Project: COVID-19

Yuhao Mao

目录

1 Introduction

COVID-19, also known as corona virus, has enforced people to stay home for more than half a year. Since its breakout, Chinese government has taken numerous attempts, including lock-down, to prevent its spread. Although now the epidemic in China has been well controlled, in other parts of the world, such as the United States, the virus still remains powerful.

Numerous scientists have tried to help. Some researchers proposed new models to predict the spread trend of the epidemic, some researchers focused on inventing a cure for the pneumonia caused by COVID-19 and others analyze the data of the epidemic to get an insight. To facilitate the analysis and make the disease data open to everyone, John Hopkins University has been maintaining a Github repository containing daily data [1]. This repository collects data from the visual dashboard of JHU CSSE[2] and stores in the format of csv file.

In this project, I will use the daily data downloaded from the repository to do some analysis.

2 Load data and preprocess

The raw data size is:

```
dim(confirmed.global)
```

```
## [1] 266 177
```

```
dim(death.global)
```

```
## [1] 266 177
```

In the raw data, dates are formatted to be columns (only display the previous ten column names), along with the country, the latitude and the longitude:

```
colnames(confirmed.global)[1:10]
```

```
##  [1] "Province/State" "Country/Region" "Lat"            "Long"
##  [5] "1/22/20"        "1/23/20"        "1/24/20"        "1/25/20"
##  [9] "1/26/20"        "1/27/20"
```

To make the data format suitable for analysis, we need to reformat it so that date is a single column.

[1] https://github.com/CSSEGISandData/COVID-19

[2] https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

```
confirmed.global <- confirmed.global %>%
  pivot_longer(-c(`Province/State`,`Country/Region`, `Lat`, `Long`),
               names_to="Date",
               values_to="count")
death.global <- death.global %>%
  pivot_longer(-c(`Province/State`,`Country/Region`, `Lat`, `Long`),
               names_to="Date",
               values_to="count")
confirmed.global$Date <- as.Date(confirmed.global$Date, "%m/%d/%y")
death.global$Date <- as.Date(death.global$Date, "%m/%d/%y")
head(confirmed.global, n=2)
```

```
## # A tibble: 2 x 6
##   `Province/State` `Country/Region`   Lat  Long Date       count
##   <chr>            <chr>            <dbl> <dbl> <date>     <dbl>
## 1 <NA>             Afghanistan         33    65 2020-01-22     0
## 2 <NA>             Afghanistan         33    65 2020-01-23     0
```

```
head(death.global, n=2)
```

```
## # A tibble: 2 x 6
##   `Province/State` `Country/Region`   Lat  Long Date       count
##   <chr>            <chr>            <dbl> <dbl> <date>     <dbl>
## 1 <NA>             Afghanistan         33    65 2020-01-22     0
## 2 <NA>             Afghanistan         33    65 2020-01-23     0
```

Let’s see we have data from when to when.

```
max(confirmed.global$Date)
```

```
## [1] "2020-07-12"
```

```
min(confirmed.global$Date)
```

```
## [1] "2020-01-22"
```

So we are getting data from 2020-01-22 to 2020-07-12. Do we have NA in columns apart from the `Province/State` column?

```r
any(is.na(confirmed.global[, 2:length(confirmed.global)]))
```

```
## [1] FALSE
```

```r
any(is.na(death.global[, 2:length(death.global)]))
```

