Covid Data Analysis

pablo

4/29/2023

Importing the Data

We read the data directly from the below URL, ad display the frist rows to get an idea of the schema.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
pth<-"https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_

filenames<-c("time_series_covid19_confirmed_US.csv",  "time_series_covid19_confirmed_global.csv", "time_

#Read Files from URL
us_cases<-read_csv(paste(pth,filenames[1],sep=''))
```

```
## Rows: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_cases<-read_csv(paste(pth,filenames[2],sep=''))
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
```

```
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
us_deaths<-read_csv(paste(pth,filenames[3],sep=''))
```

```
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_deaths<-read_csv(paste(pth,filenames[4],sep=''))
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Transpose Global Cases
global_cases <- global_cases %>%
  pivot_longer(cols=-c('Province/State','Country/Region','Lat','Long'),
               names_to='date',
               values_to='cases') %>%
  select(-c('Lat','Long'))

#Transpose Global Deaths
global_deaths <- global_deaths %>%
  pivot_longer(cols=-c('Province/State','Country/Region','Lat','Long'),
               names_to='date',
               values_to='deaths') %>%
  select(-c('Lat','Long'))

#Merge Global Cases and Deaths
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region=`Country/Region`,
         Province_State= `Province/State`) %>%
  mutate(date=mdy(date))
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

```r
#Transpose US cases
us_cases <- us_cases %>%
  pivot_longer(cols=-c(UID:Combined_Key),
               names_to='date',
               values_to='cases') %>%
  select(Admin2:cases) %>%
  mutate(date=mdy(date)) %>%
  select(-c(Lat,Long_))
head(us_cases)
```

```
## # A tibble: 6 x 6
##   Admin2  Province_State Country_Region Combined_Key         date       cases
##   <chr>   <chr>          <chr>          <chr>                <date>     <dbl>
## 1 Autauga Alabama        US             Autauga, Alabama, US 2020-01-22     0
## 2 Autauga Alabama        US             Autauga, Alabama, US 2020-01-23     0
## 3 Autauga Alabama        US             Autauga, Alabama, US 2020-01-24     0
## 4 Autauga Alabama        US             Autauga, Alabama, US 2020-01-25     0
## 5 Autauga Alabama        US             Autauga, Alabama, US 2020-01-26     0
## 6 Autauga Alabama        US             Autauga, Alabama, US 2020-01-27     0
```

```r
#Transpose US deaths
head(us_deaths)
```

```
## # A tibble: 6 x 1,155
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,146 more variables: Long_ <dbl>, Combined_Key <chr>, Population <dbl>,
## #   `1/22/20` <dbl>, `1/23/20` <dbl>, `1/24/20` <dbl>, `1/25/20` <dbl>,
## #   `1/26/20` <dbl>, `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>,
## #   `1/30/20` <dbl>, `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>,
## #   `2/3/20` <dbl>, `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>,
## #   `2/7/20` <dbl>, `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>,
## #   `2/11/20` <dbl>, `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>, ...
```

```r
us_deaths <- us_deaths %>%
  pivot_longer(cols=-c(UID:Population),
               names_to='date',
               values_to='deaths') %>%
  select(Admin2:deaths) %>%
  mutate(date=mdy(date)) %>%
  select(-c(Lat,Long_))
head(us_deaths)
```

```
## # A tibble: 6 x 7
##   Admin2 Province_State Country_Region Combined_Key Population date       deaths
##   <chr>  <chr>          <chr>          <chr>             <dbl> <date>      <dbl>
```

```
## 1 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-22      0
## 2 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-23      0
## 3 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-24      0
## 4 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-25      0
## 5 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-26      0
## 6 Autau~ Alabama        US             Autauga, Al~      55869 2020-01-27      0
```

```
#Merge US cases and deaths
us<-us_cases %>%
  full_join(us_deaths)
```

