# COVID-19 Research Group: EDA Stage 2, Graphs

Eddy D. Varela, Chenyang Sun, Jackson Bibbens

4/27/2020

```
library(data.table)
countyData = read.csv("./data/countyData.csv")
```

This dataset consists of county-wide measurements of the following quantities:

Total Deaths, Total Cases, Income, Number of Beds, 2019 Population, and date when stay-at-home orders were initiated.

The counties were chosen from states that have the highest number of infections as of 27 April 2020 (a milestone with 1,000,000 domestic cases). Note that the data was synthesized from multiple sources and then cleaned.

```
hist(log(countyData$totalDeaths+0.1), main="Distribution of log(Deaths)",
     xlab="log(Deaths)", border="blue", col="green")
```

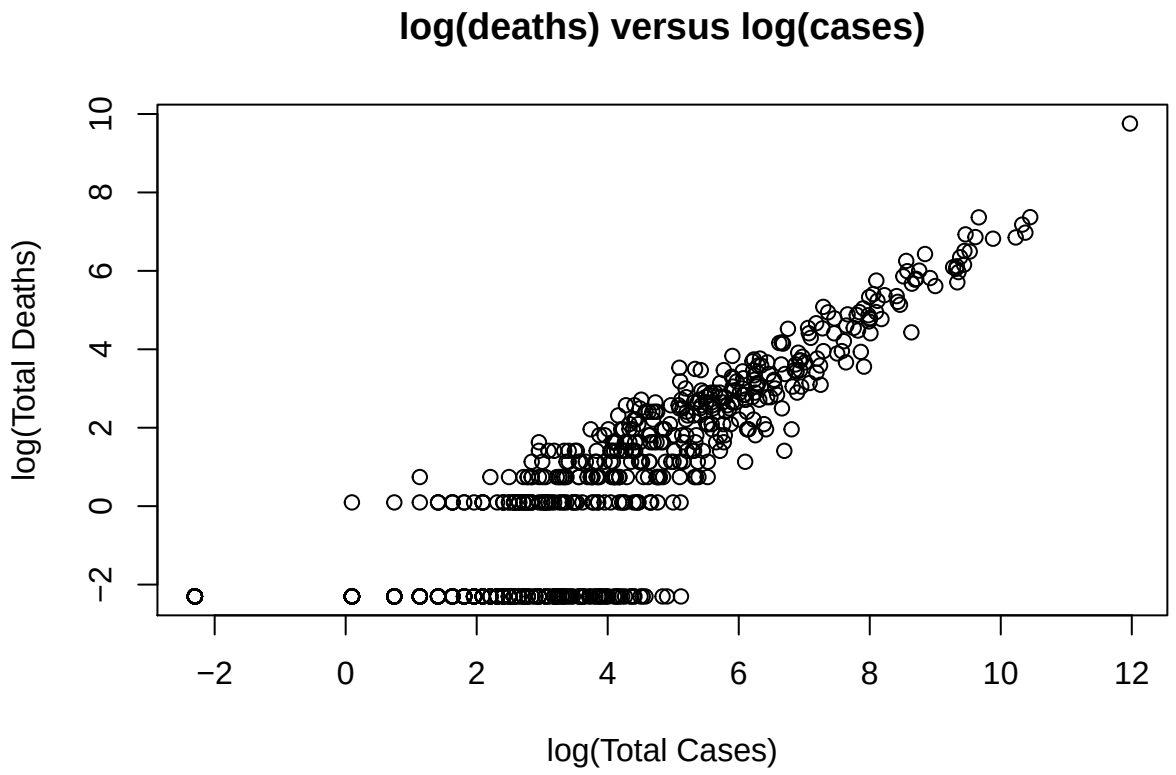
```r
summary(log(countyData$totalDeaths+0.1))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -2.30259 -2.30259  0.09531  0.14613  1.99304  9.75731        1
```

This histogram represents the distribution of log-transformed number of deaths in each county, where I have added 0.1 before taking the logarithm to i) avoid an undefined operation and ii) visually separate those with zero deaths from those with a nonzero number. The distribution appears unimodal and it is interesting to note the strong right skew even after the log-transform. There is one potential high outlier corresponding to New York.

Interestingly, the same transform gives the number of cases a slightly right-skewed and almost bell-shaped distribution rather than the strong skew of the number of deaths. This calls into question the stability of death rate as a measure, since a stable death rate would imply that the number of cases and the number of deaths would be roughly proportional and thus behave similarly under the same transform.

```r
plot(log(countyData$totalDeaths+0.1)~log(countyData$totalCases+0.1),
     main="log(deaths) versus log(cases)",xlab="log(Total Cases)",ylab="log(Total Deaths)")
```




```r
cor(log(countyData$totalDeaths+0.1),log(countyData$totalCases+0.1))
```

```
## [1] NA
```

This scatterplot show the logarithm of the number of deaths versus the logarithm of the number of cases for every county; again, 0.1 is added before taking the logarithm. There is a strong, positive, linear association

(with correlation 0.855) between log(Cases) and log(Deaths), with no obvious outliers. Notice that the scarcity of data at higher values of Cases and Deaths makes uniform uniform variance somewhat difficult to check.

One of our interests is to see which variables may be a good predictor for some measure of death, by some definition of death rate or by transformed number of deaths. It seems that the logarithm of cases would make a good predictor, though the log transform does have the disadvantage of destabilizing lower values, as seen from the wider "base" of the scatterplot.