Covid19 - Analysis

Jiadi Huang

3/20/2020

An analysis of coronavirus in China

An analysis of the current outbreak of the coronavirus happening in China using the data from https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/data. The data is retrieved from 03/11/2019

Background

Coronaviruses are a large family of viruses that are common in people and many different species of animals, including camels, cattle, cats, and bats. Rarely, animal coronaviruses can infect people and then spread between people such as with MERS-CoV, SARS-CoV, and now with this new virus. The coronavirus was first detected in China. On January 30, 2020, the International Health Regulations Emergency Committee of the World Health Organization (WHO) declared the outbreak a "public health emergency of international concernexternal icon" (PHEIC). On January 31, Health and Human Services Secretary Alex M. Azar II declared a public health emergency (PHE) for the United States to aid the nation's healthcare community in responding to COVID-19. On March 11, WHO publiclyexternal icon characterized COVID-19 as a pandemic. On March 13, the President of the United States declared the COVID-19 outbreak a national emergency (information retrieved from CDC https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html)

Question

Does a higher avilability of medical resources result in a lower death rate caused by the coronavirus?

Hypothesis

My hypothesis is that the amount of resources that is avilable to treat coronavirus has a significant impact on the deathrate caused by the coronavirus.My hypothesis is that the amount of resources that is avilable to treat coronavirus has a significant impact on the deathrate caused by the coronavirus.

I will be comparing the percent of population infected with the total death Rate. The percent of population infected will represent the medical resources avilable because with more people infected, less medical resources will be avilable to individuals.

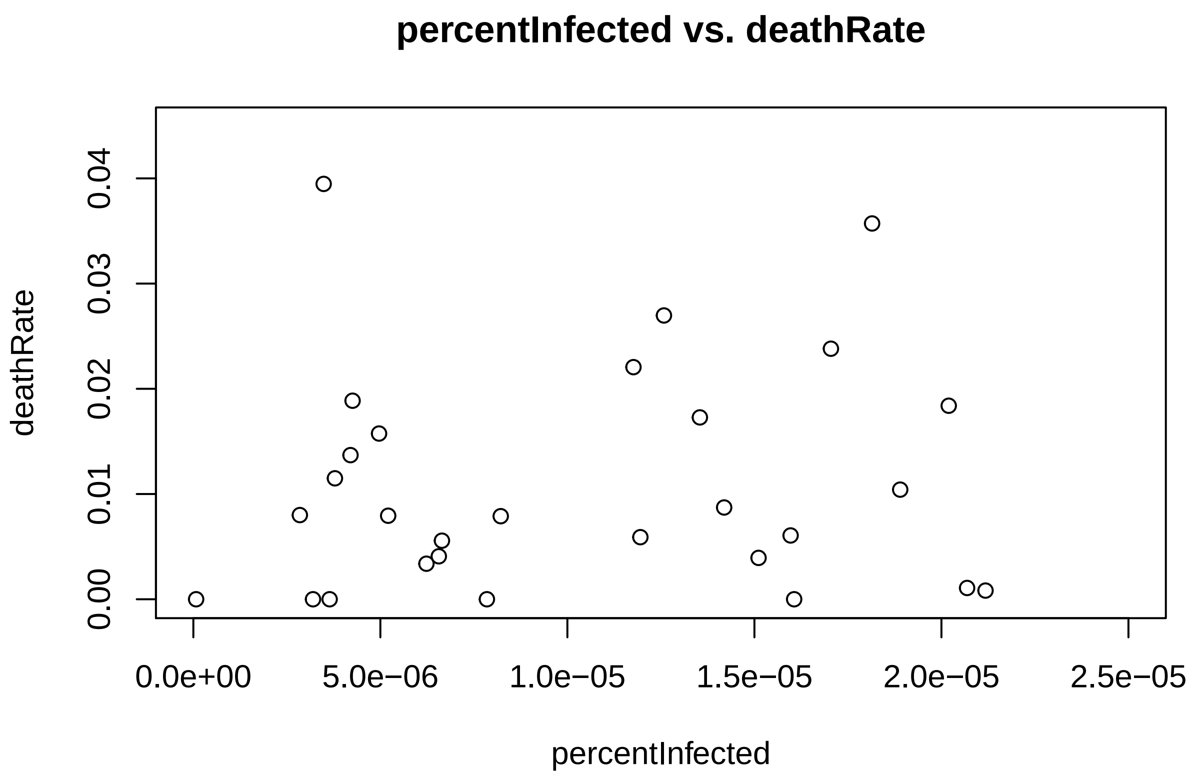
```r
library(readxl)
dat<-read_excel("covid19.xlsx")
```

```
## New names:
## * Date -> Date...2
## * Date -> Date...5
```

```
dat
```

