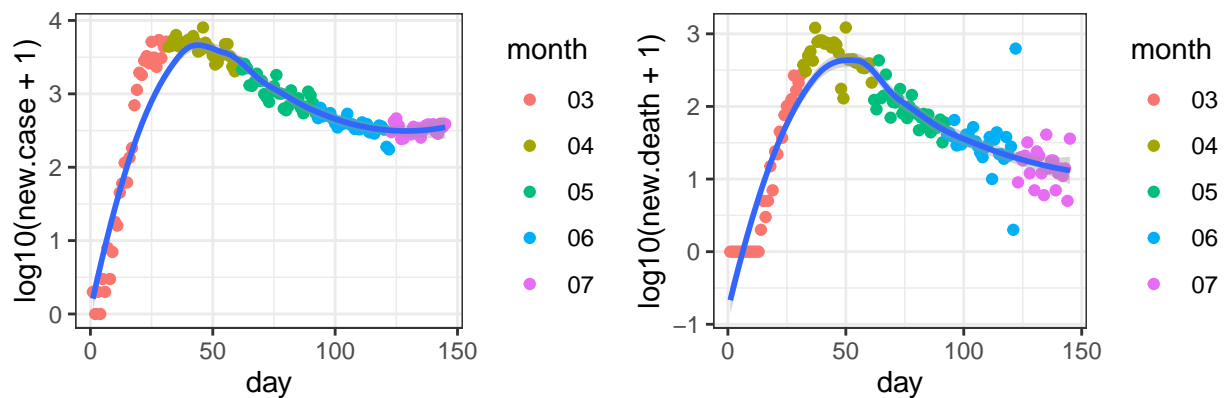
data source: https://github.com/nytimes/covid−19−data

county level data

First check the 50 counties with the largest number of deaths.

```
##              date               county                 state   fips   cases deaths
## 362130 2020-07-23      New York City           New York     NA 227517  22934
## 360893 2020-07-23               Cook           Illinois  17031 100795   4810
## 360488 2020-07-23        Los Angeles         California   6037 166848   4262
## 361600 2020-07-23              Wayne           Michigan  26163  25673   2790
## 362129 2020-07-23             Nassau           New York  36059  42816   2705
## 362054 2020-07-23              Essex         New Jersey  34013  19445   2094
## 362149 2020-07-23            Suffolk           New York  36103  42715   2042
## 362049 2020-07-23             Bergen         New Jersey  34003  20396   2034
## 361511 2020-07-23          Middlesex      Massachusetts  25017  25123   1951
## 362565 2020-07-23       Philadelphia       Pennsylvania  42101  29102   1675
## 360386 2020-07-23            Maricopa            Arizona   4013 102247   1651
## 362157 2020-07-23         Westchester           New York  36119  35666   1576
## 362056 2020-07-23             Hudson         New Jersey  34017  19464   1494
## 360591 2020-07-23           Hartford        Connecticut   9003  12145   1408
## 360590 2020-07-23          Fairfield        Connecticut   9001  17184   1401
## 362059 2020-07-23          Middlesex         New Jersey  34023  17654   1396
## 360646 2020-07-23          Miami-Dade            Florida  12086  95067   1354
## 362067 2020-07-23              Union         New Jersey  34039  16472   1343
## 362063 2020-07-23            Passaic         New Jersey  34031  17370   1238
## 361507 2020-07-23              Essex      Massachusetts  25009  16856   1157
## 361580 2020-07-23            Oakland           Michigan  26125  13711   1122
## 360594 2020-07-23          New Haven        Connecticut   9009  12835   1093
```
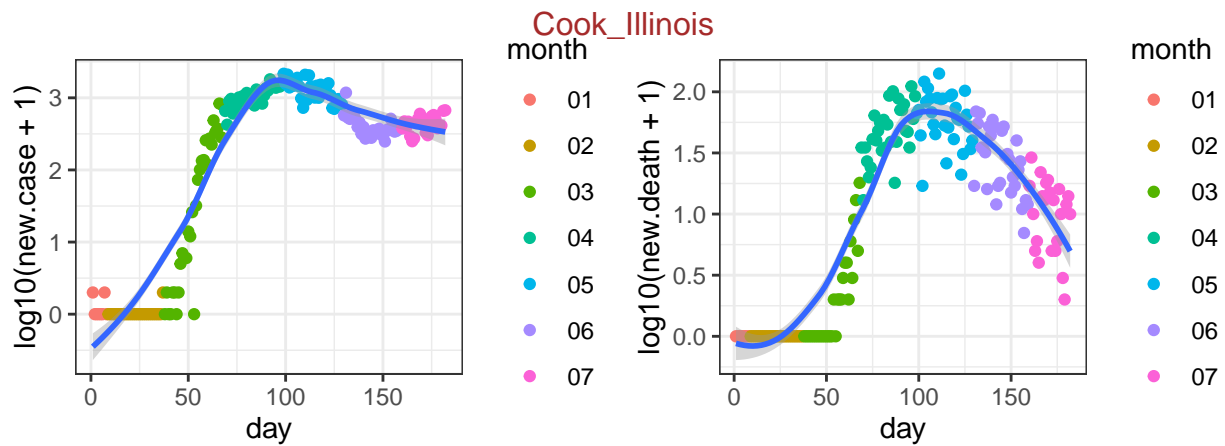
```
## 361515 2020-07-23             Suffolk      Massachusetts 25025  20732  1042
## 362062 2020-07-23               Ocean        New Jersey 34029  10144  1012
## 361513 2020-07-23             Norfolk      Massachusetts 25021   9849   974
## 361517 2020-07-23           Worcester      Massachusetts 25027  12972   974
## 361567 2020-07-23              Macomb          Michigan 26099   8889   941
## 362060 2020-07-23            Monmouth        New Jersey 34025   9848   851
## 362560 2020-07-23          Montgomery      Pennsylvania 42091   9311   842
## 362061 2020-07-23              Morris        New Jersey 34027   7169   827
## 361628 2020-07-23            Hennepin         Minnesota 27053  15537   804
## 362664 2020-07-23          Providence      Rhode Island 44007  14033   801
## 361493 2020-07-23          Montgomery          Maryland 24031  16692   781
## 361029 2020-07-23              Marion           Indiana 18097  13380   758
## 360653 2020-07-23          Palm Beach           Florida 12099  28265   727
## 361494 2020-07-23      Prince George's          Maryland 24033  21493   727
## 362537 2020-07-23            Delaware      Pennsylvania 42045   8097   720
## 361514 2020-07-23            Plymouth      Massachusetts 25023   8940   701
## 361509 2020-07-23             Hampden      Massachusetts 25013   7220   681
## 363320 2020-07-23                King        Washington 53033  13837   659
## 360501 2020-07-23           Riverside        California  6065  33093   637
## 361873 2020-07-23           St. Louis          Missouri 29189   9944   626
## 361505 2020-07-23             Bristol      Massachusetts 25005   8788   615
## 362115 2020-07-23                Erie          New York 36029   8187   614
## 362058 2020-07-23              Mercer        New Jersey 34021   7912   608
## 362973 2020-07-23              Harris             Texas 48201  61416   597
## 360603 2020-07-23 District of Columbia District of Columbia 11001  11571   581
## 362024 2020-07-23               Clark            Nevada 32003  34209   580
## 362929 2020-07-23              Dallas             Texas 48113  44087   579
## 362523 2020-07-23               Bucks      Pennsylvania 42017   6542   574
```

For these 50 counties, I check the number of new cases and the number of new deaths.
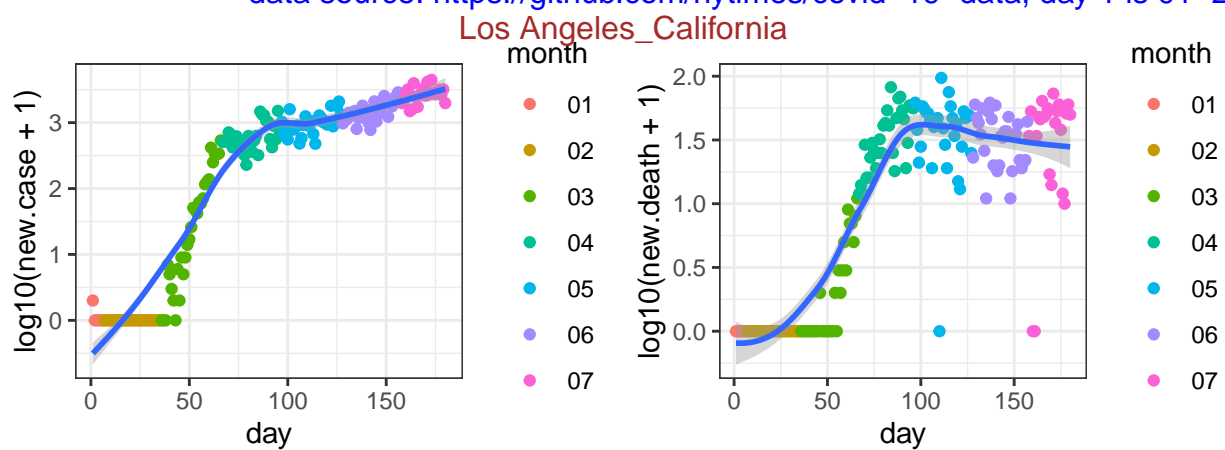

New York City_New York
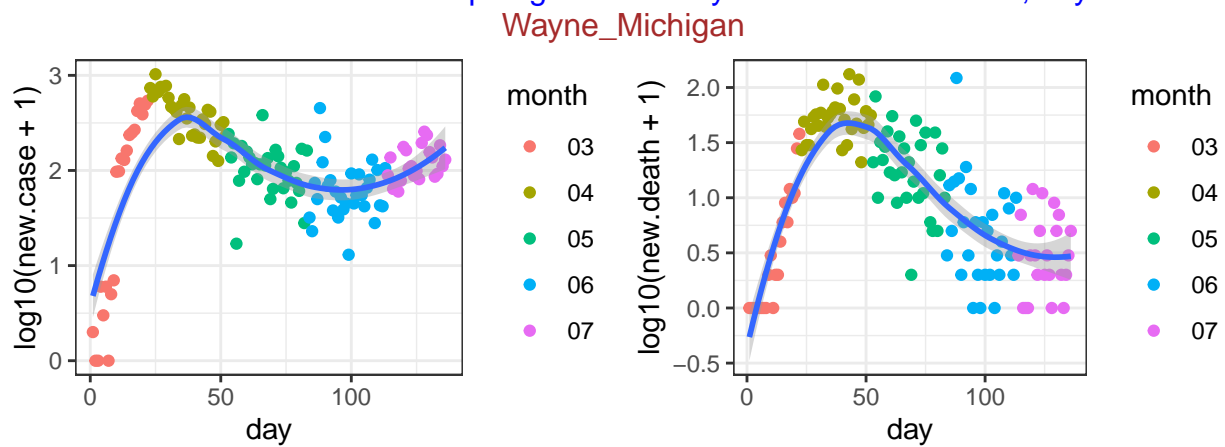
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