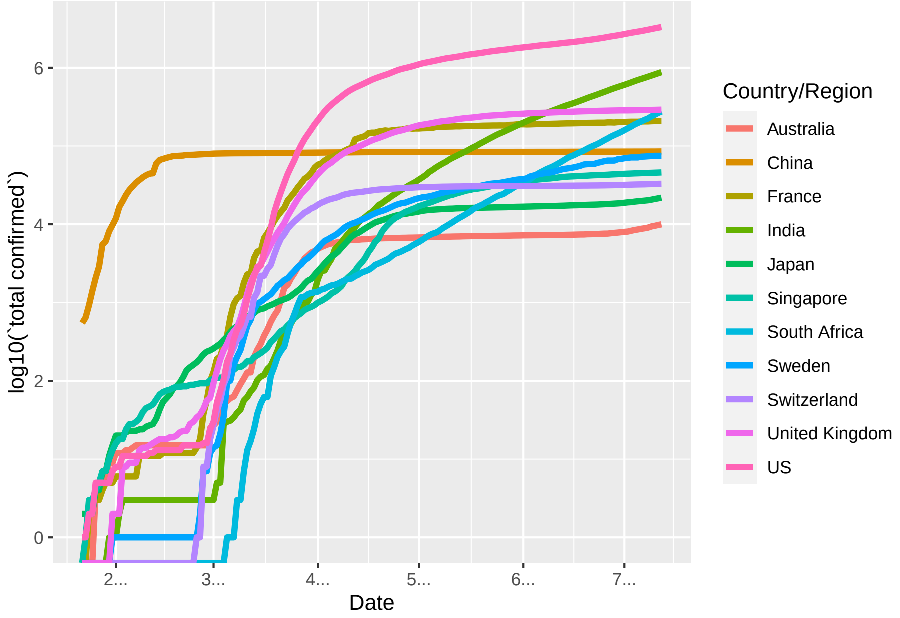
```
## [1] FALSE
```

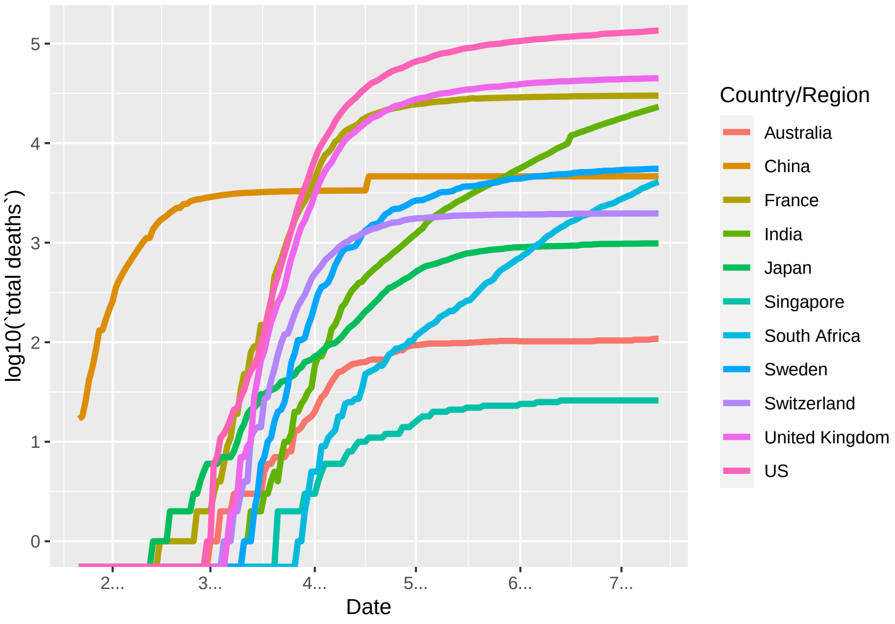
Luckily, there is no NA in the data. JHU is good at database maintaining!

3 Analysis of the data

3.1 The trend across the world.

To begin with, let's see how the confirmed cases and deaths evolve across time.



