Assignment4

Qingyuan Li

2022-10-16

1. Discuss the business problem/goal (5 points) The business problem is to explore the relationship of the accumulative cases and deaths at county level.
2. identify where the dataset was retrieved from (2 points)

The COVID-19 dataset by New York Times, 2020. (Because the whole dataset is too large, I only analyze 2020 data.) <https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties-2020.csv>

1. identify the code that imported and saved your dataset in R (3 points)
2. describe your data set (using the common attributes such as #rows, #columns, variable names, types, means, SD, min/max, NAs, etc…) (10 points)

The covid dataset has 6 columns, 884737 rows(and this is only for 2020’s record!). The variable names includes “date” “county” “state” “fips” “cases” “deaths”. date, county, state are character type, fips, cases, and deaths are integer type. Meaning: Date: the date of the record. county: the name of the county State: the name of the state Fips: the fips code of the county Cases: the accumulative number of the covid cases Deaths: the accumulative numer of the covid deaths

For the means, max, SD, NAs, please see the following summary:

summary(covid)

## date county state fips   
## Length:884737 Length:884737 Length:884737 Min. : 1001   
## Class :character Class :character Class :character 1st Qu.:18183   
## Mode :character Mode :character Mode :character Median :29215   
## Mean :31262   
## 3rd Qu.:46099   
## Max. :78030   
## NA's :8266   
## cases deaths   
## Min. : 0 Min. : 0.0   
## 1st Qu.: 36 1st Qu.: 0.0   
## Median : 228 Median : 4.0   
## Mean : 1952 Mean : 53.6   
## 3rd Qu.: 993 3rd Qu.: 21.0   
## Max. :770915 Max. :25144.0   
## NA's :18761

1. discuss any data preparation, missing values and errors (10 points) (if the dataset was clean and there is no prep in the code, include a comment that explains what likely data preparation was done. What are the common issues with raw data?)

What I did to prepare the data is to clean it a little bit to get rid of fips NAs(that is, unrecognized counties), and also transform the character type of date into datetime type. There are no obvious error that I could detect from my knowledge.

1. discuss the modeling (10 points)

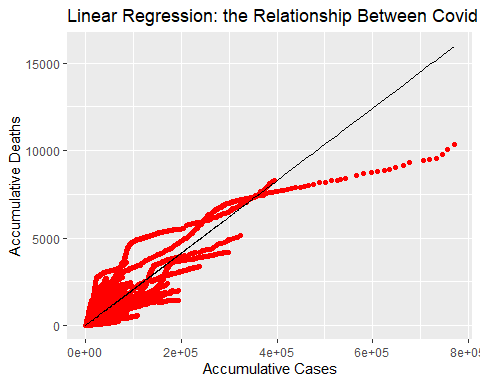
The model basically omit the “number of days” and focus on the number of covid cases and deaths, accummulatively. This treats different records equally and get rid of the limitation of “time series” or “panel data”(because the spread of virus doesn’t happen simultaneously in every states, and also doesn’t distribute evenly in many different regions). We just want to see while the covid cases accumulative number goes up, how the number of deaths changes, at every scale.

Therefore, I use a simple linear regression model, a quadratic regression model, and a log-linear model(we use log(x+1)) to study this issue, and let’s see which one fits the best.

1. produce and discuss the output (10 points)
2. provide explanation with any visuals (10 points) (if there are no visualizations in your code, discuss a visualization that would be useful to this project)

Below is the plot that shows the relationship between accumulative covid cases and deaths in 2020. Most of the data are aggregated in very small area(the bottom left corner) and yield an average pattern of linear relationship. The outliers are very obvious on the diagram - those that have large number of cases(and potentially from very large, dense urban areas). In those counties that have more population, the share of COVID-19 deaths seemed to be smaller than the general rate. In most counties, however, the death count’s relationship to accumulative cases is almost linear.

theplot



1. Your document should clearly and concisely communicate the project (10 points)

Conclusion:

Among the 3 models tested, the linear regression model describes the relationship best. In most counties, the accumulative number of COVID cases and deaths have a linear relationship.

Only in some large or special counties, the death rate(covid death/covid case) went high and then became low, a lot lower than the average. This somehow shows us the functionality of health infrastructure in large cities.