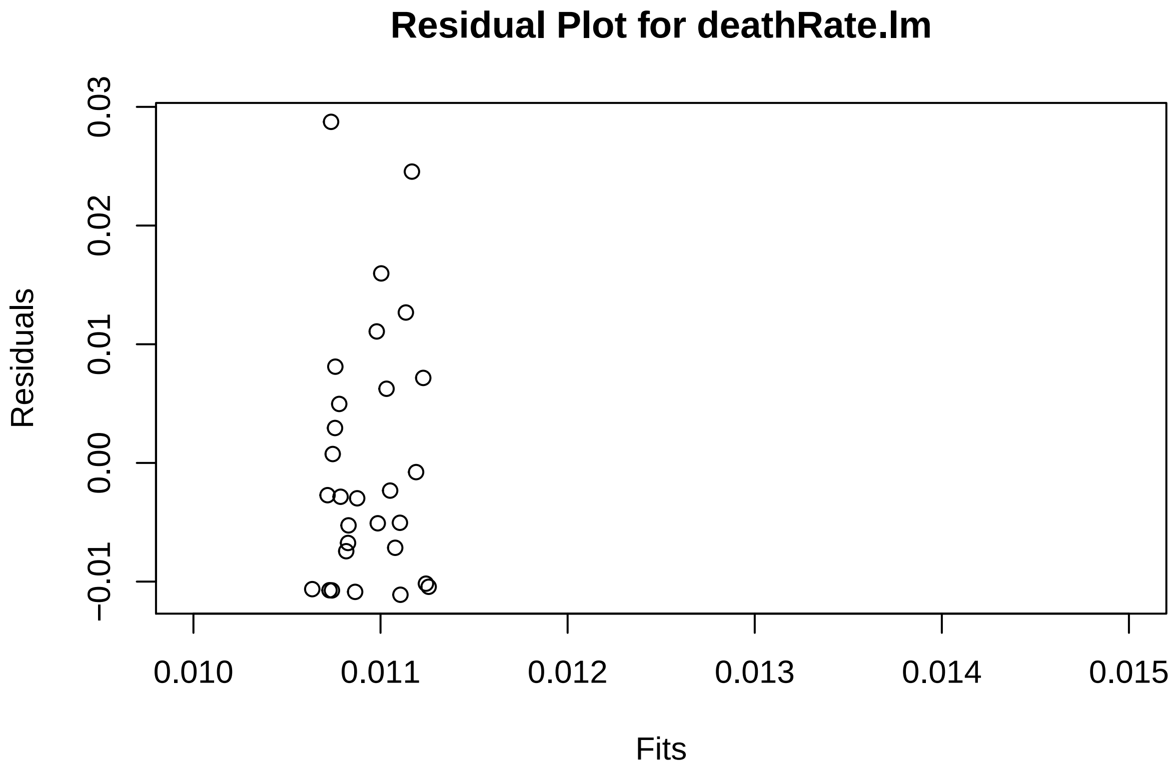
```
## # A tibble: 30 x 10
##    Number Date...2            Province Country Date...5 `Number of Infe~ Death
##     <dbl> <dttm>              <chr>    <chr>   <chr>               <dbl> <dbl>
##  1      1 2020-03-11 00:00:00 Hubei    Mainla~ 2020-03~            67773  3046
##  2      2 2020-03-11 00:00:00 Guangdo~ Mainla~ 2020-03~             1356     8
##  3      3 2020-03-11 00:00:00 Henan    Mainla~ 2020-03~             1273    22
##  4      4 2020-03-11 00:00:00 Zhejiang Mainla~ 2020-03~             1215     1
##  5      5 2020-03-11 00:00:00 Hunan    Mainla~ 2020-03~             1018     4
##  6      6 2020-03-11 00:00:00 Anhui    Mainla~ 2020-03~              990     6
##  7      7 2020-03-11 00:00:00 Jiangxi  Mainla~ 2020-03~              935     1
##  8      8 2020-03-11 00:00:00 Shandong Mainla~ 2020-03~              760     6
##  9      9 2020-03-11 00:00:00 Jiangsu  Mainla~ 2020-03~              631     0
## 10     10 2020-03-11 00:00:00 Chongqi~ Mainla~ 2020-03~              576     6
## # ... with 20 more rows, and 3 more variables: Recovered <dbl>,
## #   percentInfected <dbl>, deathRate <dbl>
```

```
percentInfected<-dat$percentInfected
lnpercentInfected = log(percentInfected)
deathRate<-dat$deathRate
lndeathRate = log(deathRate)
plot(percentInfected, deathRate, xlab="percentInfected",ylab="deathRate", main="percentInfected vs. dea
```