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

Los Angeles_California

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

Wayne_Michigan

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

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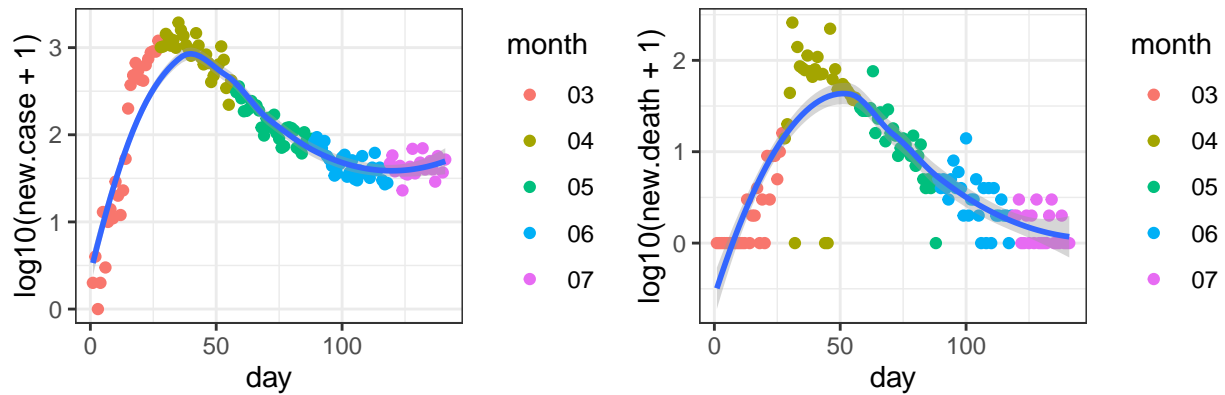
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

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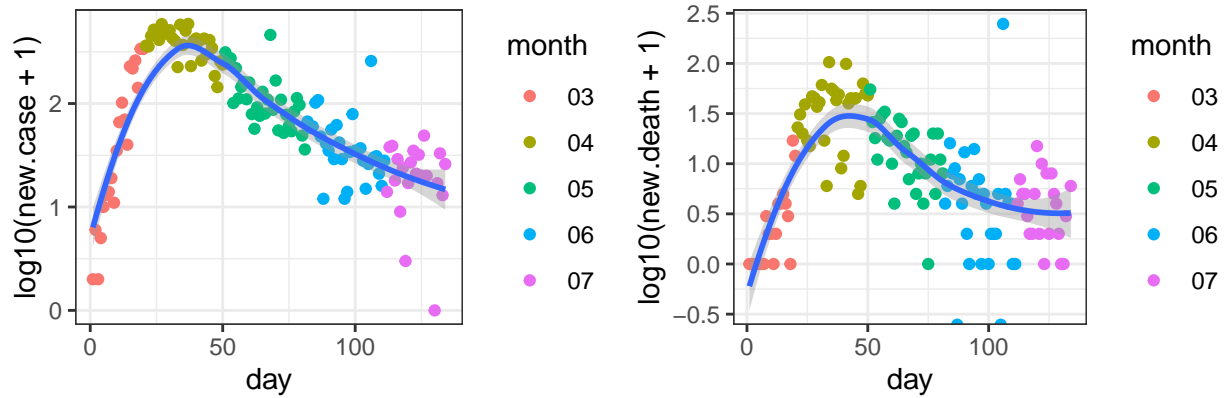
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12

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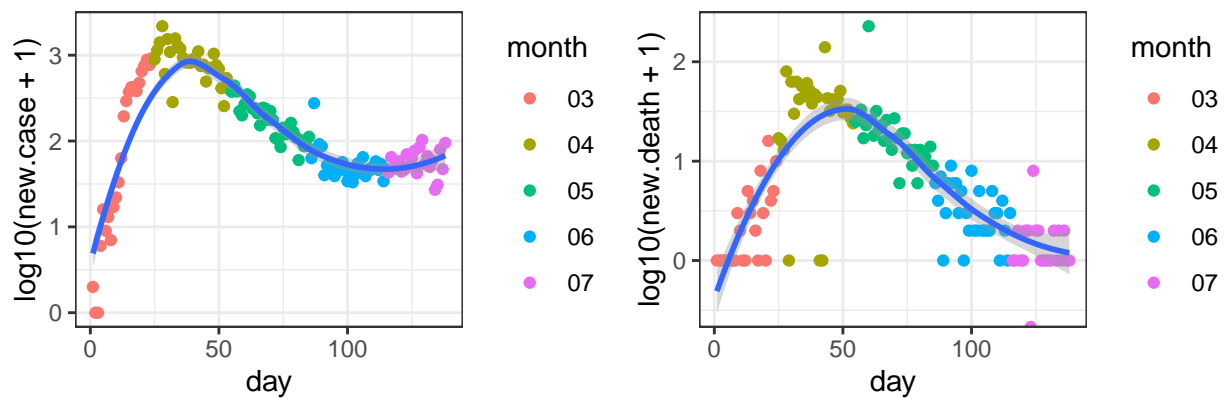
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08

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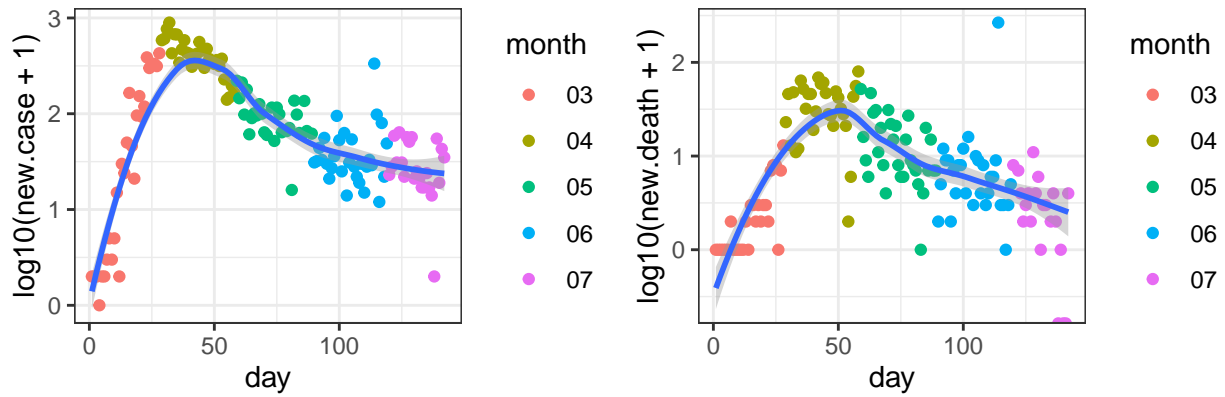
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04

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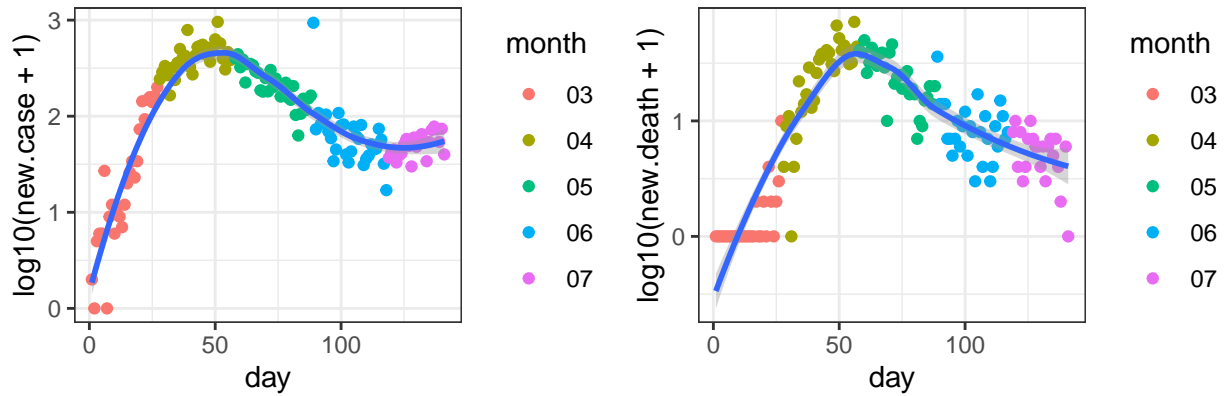
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

Philadelphia_Pennsylvania

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

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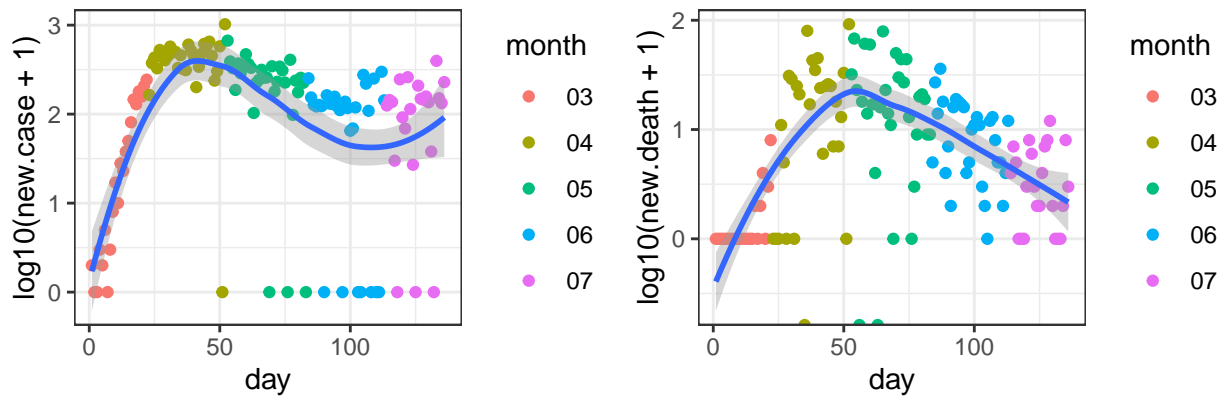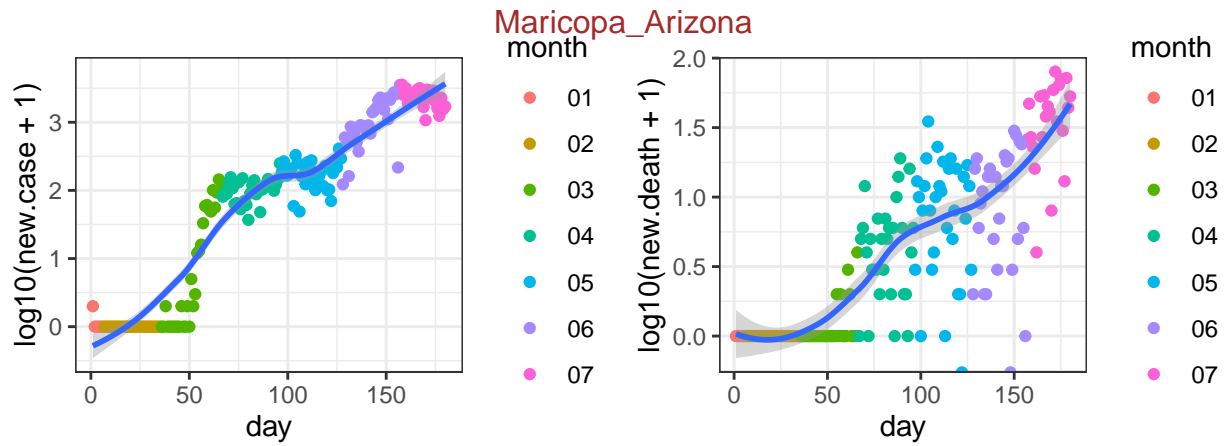
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-26