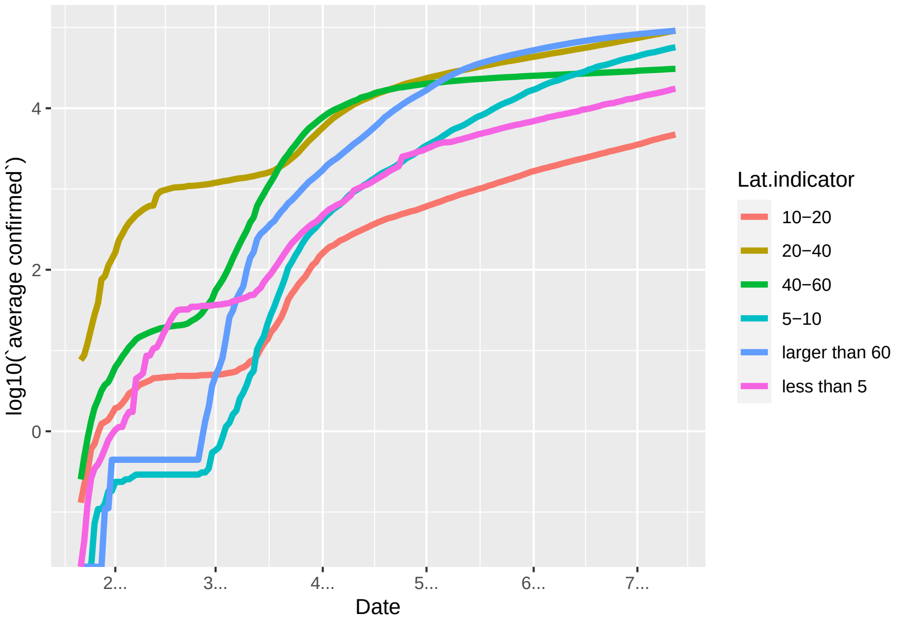


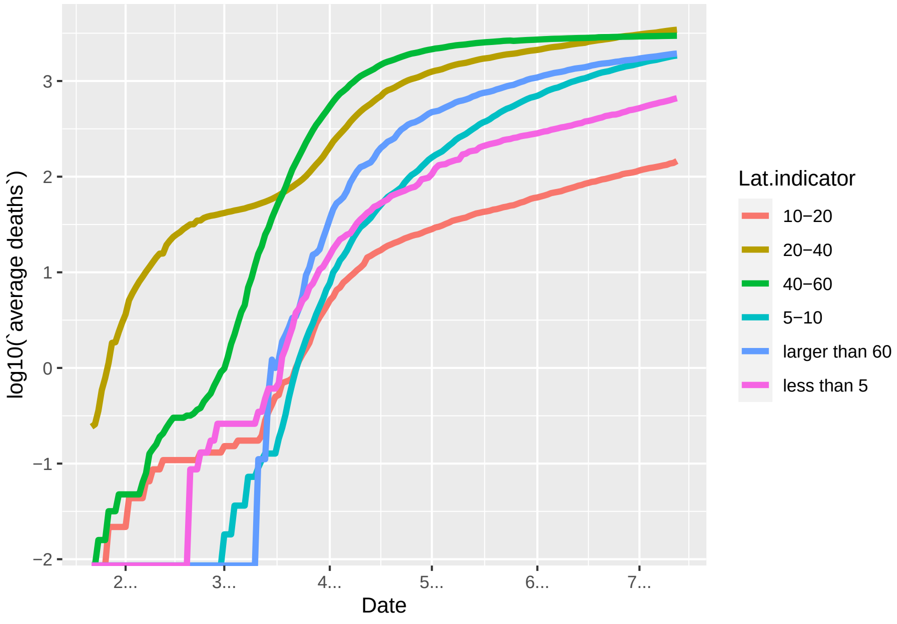
We can conclude that in the early phase of the epidemic in every country, the situation is out of control and the confirmed cases and deaths grow exponentially, following the exponential growth curve. So far, a lot of countries have managed to control the epidemic but some countries, such as the United States, India and South Africa, still has an exponential growth of infestors although the power is smaller than outbreak. China and European countries have almost no new cases now.

3.2 Trends in different parts of the world.

Many diseases show interesting phenomenons by temperature, which is roughly represented by the latitude. We discuss if latitude makes difference in the COVID infests.