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

```
head(us)
```

```
## # A tibble: 6 x 8
##   Admin2  Province_State Country_Region Combined_Key date       cases Population
##   <chr>   <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
## 1 Autauga Alabama        US             Autauga, Al~ 2020-01-22     0      55869
## 2 Autauga Alabama        US             Autauga, Al~ 2020-01-23     0      55869
## 3 Autauga Alabama        US             Autauga, Al~ 2020-01-24     0      55869
## 4 Autauga Alabama        US             Autauga, Al~ 2020-01-25     0      55869
## 5 Autauga Alabama        US             Autauga, Al~ 2020-01-26     0      55869
## 6 Autauga Alabama        US             Autauga, Al~ 2020-01-27     0      55869
## # i 1 more variable: deaths <dbl>
```

```
#Create a combined key for global
global <- global %>%
  unite("Combined_Key",
        c(Province_State,Country_Region),
        sep=",",
        na.rm=TRUE,
        remove= FALSE)
#Import Population
uid<-read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_
  select(-c(Lat, Long_,Combined_Key,code3,iso2,iso3,Admin2))
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(uid)
```

```
## # A tibble: 6 x 5
##     UID FIPS  Province_State Country_Region Population
```

```
##   <dbl> <chr> <chr>          <chr>               <dbl>
## 1     4 <NA>  <NA>           Afghanistan      38928341
## 2     8 <NA>  <NA>           Albania           2877800
## 3    10 <NA>  <NA>           Antarctica             NA
## 4    12 <NA>  <NA>           Algeria          43851043
## 5    20 <NA>  <NA>           Andorra             77265
## 6    24 <NA>  <NA>           Angola           32866268
```

```
global<-global %>%
  left_join(uid,by=c('Province_State','Country_Region')) %>%
  select(-c(UID,FIPS)) %>%
  select(c(Province_State,Country_Region,date, cases,deaths,Population,Combined_Key))

head(global)
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths Population Combined_Key
##   <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
## 1 <NA>           Afghanistan    2020-01-22     0      0   38928341 Afghanistan
## 2 <NA>           Afghanistan    2020-01-23     0      0   38928341 Afghanistan
## 3 <NA>           Afghanistan    2020-01-24     0      0   38928341 Afghanistan
## 4 <NA>           Afghanistan    2020-01-25     0      0   38928341 Afghanistan
## 5 <NA>           Afghanistan    2020-01-26     0      0   38928341 Afghanistan
## 6 <NA>           Afghanistan    2020-01-27     0      0   38928341 Afghanistan
```

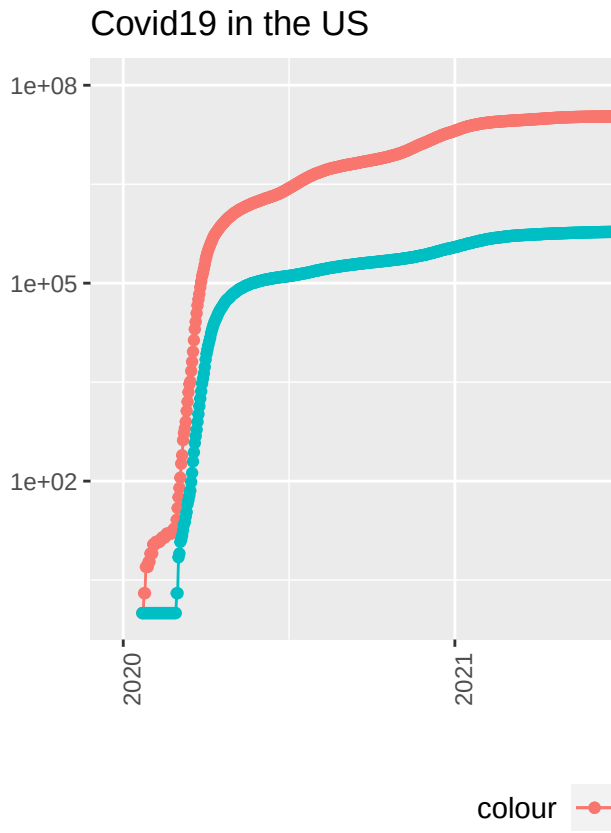
Visualization

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths deaths_per_mill
##   <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
## 1 Alabama        US             2020-01-22     0      0               0
## 2 Alabama        US             2020-01-23     0      0               0
## 3 Alabama        US             2020-01-24     0      0               0
## 4 Alabama        US             2020-01-25     0      0               0
## 5 Alabama        US             2020-01-26     0      0               0
## 6 Alabama        US             2020-01-27     0      0               0
## # i 1 more variable: Population <dbl>
```