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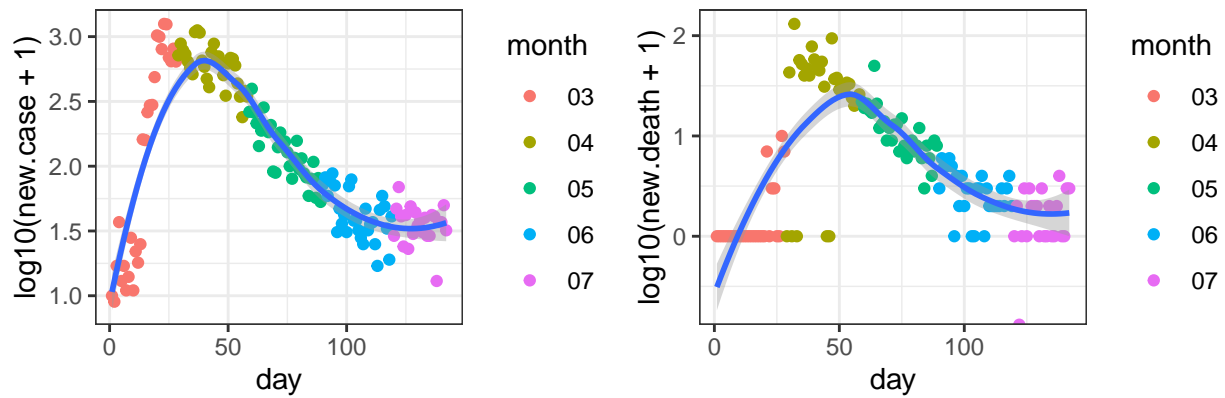
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04

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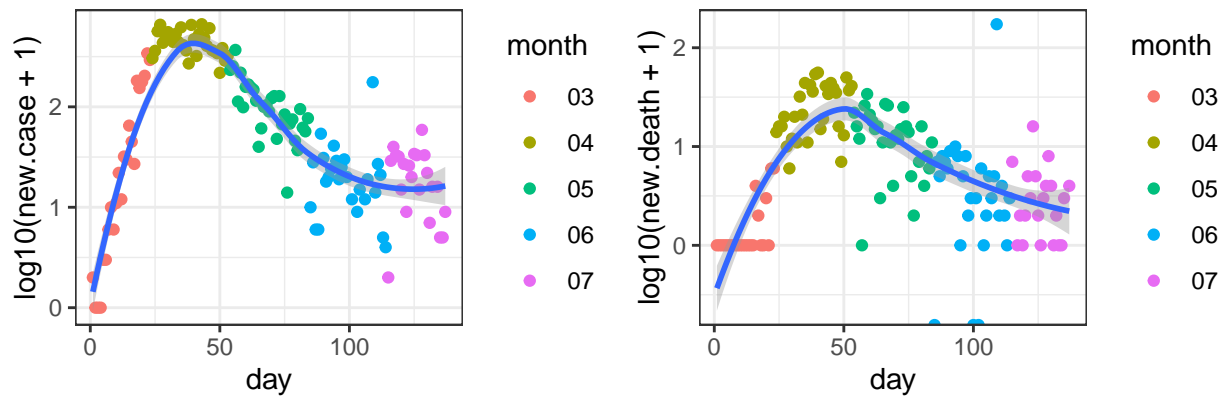
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09

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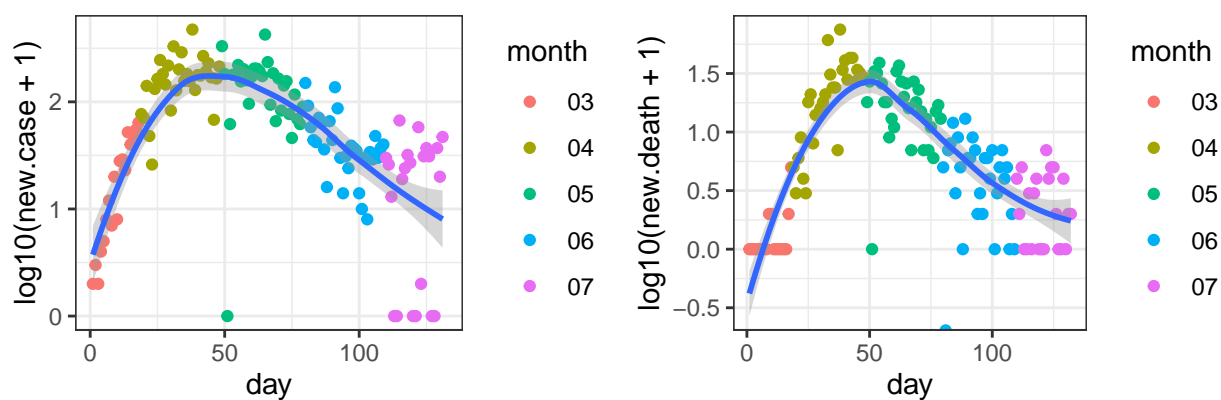
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

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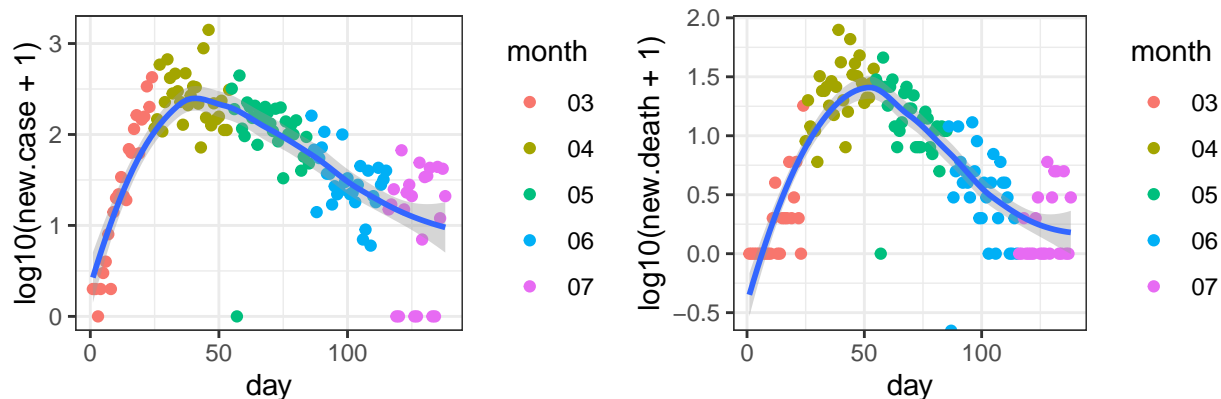
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

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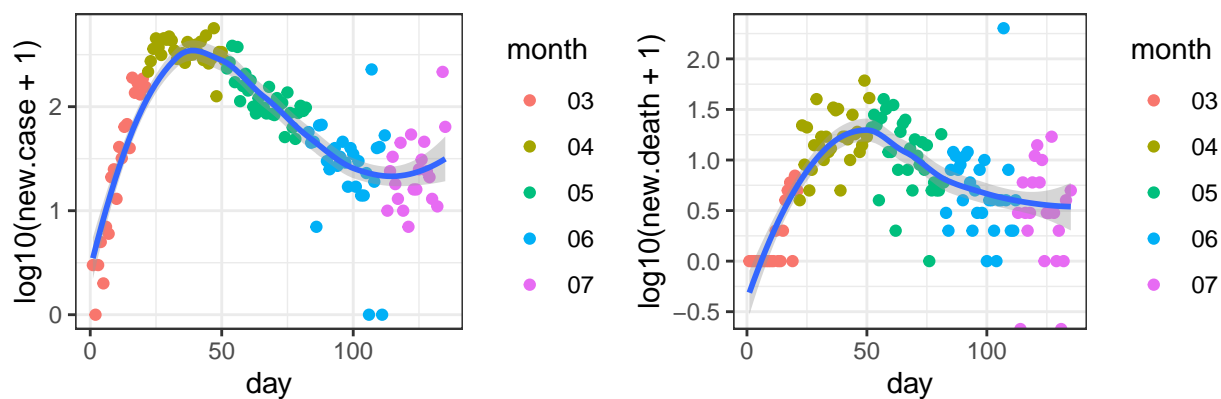
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Miami–Dade_Florida
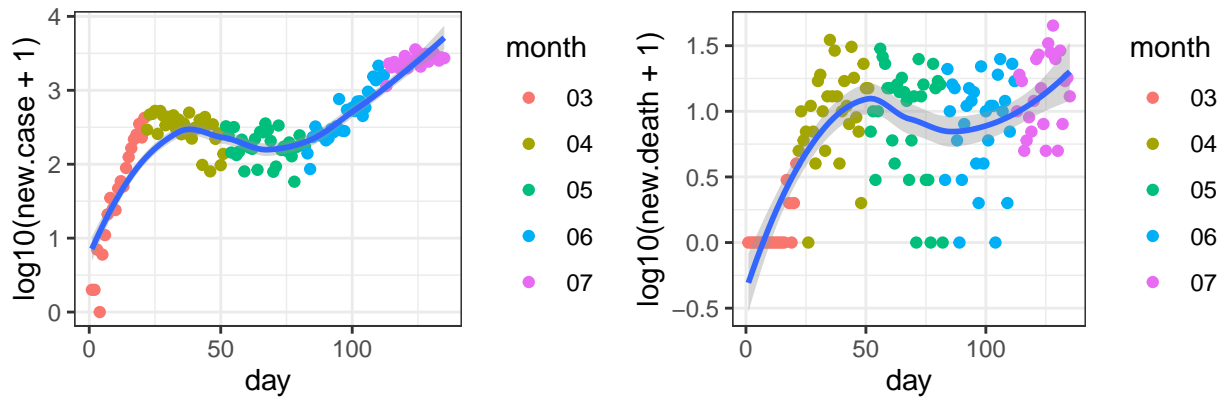
data source: https://github.com/nytimes/covid-19-data, day 1 is 03–11

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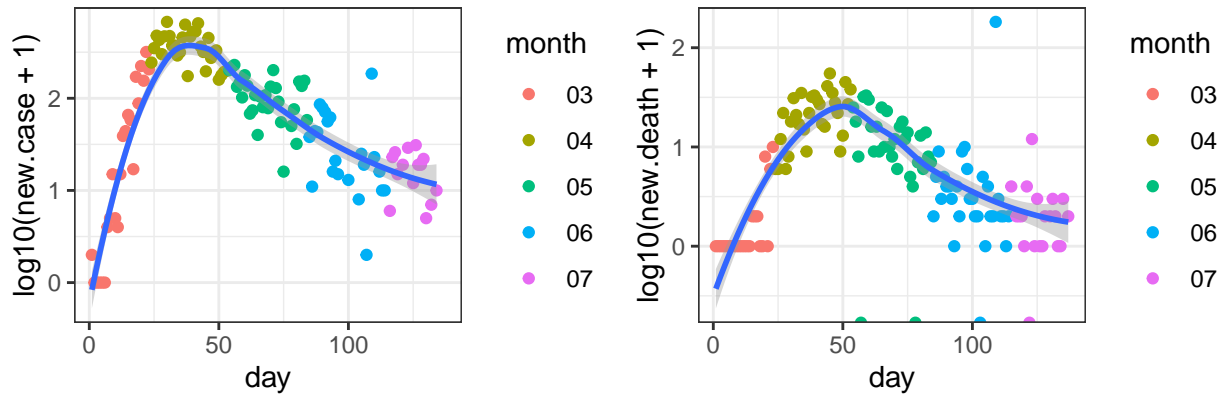
data source: https://github.com/nytimes/covid-19-data, day 1 is 03–09