We visualize the impact of latitude below.





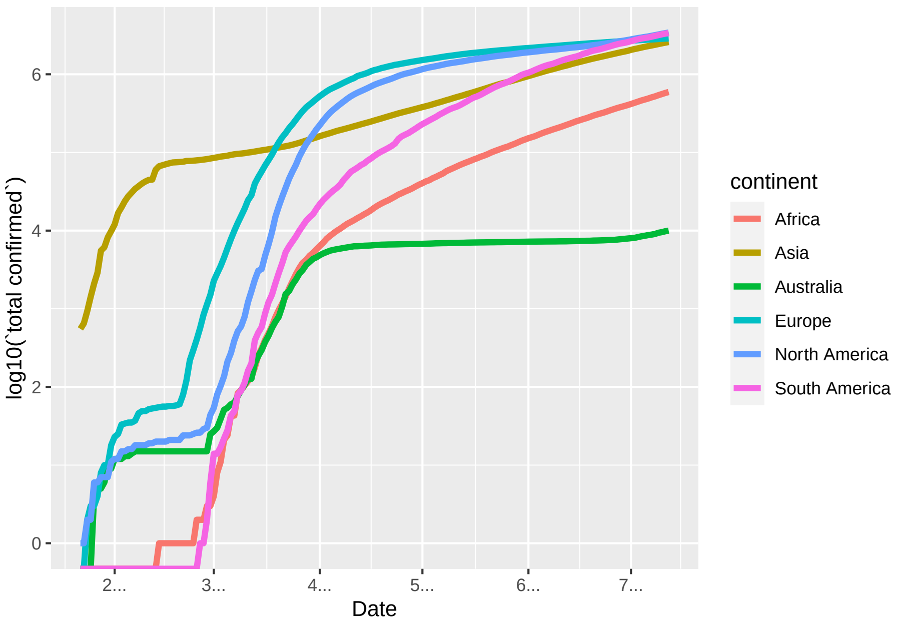
It seems that hot areas (latitude smaller than 20) and cold areas (latitude larger than 60) have less infestors. Of course, this conclusion can be biased because some countries with large number of infestors, such as the US and India, are located between 20-40.

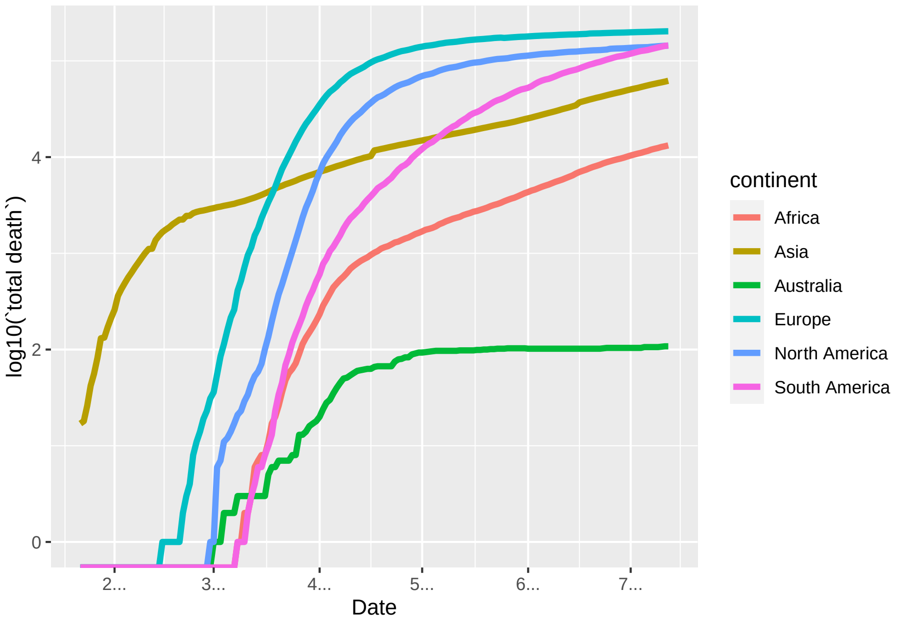
3.3 Difference across continents.