Shows some form of linear relationship

```r
deathRate.lm<-lm(deathRate~percentInfected, data=dat)
Fits<-fitted.values(deathRate.lm)
Resids<-residuals(deathRate.lm)
plot(Fits,Resids,xlab="Fits", ylab="Residuals", main="Residual Plot for deathRate.lm", xlim=c(0.01,0.015
```

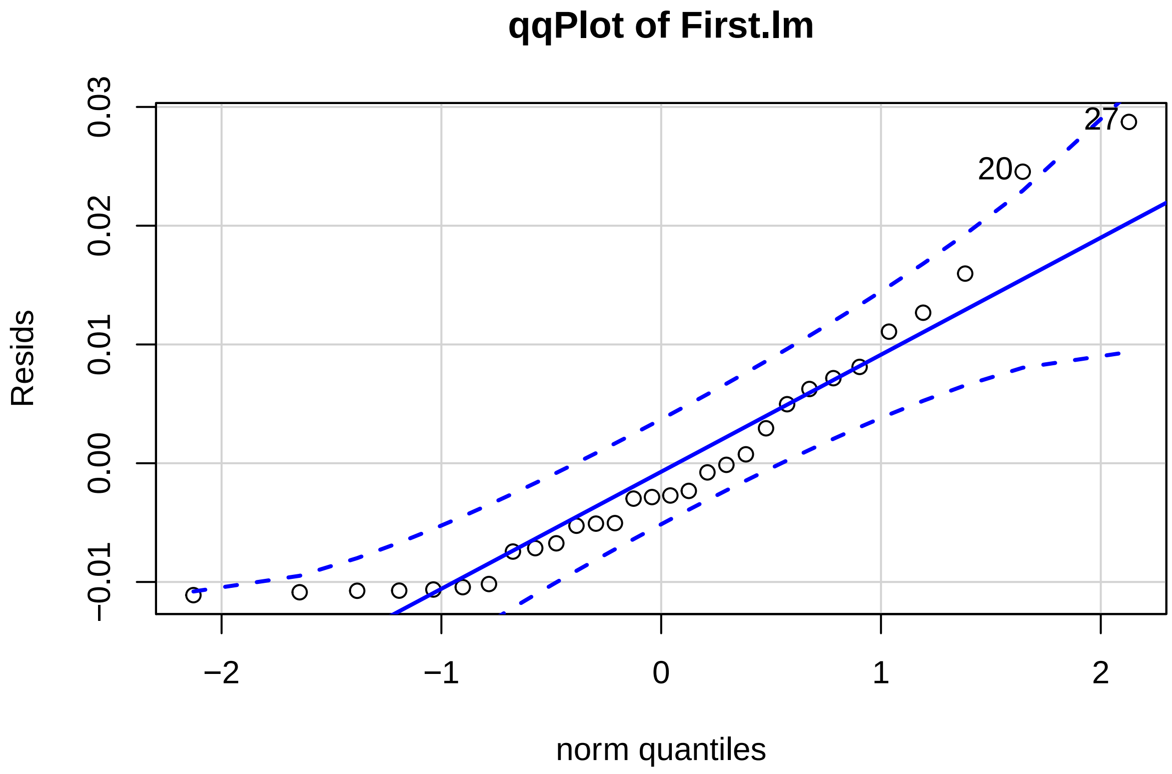


Homoscedasticity: Constant Variance is shown in the data

```
library(car)
```

```
## Loading required package: carData
```

```
qqPlot(Resids, main="qqPlot of First.lm")
```



```
## [1] 27 20
```

Normality is shown through the qqplot. Linear regression is an appropriate model to model the data.

```
summary(deathRate.lm)
```

```
##
## Call:
## lm(formula = deathRate ~ percentInfected, data = dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.011106 -0.007366 -0.002783  0.005930  0.028738
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.01063    0.00201   5.290 1.25e-05 ***
## percentInfected 29.47875    9.40635   3.134  0.00402 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01072 on 28 degrees of freedom
## Multiple R-squared:  0.2597, Adjusted R-squared:  0.2332
## F-statistic: 9.821 on 1 and 28 DF,  p-value: 0.004023
```

Conclusion

with a p value of 0.00402, the variable percentInfected is signficant in predicting the deathRate. The r-squared value is 0.2597 which memans the percent of population infected explains 25.97% of the variation of the death Rate variable. The given information shows that the percent infected has a great impact on the death rate caused by coronavirus. The data supports my claim that the amount of resources that is avilable to treat coronavirus has a significant impact on the deathrate caused by the coronavirus. More medical resources for treating the coronavirus will result in a lower death rate.