Passaic_New Jersey
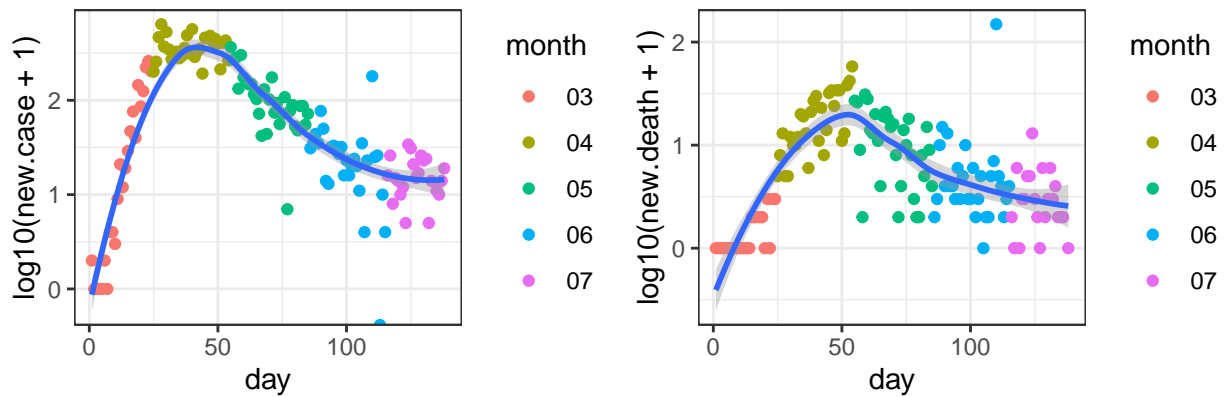
data source: https://github.com/nytimes/covid-19-data, day 1 is 03–08

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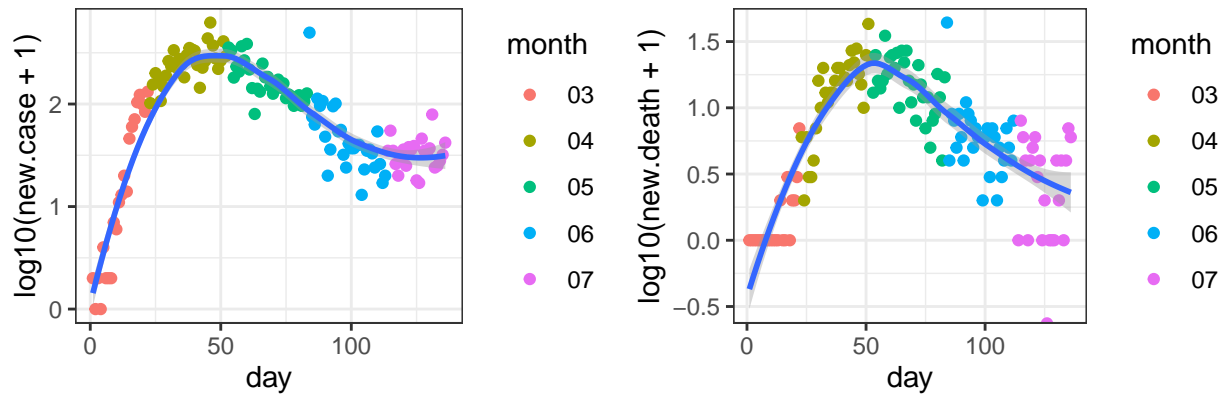
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

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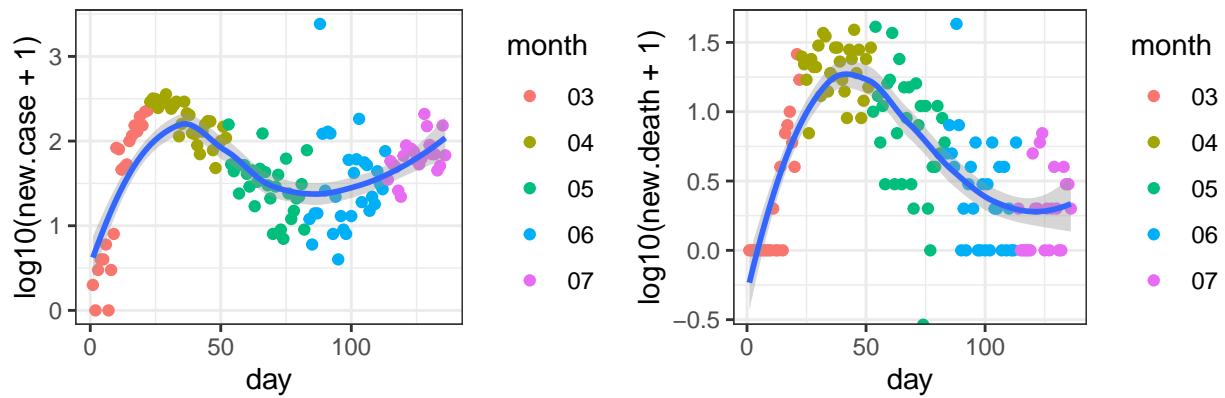
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

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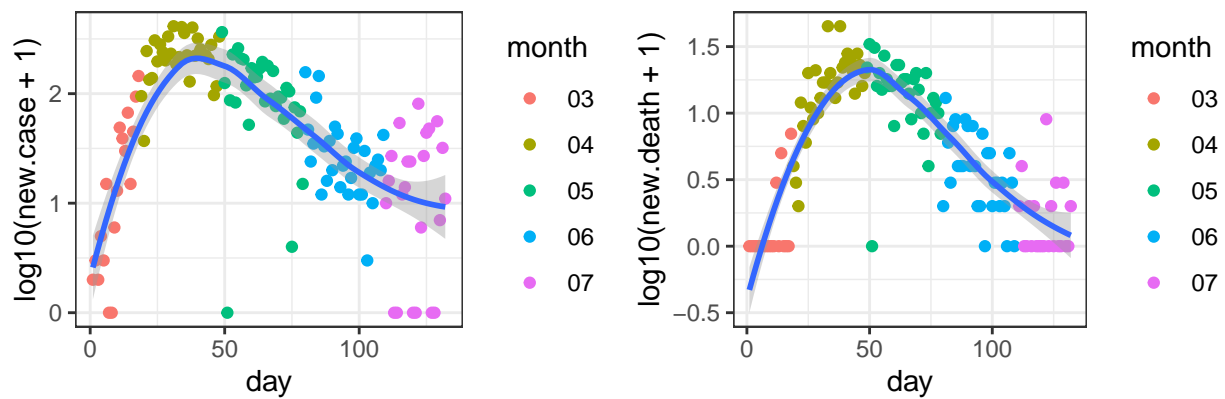
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14

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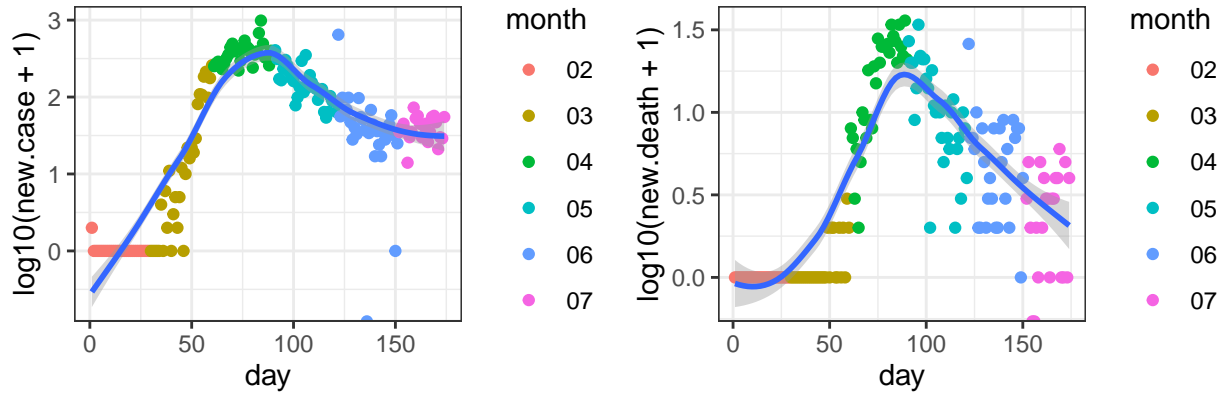
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01

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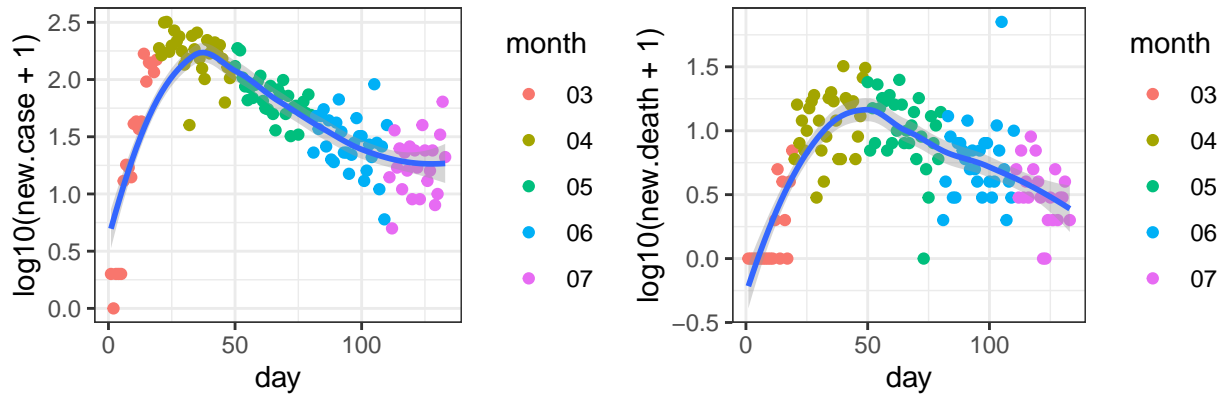
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13

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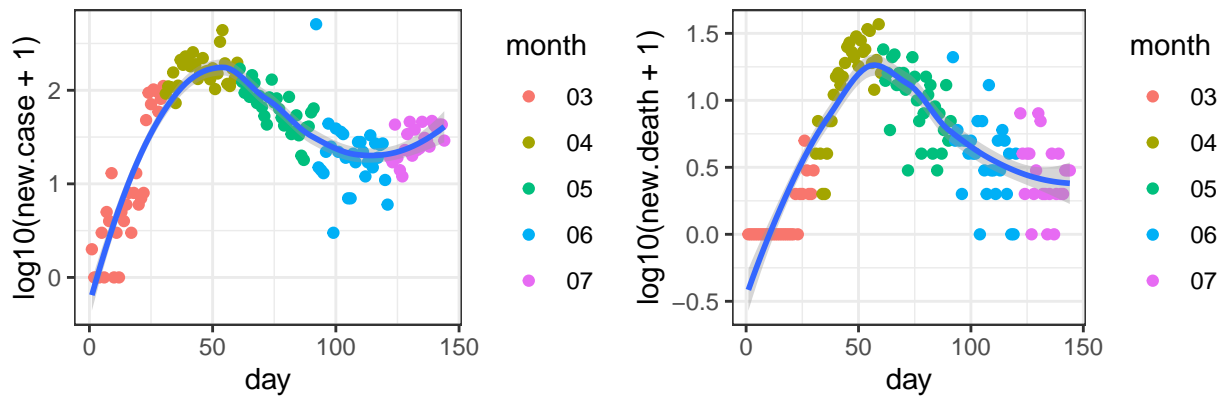
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02