In general, Australia, Europe and North America mainly consist of developed countries that have robust medical systems and countries in Africa are poor and their medical systems are fragile. To see if it makes a difference in the spread of COVID-19, we plot the trends according to continent.

First of all, we only have longitude and latitude of each region, so we need to map the coordinates to continents. To do this, we need two R packages: `sp` and `rworldmap`.

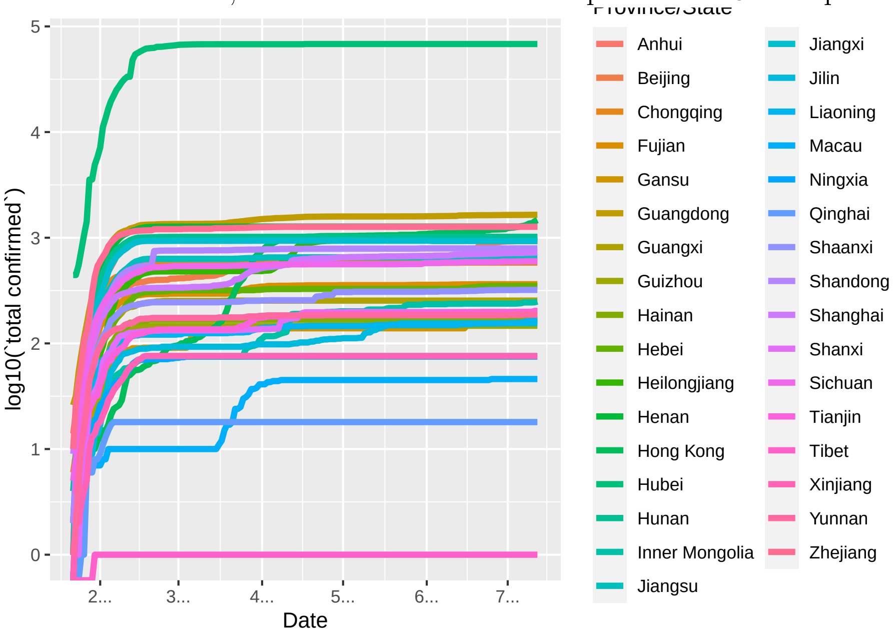
```r
library(sp)
library(rworldmap)

coords2continent = function(points)
{
  countriesSP <- getMap(resolution='low')
  pointsSP <-  SpatialPoints(
    points,proj4string=CRS(proj4string(countriesSP)))
  indices = over(pointsSP, countriesSP)
  indices$REGION
}
confirmed.global$continent <-
  coords2continent(as.data.frame(confirmed.global[,c("Long", "Lat")]))
death.global$continent <-
  coords2continent(as.data.frame(death.global[,c("Long", "Lat")]))
```