```
## `summarise()` has grouped output by 'Country_Region'. You can override using
## the `.groups` argument.
```

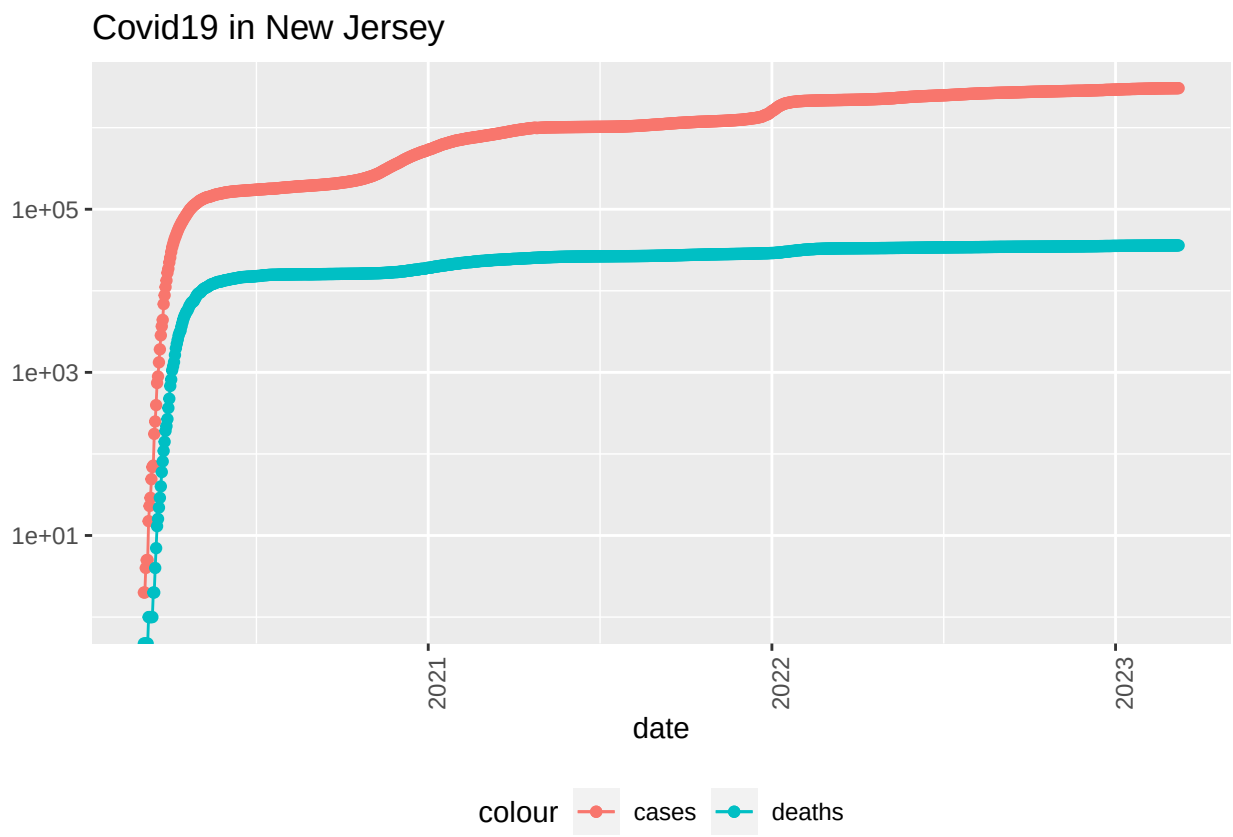
```
## # A tibble: 6 x 6
##   Country_Region date       cases deaths deaths_per_mill Population
##   <chr>          <date>     <dbl>  <dbl>           <dbl>      <dbl>
## 1 US             2020-01-22     1      1         0.00300  332875137
## 2 US             2020-01-23     1      1         0.00300  332875137
## 3 US             2020-01-24     2      1         0.00300  332875137
## 4 US             2020-01-25     2      1         0.00300  332875137
## 5 US             2020-01-26     5      1         0.00300  332875137
## 6 US             2020-01-27     5      1         0.00300  332875137
```



Let's see the cumulative Covid Cases and Deaths in the US through time

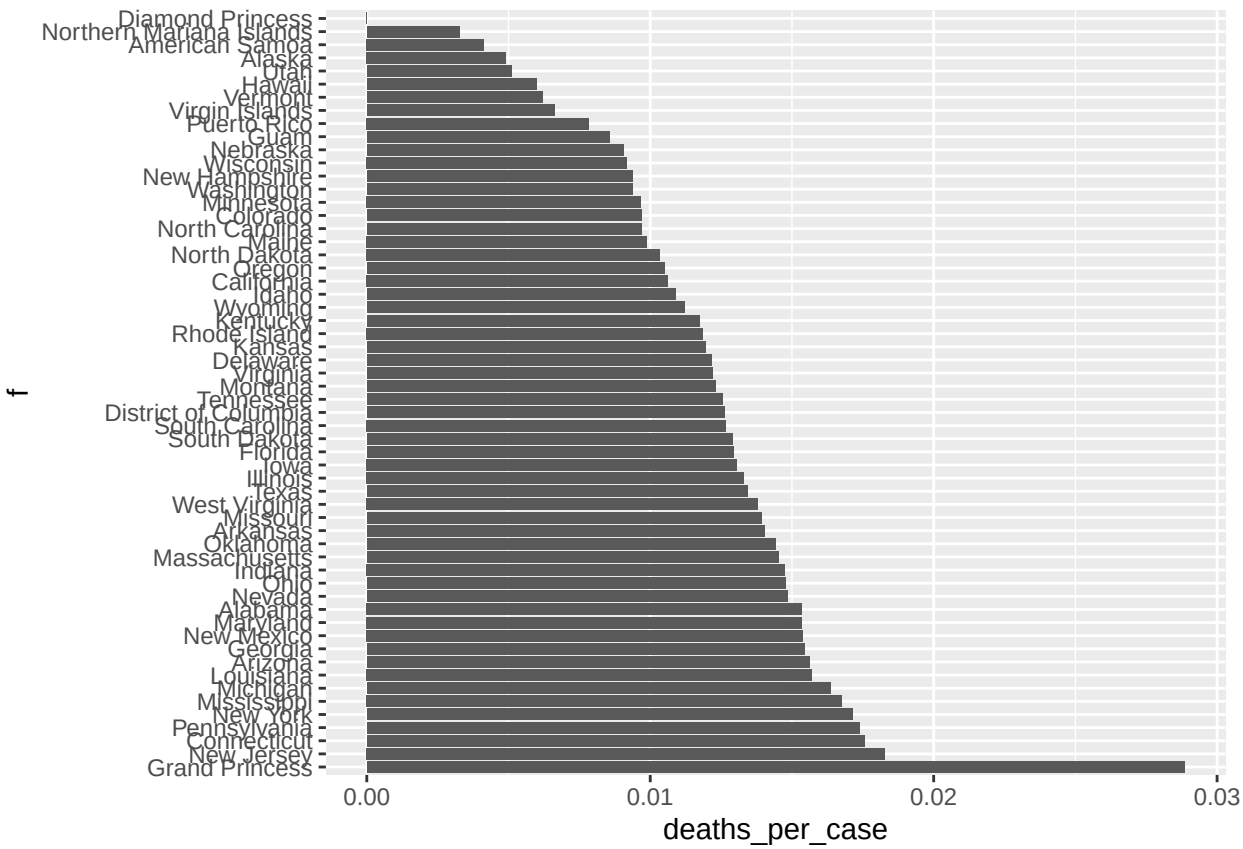
Now, for a specific state, lets say New Jersey

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```



#Model:

We are going to see if there is a significant difference of deaths per case across states in the US To do so, lets fit an linear model for death as a function of cases with Province/State as a factor.



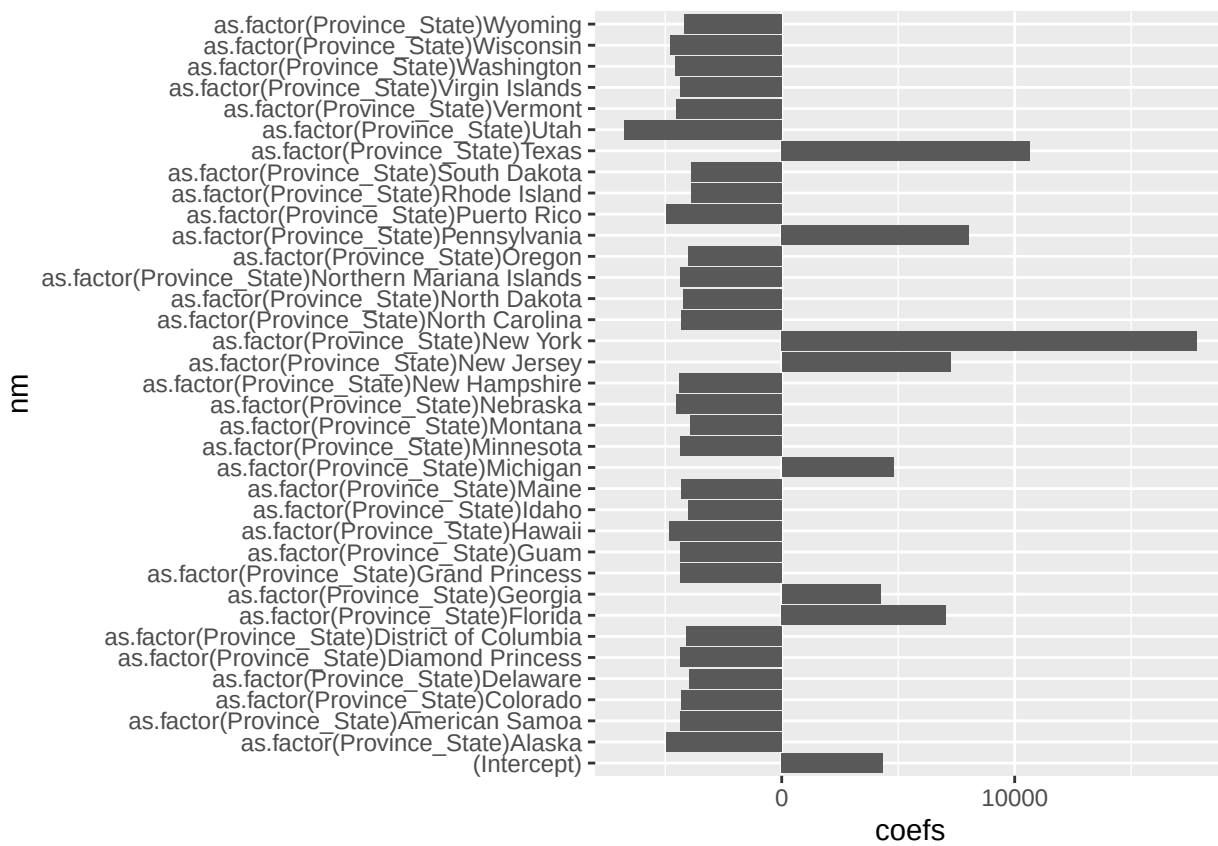
From the previous plot we expect Grand Princess and New Jersey to have the stronger effect. So these have the highest death per case. The coefficient in our Province_Region will adjust upward the relation

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.

##
## Call:
## lm(formula = deaths ~ cases + as.factor(Province_State), data = US_by_state)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22155.3   -528.8     68.0   1031.4  20594.2
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     4.338e+03  9.656e+01  44.931
```

```
## cases                                          9.670e-03  1.259e-05 768.127
## as.factor(Province_State)Alaska               -4.975e+03  1.361e+02 -36.556
## as.factor(Province_State)American Samoa       -4.351e+03  1.362e+02 -31.943
## as.factor(Province_State)Arizona               2.595e+03  1.360e+02  19.087
## as.factor(Province_State)Arkansas             -2.235e+03  1.359e+02 -16.445
## as.factor(Province_State)California            