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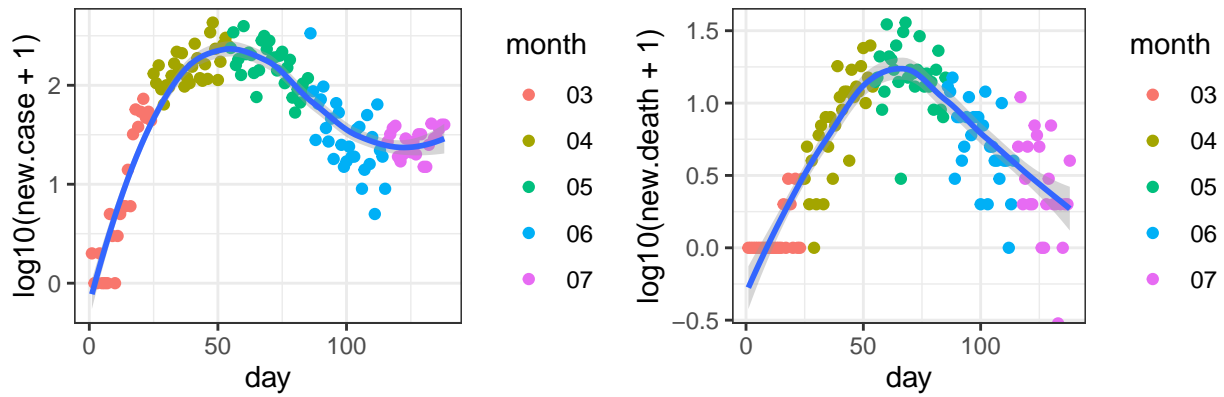
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

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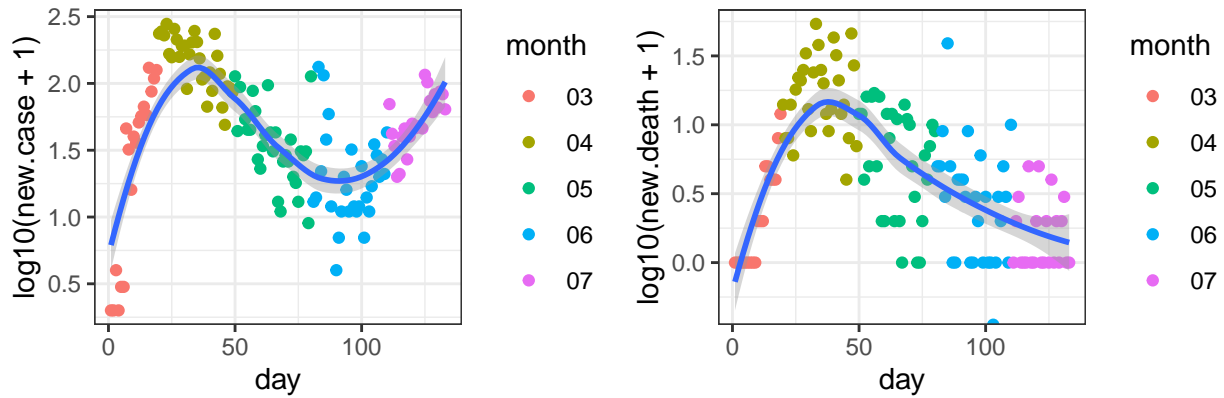
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

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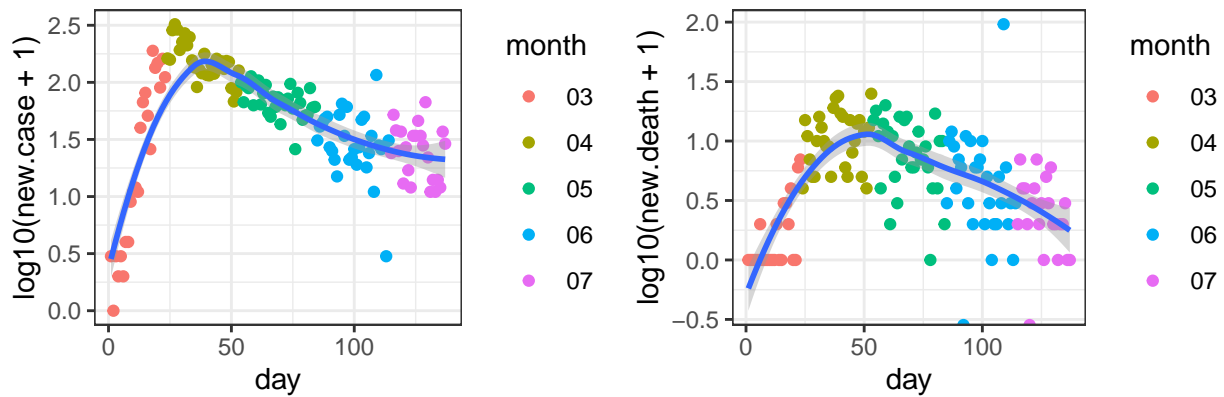
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

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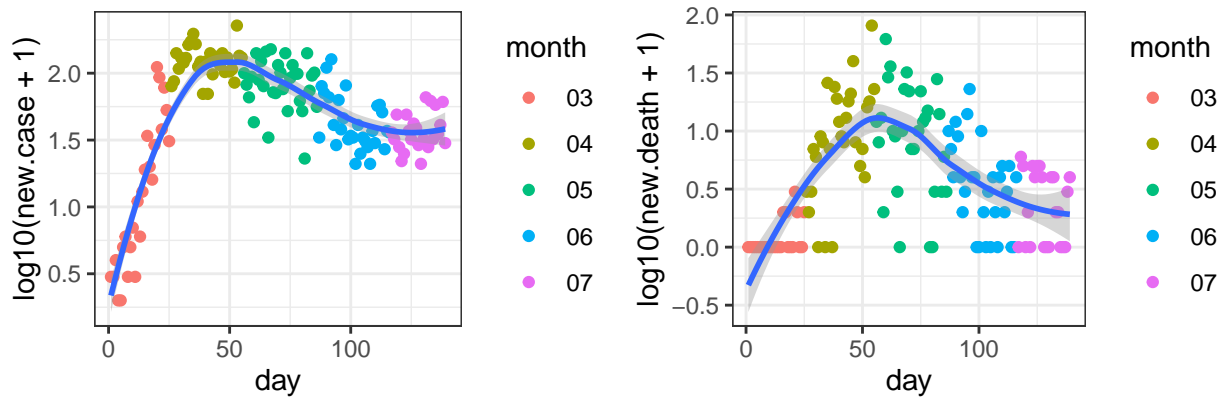
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07

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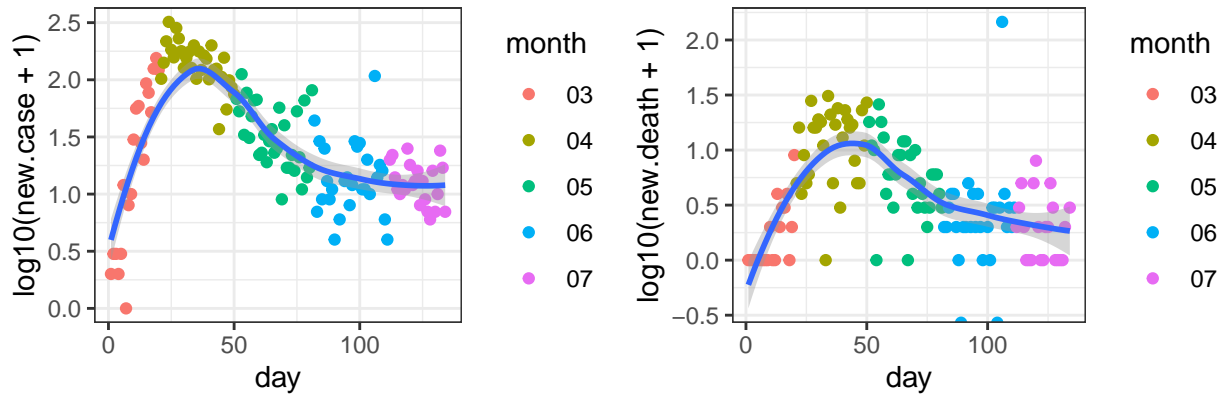
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12

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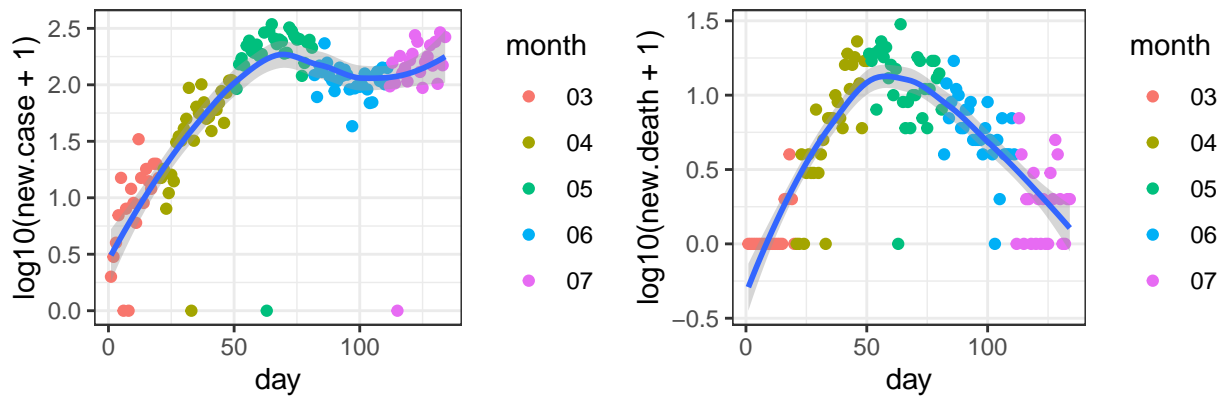
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12

Providence_Rhode Island
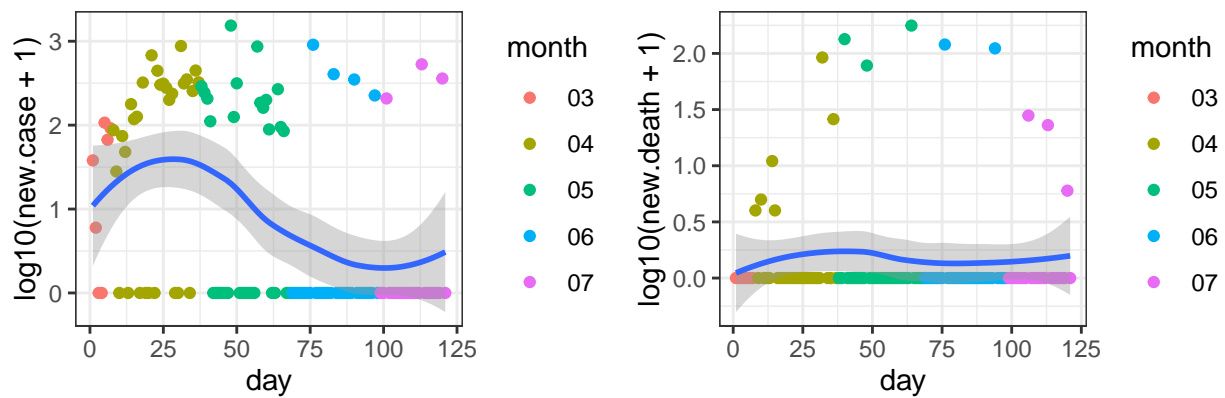
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-25