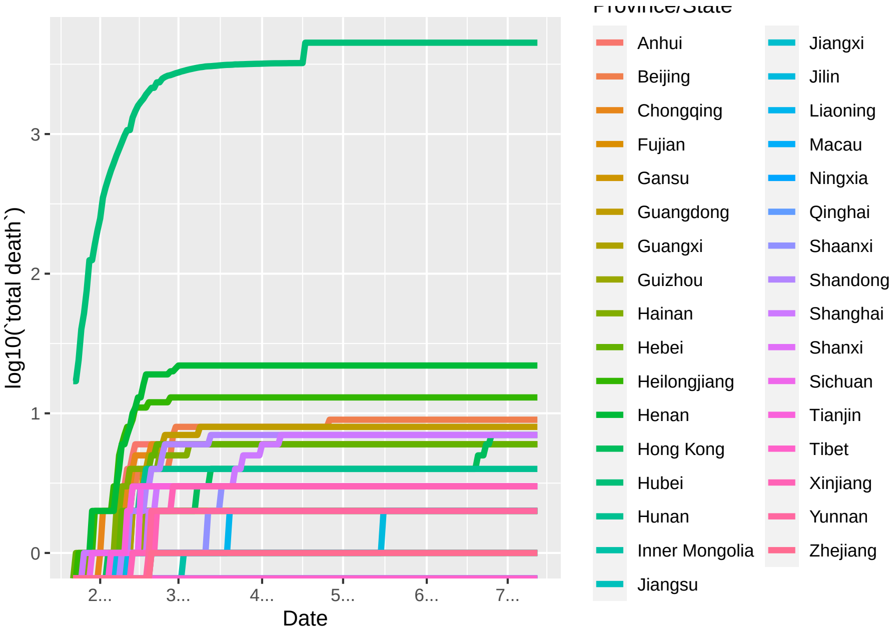

log10(`total confirmed`)
0
2
4
6
2...
3...
4...
5...
6...
7...
Date
continent
Africa
Asia
Australia
Europe
North America
South America


log10(`total death`)
0
2
4
2...
3...
4...
5...
6...
7...
Date
continent
Africa
Asia
Australia
Europe
North America
South America

3.4 Trends inside China.

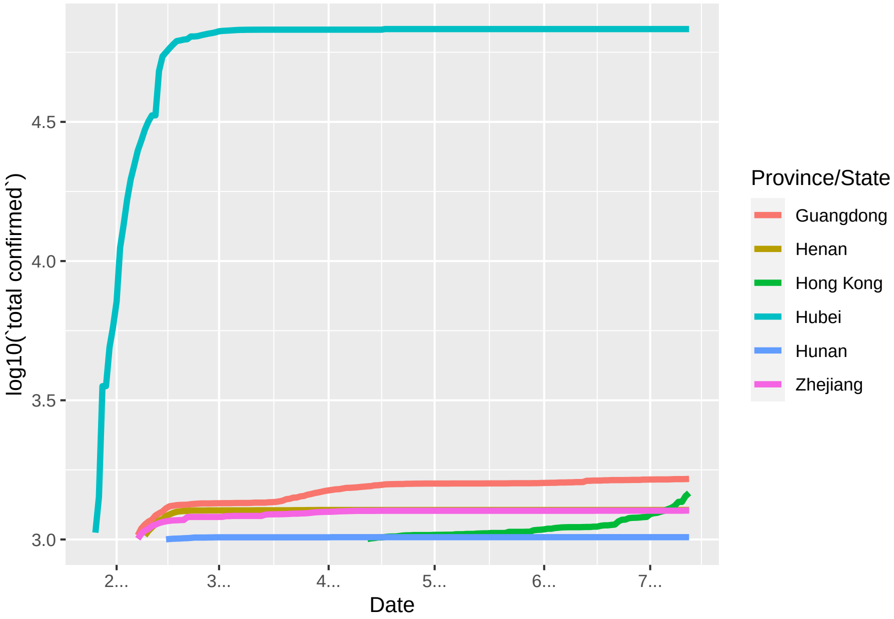
In this section, we will visualize how the epidemic in China spreads.