7.913e+02  1.479e+02   5.352
## as.factor(Province_State)Colorado             -4.319e+03  1.359e+02 -31.784
## as.factor(Province_State)Connecticut          -8.367e+02  1.359e+02  -6.156
## as.factor(Province_State)Delaware             -3.965e+03  1.361e+02 -29.133
## as.factor(Province_State)Diamond Princess     -4.339e+03  1.362e+02 -31.853
## as.factor(Province_State)District of Columbia -4.105e+03  1.361e+02 -30.150
## as.factor(Province_State)Florida               7.039e+03  1.401e+02  50.241
## as.factor(Province_State)Georgia               4.238e+03  1.362e+02  31.122
## as.factor(Province_State)Grand Princess       -4.336e+03  1.362e+02 -31.836
## as.factor(Province_State)Guam                 -4.365e+03  1.362e+02 -32.047
## as.factor(Province_State)Hawaii               -4.833e+03  1.361e+02 -35.511
## as.factor(Province_State)Idaho                -4.025e+03  1.360e+02 -29.588
## as.factor(Province_State)Illinois              2.414e+03  1.366e+02  17.679
## as.factor(Province_State)Indiana               6.567e+02  1.359e+02   4.833
## as.factor(Province_State)Iowa                 -2.795e+03  1.359e+02 -20.565
## as.factor(Province_State)Kansas               -3.338e+03  1.359e+02 -24.557
## as.factor(Province_State)Kentucky             -2.818e+03  1.359e+02 -20.739
## as.factor(Province_State)Louisiana             3.197e+01  1.359e+02   0.235
## as.factor(Province_State)Maine                -4.312e+03  1.361e+02 -31.679
## as.factor(Province_State)Maryland             -8.337e+02  1.359e+02  -6.135
## as.factor(Province_State)Massachusetts         4.802e+02  1.359e+02   3.534
## as.factor(Province_State)Michigan              4.805e+03  1.361e+02  35.309
## as.factor(Province_State)Minnesota            -4.340e+03  1.359e+02 -31.940
## as.factor(Province_State)Mississippi          -1.037e+03  1.359e+02  -7.629
## as.factor(Province_State)Missouri             -7.730e+02  1.359e+02  -5.689
## as.factor(Province_State)Montana              -3.928e+03  1.361e+02 -28.861
## as.factor(Province_State)Nebraska             -4.511e+03  1.360e+02 -33.167
## as.factor(Province_State)Nevada               -2.127e+03  1.359e+02 -15.646
## as.factor(Province_State)New Hampshire        -4.385e+03  1.361e+02 -32.225
## as.factor(Province_State)New Jersey            7.245e+03  1.361e+02  53.244
## as.factor(Province_State)New Mexico           -2.626e+03  1.360e+02 -19.307
## as.factor(Province_State)New York              1.781e+04  1.387e+02 128.460
## as.factor(Province_State)North Carolina       -4.297e+03  1.362e+02 -31.547
## as.factor(Province_State)North Dakota         -4.242e+03  1.361e+02 -31.172
## as.factor(Province_State)Northern Mariana Islands -4.367e+03  1.362e+02 -32.062
## as.factor(Province_State)Ohio                  3.536e+03  1.362e+02  25.954
## as.factor(Province_State)Oklahoma             -1.472e+03  1.359e+02 -10.831
## as.factor(Province_State)Oregon               -4.011e+03  1.360e+02 -29.506
## as.factor(Province_State)Pennsylvania          8.041e+03  1.363e+02  59.003
## as.factor(Province_State)Puerto Rico          -4.979e+03  1.360e+02 -36.620
## as.factor(Province_State)Rhode Island         -3.883e+03  1.361e+02 -28.539
## as.factor(Province_State)South Carolina       -1.809e+03  1.359e+02 -13.314
## as.factor(Province_State)South Dakota         -3.875e+03  1.361e+02 -28.471
## as.factor(Province_State)Tennessee            -9.313e+02  1.360e+02  -6.849
## as.factor(Province_State)Texas                 1.066e+04  1.418e+02  75.163
## as.factor(Province_State)Utah                 -6.774e+03  1.359e+02 -49.848
## as.factor(Province_State)Vermont              -4.545e+03  1.362e+02 -33.381
## as.factor(Province_State)Virgin Islands       -4.367e+03  1.362e+02 -32.063
## as.factor(Province_State)Virginia             -1.847e+03  1.359e+02 -13.590
```

```
## as.factor(Province_State)Washington              -4.569e+03  1.359e+02 -33.625
## as.factor(Province_State)West Virginia           -3.211e+03  1.360e+02 -23.608
## as.factor(Province_State)Wisconsin               -4.792e+03  1.359e+02 -35.261
## as.factor(Province_State)Wyoming                 -4.202e+03  1.361e+02 -30.868
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## cases                                             < 2e-16 ***