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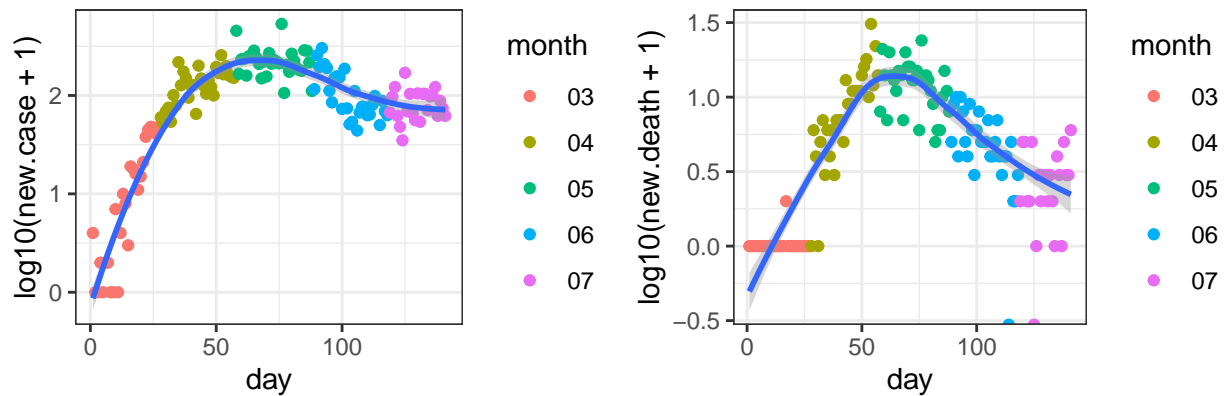
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

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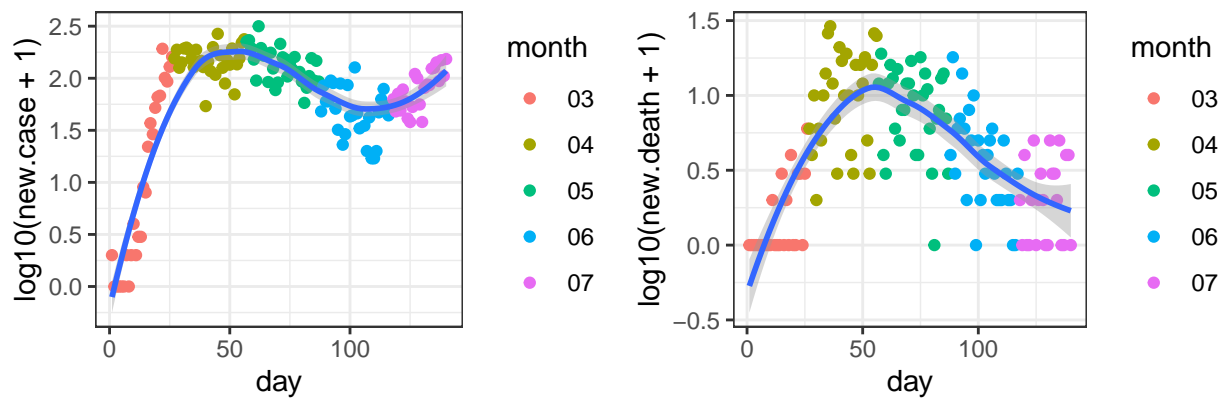
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06

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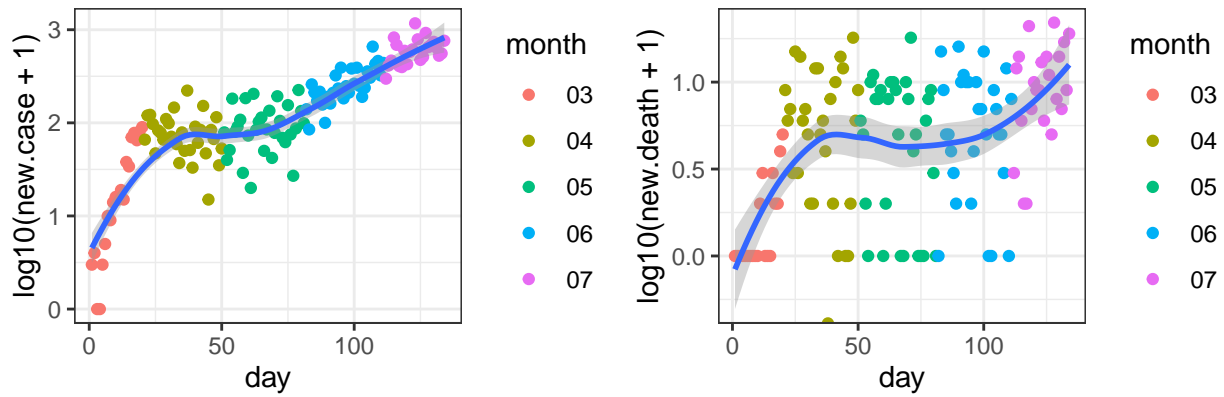
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12

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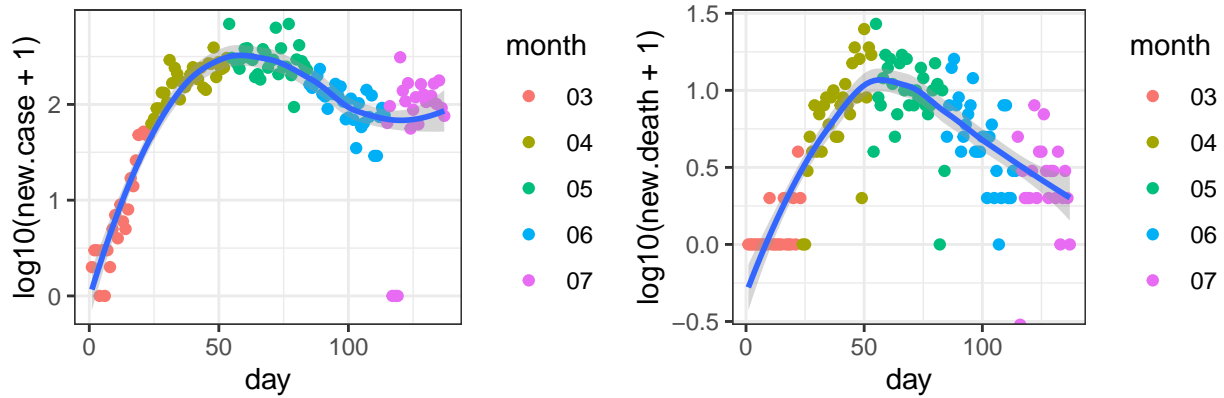
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09

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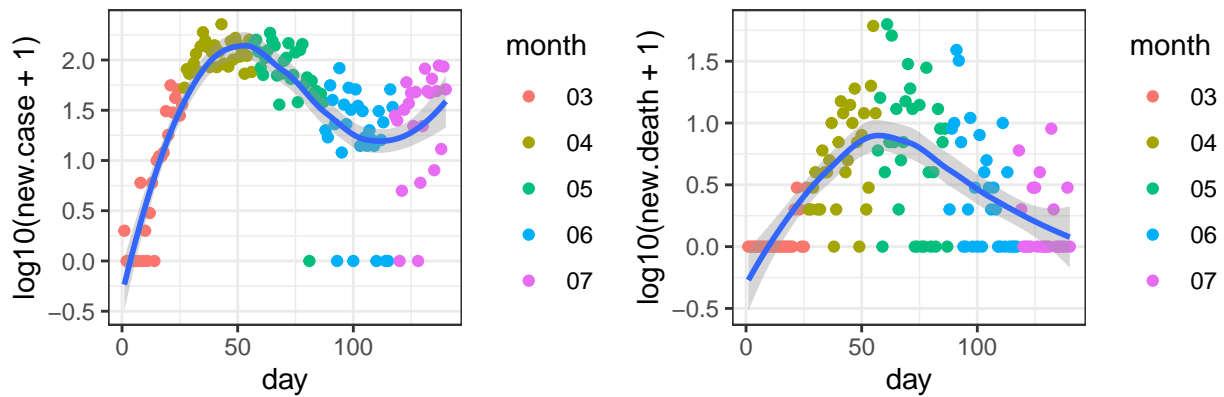
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06

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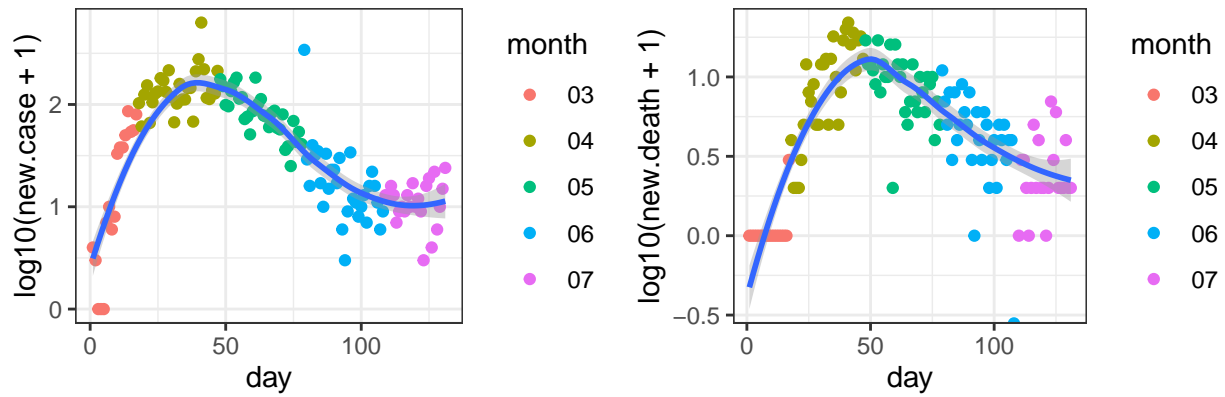
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

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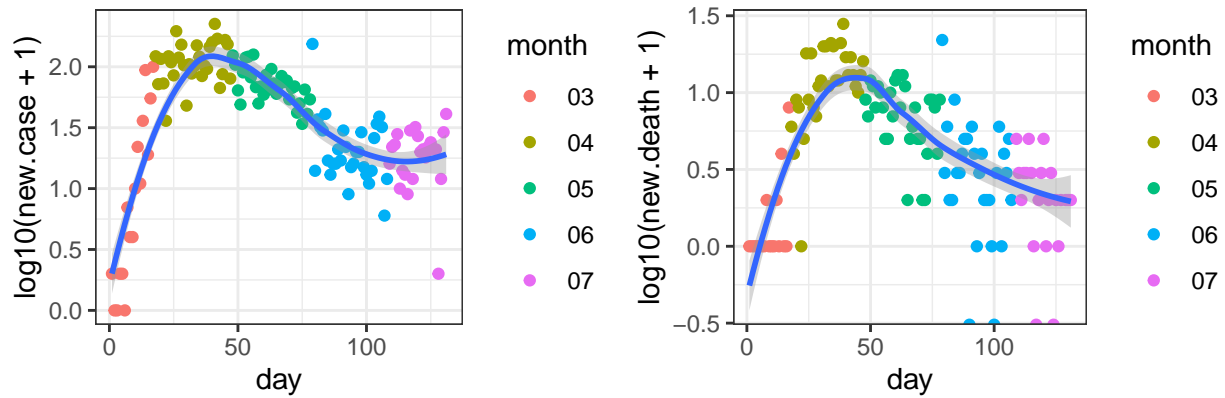
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