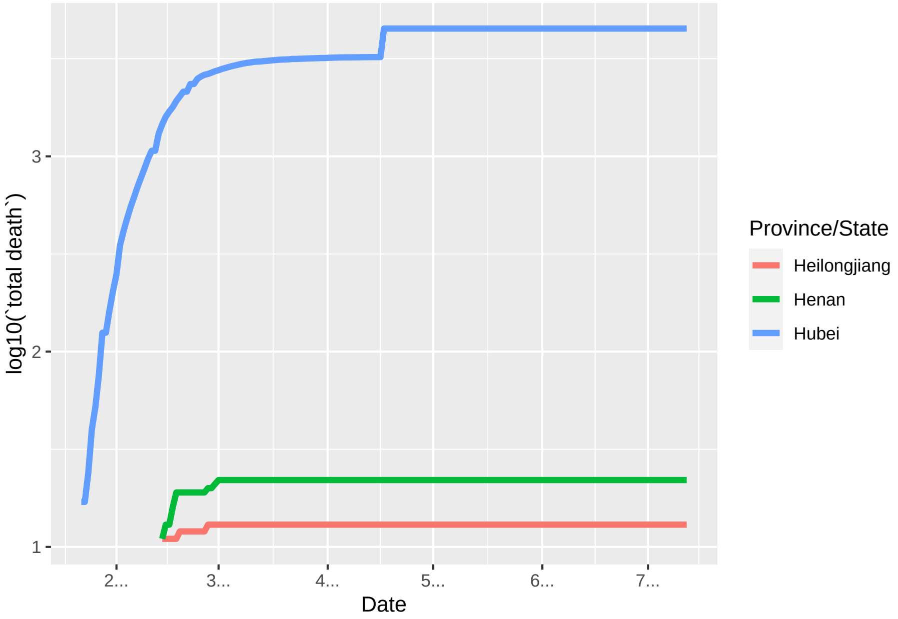
Clearly, after April, there are almost no new cases inside China mainland. It is not surprising to see Hubei has the superior number of infestors because the epidemic broke out there. This plot, although complete, contains too much information for us, so we decompose it below.

We filter out provinces with low number of cases.



The plot above shows provinces which have confirmed cases greater than 1000. Apart from Hubei, Henan and Hunan which is adjacent to Hubei also suffer from the spread. Besides, rich areas (Guangdong, Hong Kong and Zhejiang), due to their convenient transportation, also have relatively large number of cases.

Next, we show provinces with over ten deaths.



There are only three provinces that have more than ten deaths. What is surprising is that despite having low number of confirmed cases, Heilongjiang already has ten deaths in the middle Feburary, along with Henan. That could be resulted from the cold weather and the prevalent flu there.

4 Statistic models for the epidemic

Our goal is to estimate the power of medical testing systems in the early state of each continent. From plot in section 3.3 we know that during the whole March, every continent has a log linear growth curve, which means it is an exponential growth, i.e. the J curve. This kind of curve appears when the disease is totally out of control and is free to get the resources to spread. However, since this data is the number of confirmed cases, it does not fully reflect the real number of infestors. Instead, it shows the growth in the number of test because in that period the disease was spreading so fast that it results in an approximately constant positive rate in the test. Therefore, we can use the power constant of the growth in confirmed cases to estimate the product of the strength of medical testing systems and the infest rate.