## as.factor(Province_State)Alaska                   < 2e-16 ***
## as.factor(Province_State)American Samoa           < 2e-16 ***
## as.factor(Province_State)Arizona                  < 2e-16 ***
## as.factor(Province_State)Arkansas                 < 2e-16 ***
## as.factor(Province_State)California              8.74e-08 ***
## as.factor(Province_State)Colorado                 < 2e-16 ***
## as.factor(Province_State)Connecticut             7.52e-10 ***
## as.factor(Province_State)Delaware                 < 2e-16 ***
## as.factor(Province_State)Diamond Princess         < 2e-16 ***
## as.factor(Province_State)District of Columbia     < 2e-16 ***
## as.factor(Province_State)Florida                  < 2e-16 ***
## as.factor(Province_State)Georgia                  < 2e-16 ***
## as.factor(Province_State)Grand Princess           < 2e-16 ***
## as.factor(Province_State)Guam                     < 2e-16 ***
## as.factor(Province_State)Hawaii                   < 2e-16 ***
## as.factor(Province_State)Idaho                    < 2e-16 ***
## as.factor(Province_State)Illinois                 < 2e-16 ***
## as.factor(Province_State)Indiana                 1.35e-06 ***
## as.factor(Province_State)Iowa                     < 2e-16 ***
## as.factor(Province_State)Kansas                   < 2e-16 ***
## as.factor(Province_State)Kentucky                 < 2e-16 ***
## as.factor(Province_State)Louisiana                0.81398
## as.factor(Province_State)Maine                    < 2e-16 ***
## as.factor(Province_State)Maryland                8.54e-10 ***
## as.factor(Province_State)Massachusetts            0.00041 ***
## as.factor(Province_State)Michigan                 < 2e-16 ***
## as.factor(Province_State)Minnesota                < 2e-16 ***
## as.factor(Province_State)Mississippi             2.40e-14 ***
## as.factor(Province_State)Missouri                1.28e-08 ***
## as.factor(Province_State)Montana                  < 2e-16 ***
## as.factor(Province_State)Nebraska                 < 2e-16 ***
## as.factor(Province_State)Nevada                   < 2e-16 ***
## as.factor(Province_State)New Hampshire            < 2e-16 ***
## as.factor(Province_State)New Jersey               < 2e-16 ***
## as.factor(Province_State)New Mexico               < 2e-16 ***
## as.factor(Province_State)New York                 < 2e-16 ***
## as.factor(Province_State)North Carolina           < 2e-16 ***