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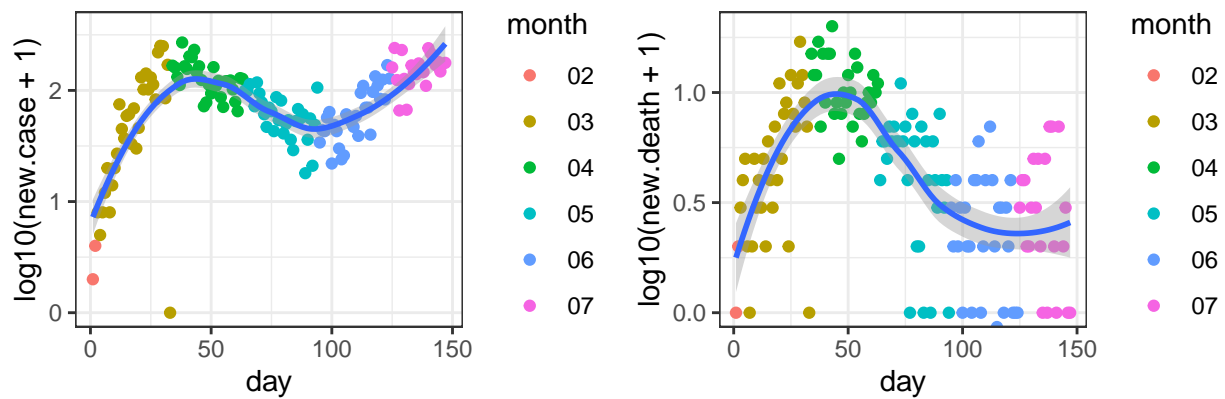
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

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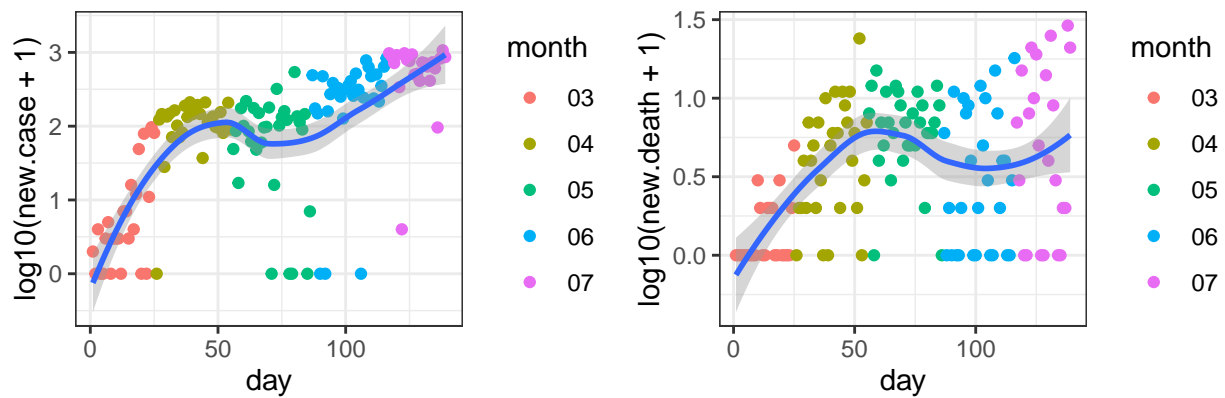
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07

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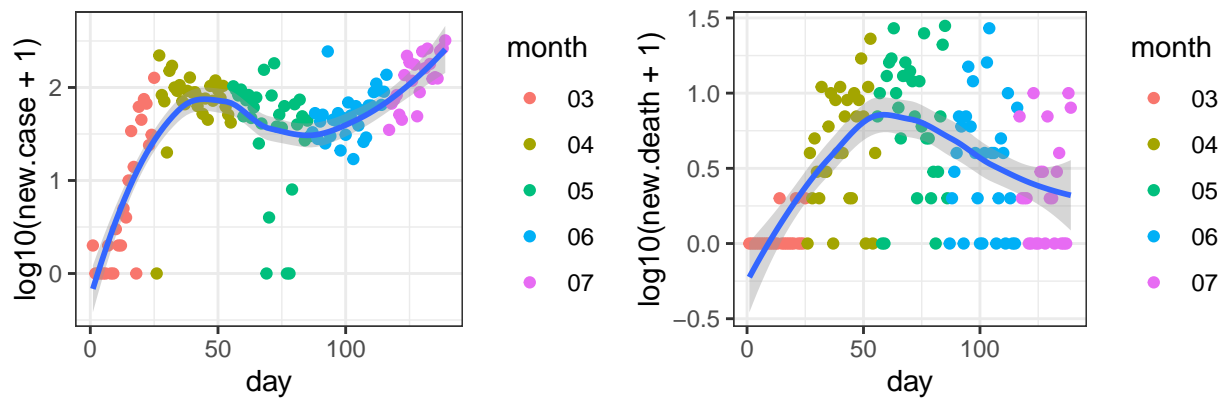
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07

Bristol_Massachusetts

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

33
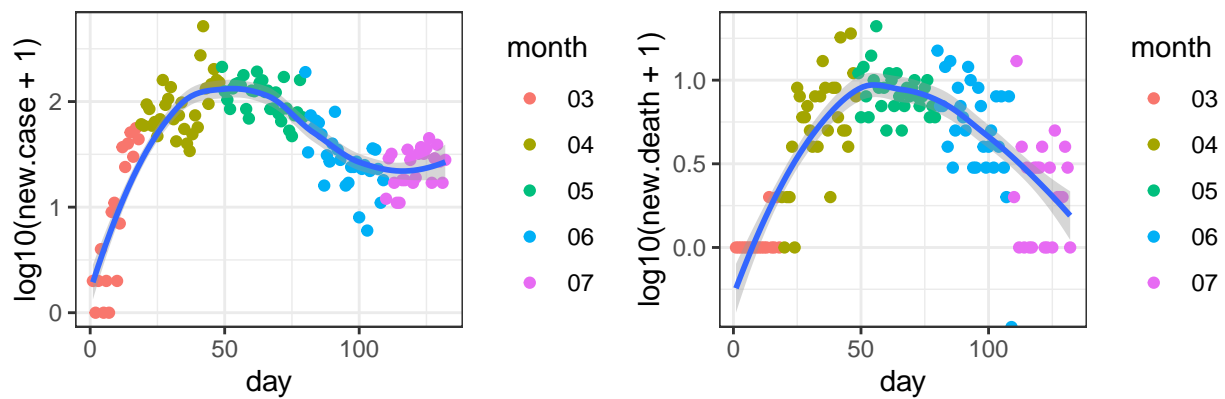
Erie_New York
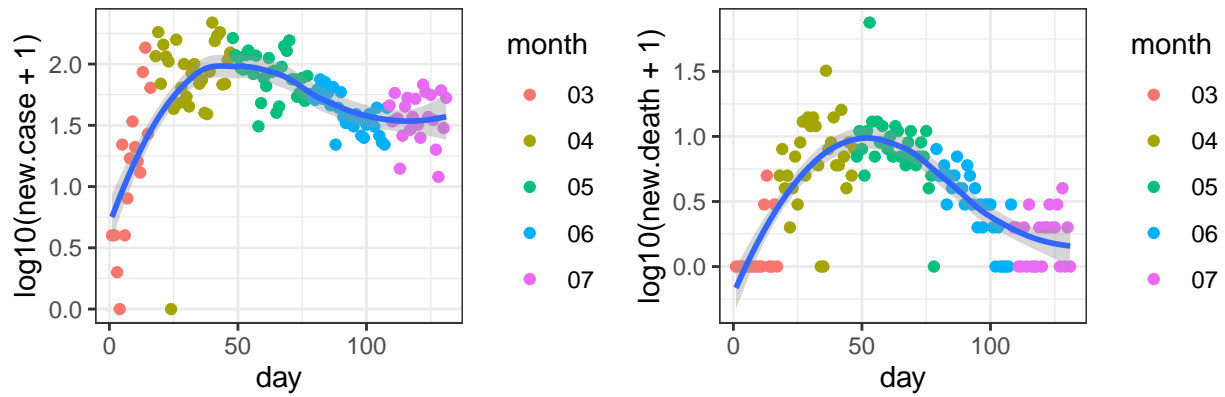
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15

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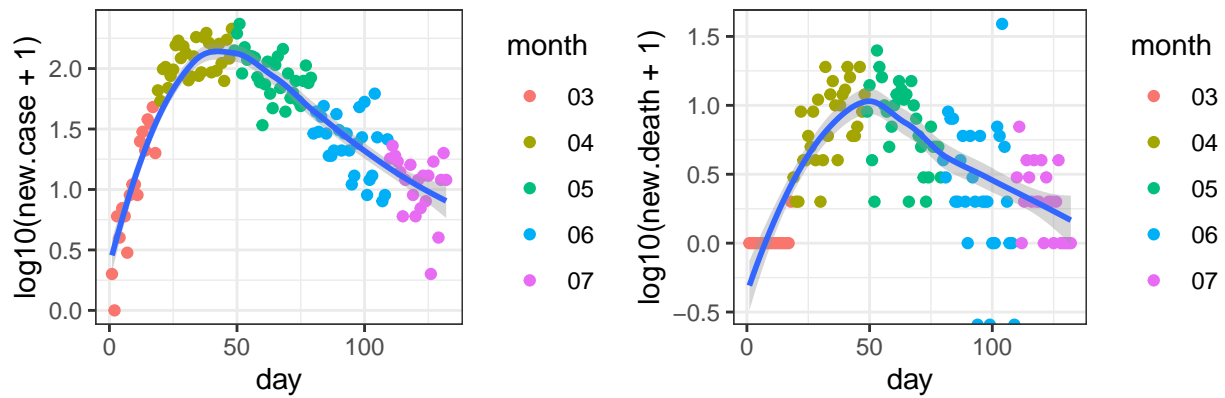
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14

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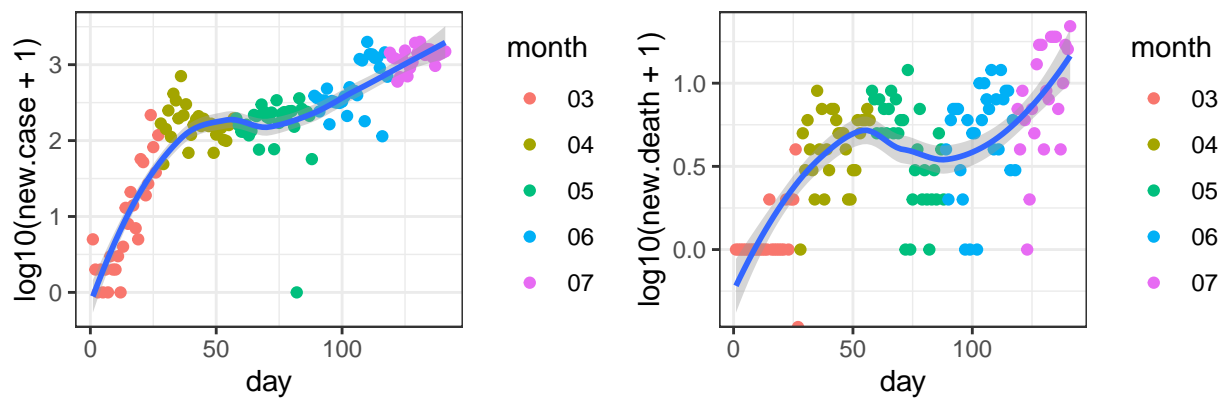
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