What we are going to do is to extract the data from March 1st to April 1st and apply a linear model on log number of confirmed cases against the date.

```
march.confirmed <- confirmed.global %>%
  filter(!is.na(continent)) %>%
  group_by(`Date`,`continent`) %>%
  dplyr::summarize(`total confirmed`=sum(count)) %>%
  ungroup() %>%
  filter(Date>=as.Date('2020-03-01') &
           Date<=as.Date('2020-04-01'))
```

```
## `summarise()` regrouping output by 'Date' (override with `.groups` argument)
```

```
tb <- table(march.confirmed$continent)
continents <- (as.data.frame(tb) %>% filter(Freq>0))[,1]
regression.by.continent <- function(df){
  result <- lm(log(`total confirmed`)~Date, data=df)
  summary(result)
}
slopes <- c()
for(cont in continents){
  print(paste("Result for", cont))
  df <- march.confirmed %>% filter(continent==cont)
  result <- regression.by.continent(df)
  print(result)
  slopes <- c(slopes, result$coefficients[[2]])
}
```

```
## [1] "Result for Africa"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0001 -0.1117  0.1335  0.2083  0.4563
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.123e+03  1.143e+02  -36.06   <2e-16 ***
## Date         2.252e-01  6.236e-03   36.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3257 on 30 degrees of freedom
## Multiple R-squared:  0.9775, Adjusted R-squared:  0.9768
## F-statistic:  1304 on 1 and 30 DF,  p-value: < 2.2e-16
##
## [1] "Result for Asia"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.03379 -0.02503 -0.00760  0.01945  0.06901
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.416e+02  9.966e+00  -34.28   <2e-16 ***
## Date         1.926e-02  5.435e-04   35.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02839 on 30 degrees of freedom
## Multiple R-squared:  0.9767, Adjusted R-squared:  0.9759
## F-statistic:  1256 on 1 and 30 DF,  p-value: < 2.2e-16
##
## [1] "Result for Australia"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37380 -0.12025  0.00454  0.11343  0.31521
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.352e+03  5.886e+01  -56.95   <2e-16 ***
## Date         1.831e-01  3.210e-03   57.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1677 on 30 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9906
## F-statistic:  3255 on 1 and 30 DF,  p-value: < 2.2e-16
##
## [1] "Result for Europe"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5214 -0.1898  0.0475  0.2398  0.3087
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.266e+03  9.085e+01  -35.95   <2e-16 ***
## Date         1.787e-01  4.954e-03   36.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2588 on 30 degrees of freedom
## Multiple R-squared:  0.9775, Adjusted R-squared:  0.9767
## F-statistic:  1301 on 1 and 30 DF,  p-value: < 2.2e-16
##
## [1] "Result for North America"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62366 -0.12090  0.02733  0.14256  0.34632
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.128e+03  8.327e+01  -61.58   <2e-16 ***
## Date         2.801e-01  4.541e-03   61.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2372 on 30 degrees of freedom
## Multiple R-squared:  0.9922, Adjusted R-squared:  0.9919
## F-statistic:  3805 on 1 and 30 DF,  p-value: < 2.2e-16
##
## [1] "Result for South America"
##
## Call:
## lm(formula = log(`total confirmed`) ~ Date, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71016 -0.21033 -0.03295  0.27643  0.49754
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.805e+03  1.124e+02  -42.75   <2e-16 ***
## Date         2.624e-01  6.130e-03   42.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3202 on 30 degrees of freedom
## Multiple R-squared:  0.9839, Adjusted R-squared:  0.9834
## F-statistic:  1832 on 1 and 30 DF,  p-value: < 2.2e-16
```

Not surprisingly (as we have seen from the visualization), all coefficients are statistically significant and R squares are all large (around 0.98). Now we compare their slopes, which is the power constant of the growth.

```
##        slope     continent
## 1 0.22516802        Africa
## 2 0.01926155          Asia
## 3 0.18314731     Australia
## 4 0.17869949        Europe
## 5 0.28011597 North America
## 6 0.26240738 South America
```

Among all continents, Asia has the smallest power constant. That is partially because rich countries in Asia attach great importance to the epidemic and have been dealing with it for a month while poor countries lack the ability to test. It is not surprising that Europe, Australia and Africa have smaller slope than America. The reasons are three-folded. Firstly, Australia and Africa have a vast land and relatively small entrants compared to Europe and America, so number of infests are small. Secondly, countries in Europe, such as Italy that has been known in that period as the first follower of Chinese government's lock-down policy, suffer from breakdown of medical systems, which is a clear sign of medical testing system deficiency. Thirdly, countries in America, e.g. Canada and the US, have strong economic power as well as medical system. Therefore, they are able to give more tests in the early period of the spread.

5 Conclusion

In this project, I use the daily data collected from JHU CSSE project to do analysis based on R. I use functions in the `tidyverse` package to reformat the data and `ggplot` package to visualize it. Furthermore, I use the linear model to estimate the medical power of each continent.