## as.factor(Province_State)North Dakota             < 2e-16 ***
## as.factor(Province_State)Northern Mariana Islands  < 2e-16 ***
## as.factor(Province_State)Ohio                     < 2e-16 ***
## as.factor(Province_State)Oklahoma                 < 2e-16 ***
## as.factor(Province_State)Oregon                   < 2e-16 ***
## as.factor(Province_State)Pennsylvania             < 2e-16 ***
## as.factor(Province_State)Puerto Rico              < 2e-16 ***
## as.factor(Province_State)Rhode Island             < 2e-16 ***
## as.factor(Province_State)South Carolina           < 2e-16 ***
## as.factor(Province_State)South Dakota             < 2e-16 ***
```

```
## as.factor(Province_State)Tennessee                7.50e-12 ***
## as.factor(Province_State)Texas                     < 2e-16 ***
## as.factor(Province_State)Utah                      < 2e-16 ***
## as.factor(Province_State)Vermont                   < 2e-16 ***
## as.factor(Province_State)Virgin Islands            < 2e-16 ***
## as.factor(Province_State)Virginia                  < 2e-16 ***
## as.factor(Province_State)Washington                < 2e-16 ***
## as.factor(Province_State)West Virginia             < 2e-16 ***
## as.factor(Province_State)Wisconsin                 < 2e-16 ***
## as.factor(Province_State)Wyoming                   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3248 on 66235 degrees of freedom
## Multiple R-squared:  0.9626, Adjusted R-squared:  0.9626
## F-statistic: 2.943e+04 on 58 and 66235 DF,  p-value: < 2.2e-16
```



After controlling for number of cases, the deadliest ones are NJ, NY and Texas.

BIAS

There are three sources of bias that i can identify, all related with the prevalence of testing.

1. The first is fundamental almost tautological, the results depend on the prevalence of testing. If you

have low testing rates then you are most likely testing the patients that have worse prognosis, so the deaths per confirmed case will increase.

2. In the US the testing/vaccines got politicized and response to the pandemic got split across political spectrum like red states vs blue states.

3. Wealthier, mega cities in both coasts have very high density which helps transmission and overwhelmed the medical response.