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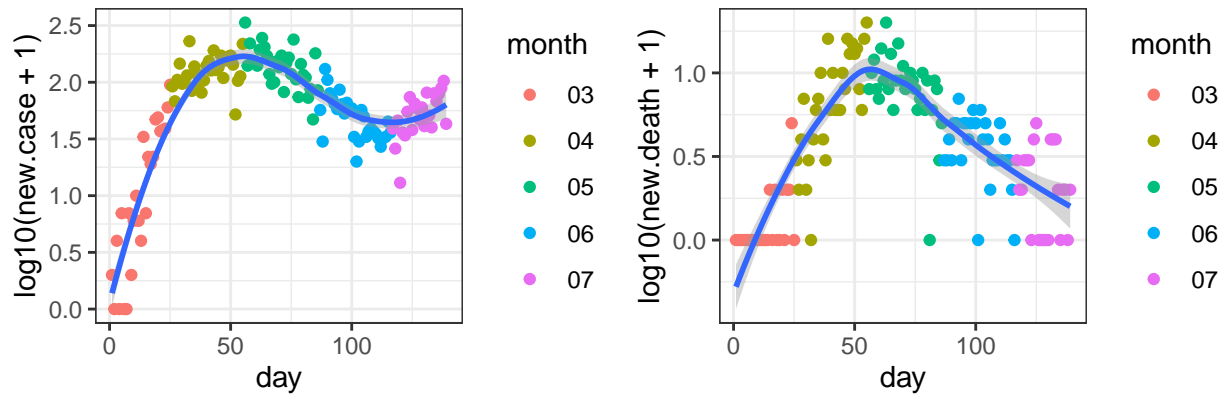
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

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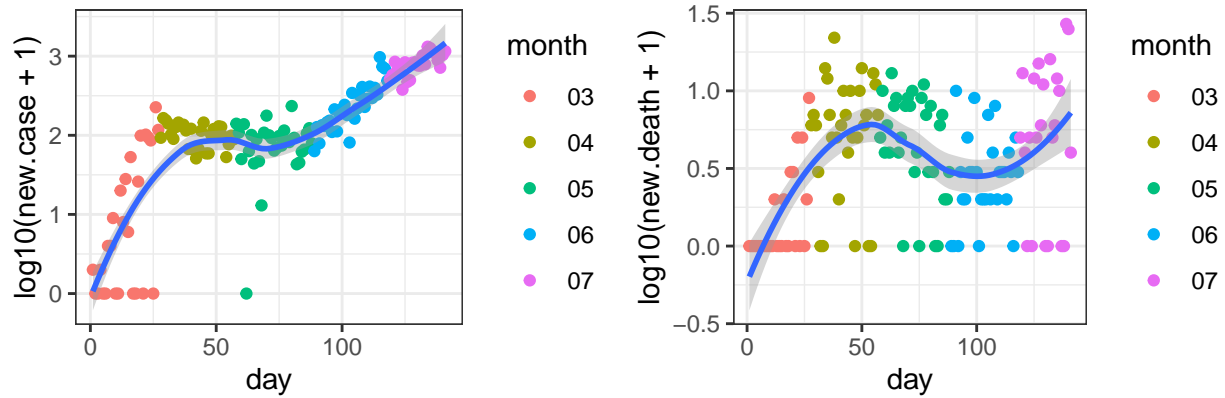
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Dallas_Texas

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

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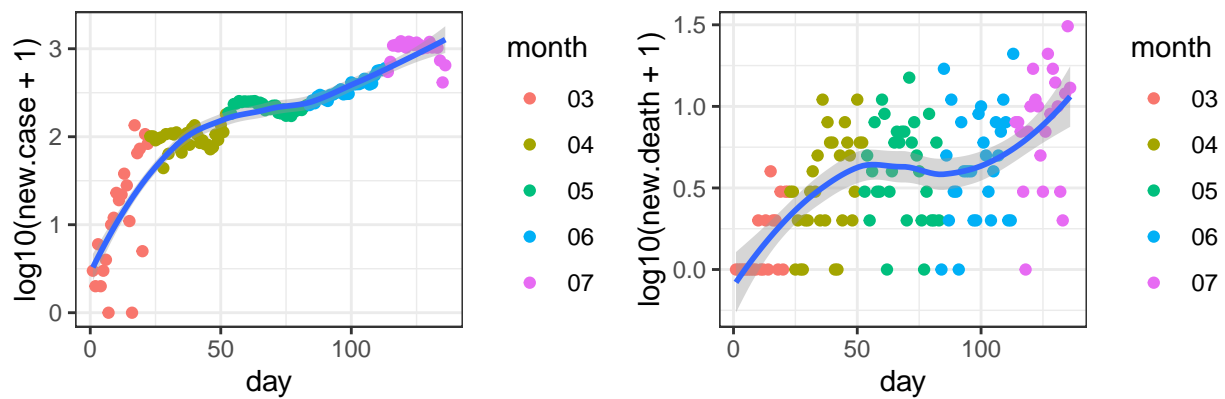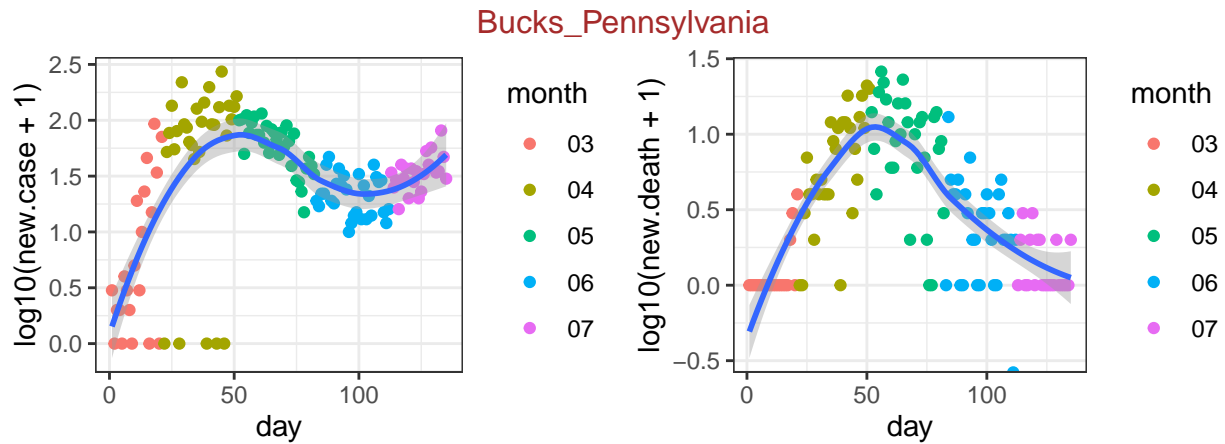
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much 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.

The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.
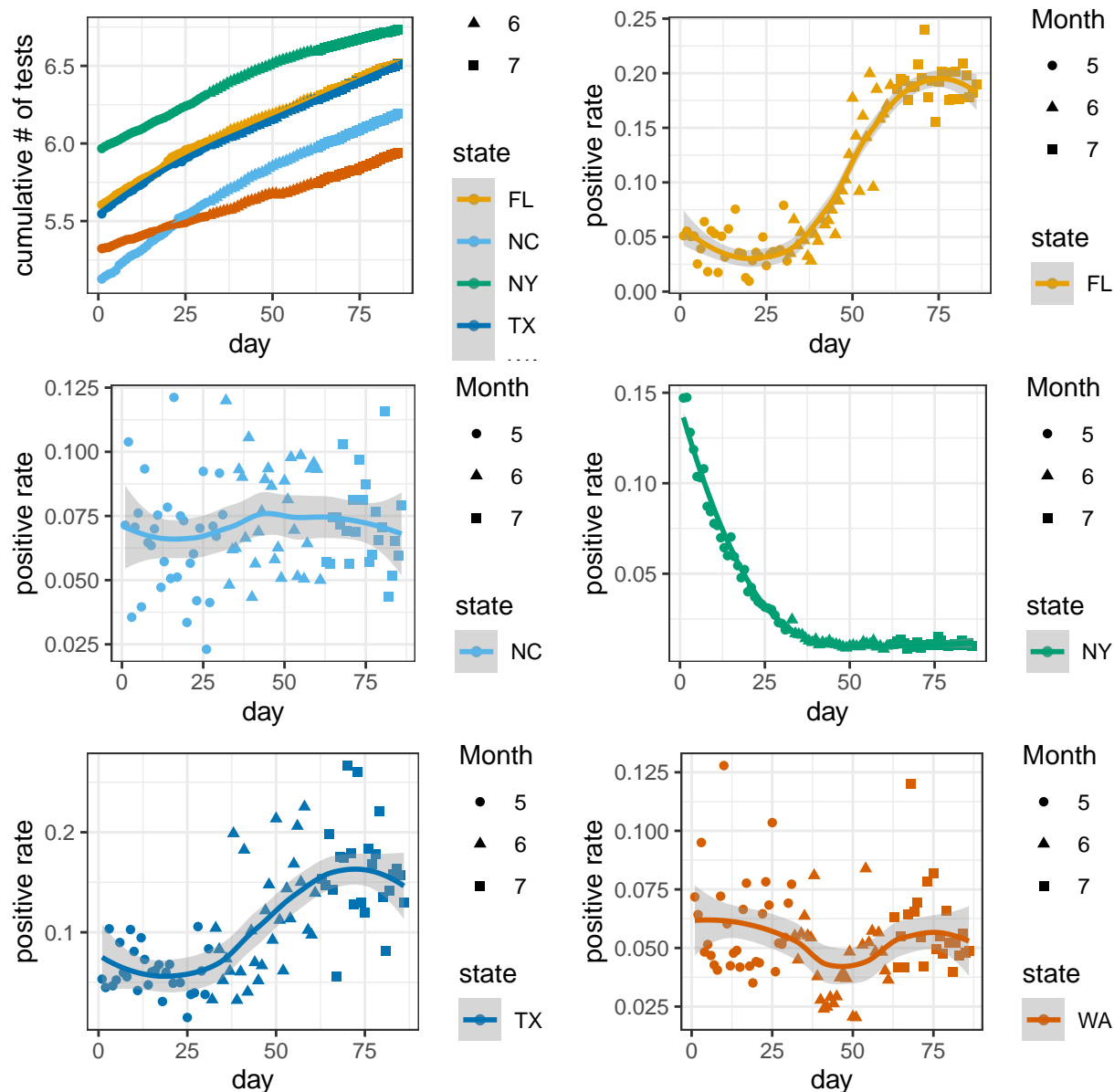
github.com/COVID19Tracking/, positive rate on 0724: 0.19(FL) 0.08(NC) 0.01(NY) 0.13(TX) 0.05(WA)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## 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_3.3.1
##
## 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    lattice_0.20-38  nlme_3.1-144     evaluate_0.14
##  [9] lifecycle_0.2.0  tibble_3.0.1     gtable_0.3.0     mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.6      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [21] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [25] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [29] rmarkdown_2.1    purrr_0.3.3      farver_2.0.3     splines_3.6.2
## [33] scales_1.1.0     ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1
## [37] colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3     stringi_1.4.5
## [41] munsell_0.5.0    crayon_1.3.4
```