Untitled

Mustofa

4/5/2020

```
#data source"https://data.world/covid-19-data-resource-hub/covid-19-case-counts"

library(magrittr) # needs to be run every time you start R and want to use %>%
library(dplyr)    # alternatively, this also loads %>%
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- read.csv("file:///Users/mustofa/Desktop/COVID19.csv", header=TRUE)
colnames(df)
```

```
##  [1] "Case_Type"         "Cases"             "Difference"
##  [4] "Date"              "Country_Region"    "Province_State"
##  [7] "Admin2"            "Combined_Key"      "FIPS"
## [10] "Lat"               "Long"              "Table_Names"
## [13] "Prep_Flow_Runtime"
```

```
df_us_s <- subset(df, Country_Region =="US")
#df_us_s <- subset(df_us, Province_State =="New York")

df_us_s1=df_us_s %>%
    group_by(Date, Case_Type) %>%
summarise(Cum_num = sum(Cases))

library(reshape)
```

```
##
## Attaching package: 'reshape'
```

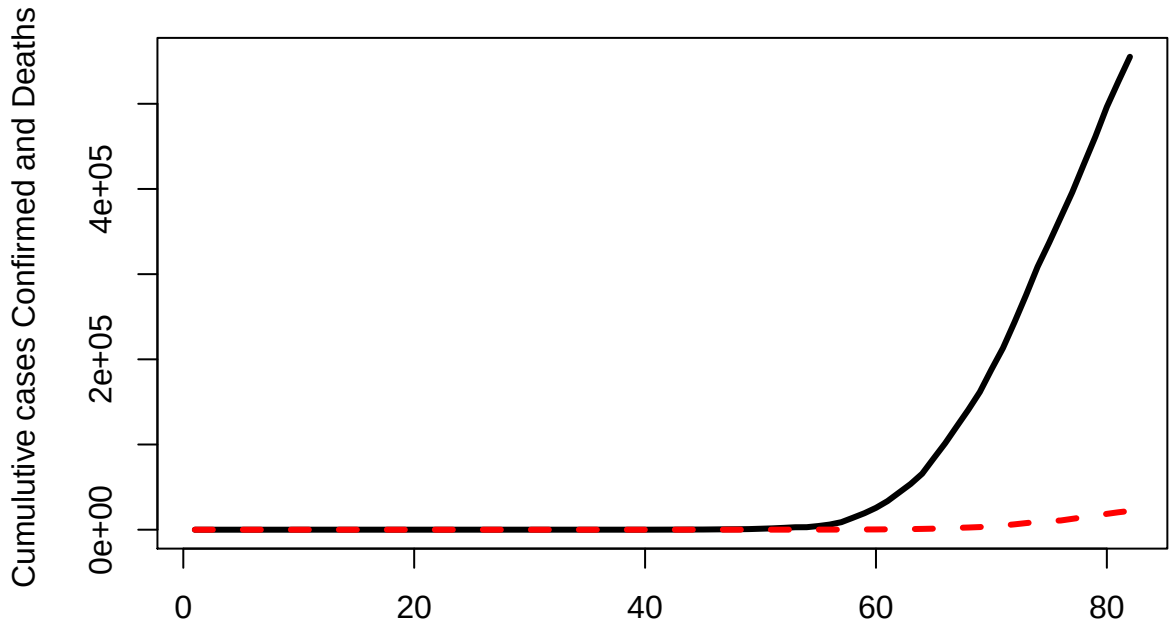
```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
df_us_s2 <- melt(df_us_s1, id=(c("Date", "Case_Type")))
df_us_s3 = cast(df_us_s2, Date~Case_Type)
df_us_s4=df_us_s3[order(as.Date(df_us_s3$Date, format="%m/%d/%Y")),]
df_us_s5=df_us_s4 %>% as_tibble() %>% mutate(New_Confirmed = Confirmed-lag(df_us_s4$Confirmed),
New_Deaths = Deaths-lag(df_us_s4$Deaths),
Lead_Deaths_7=lead(df_us_s4$Deaths, n=7L),
Lead_Deaths_14=lead(df_us_s4$Deaths, n=14L))
```

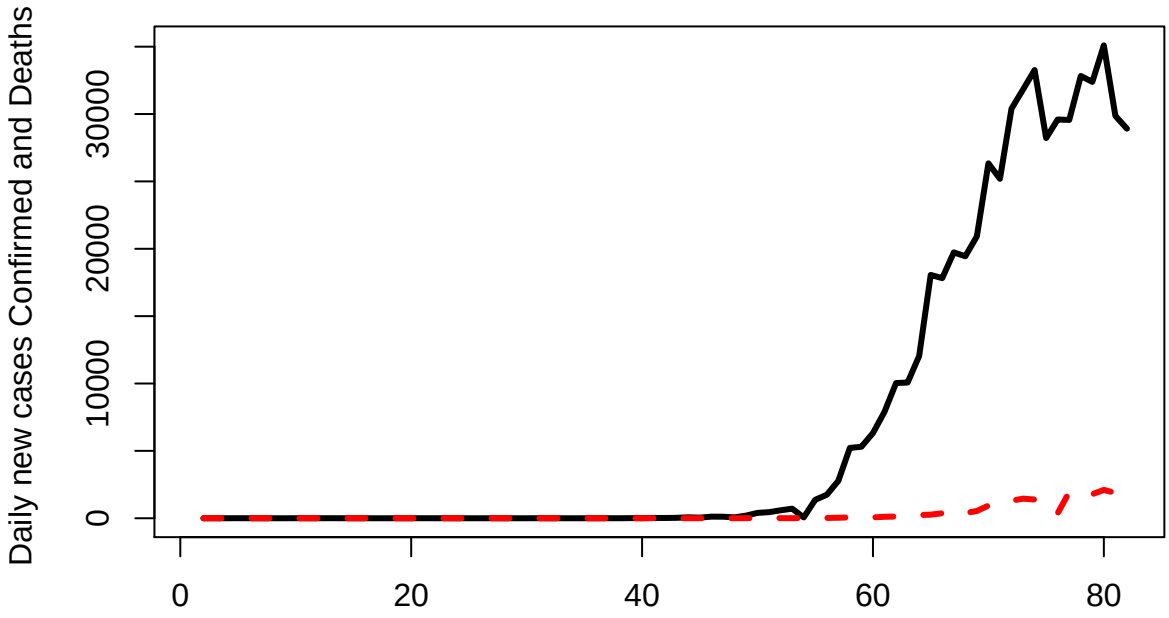
```
df_us_s6=df_us_s5 %>% as_tibble() %>% mutate(
                          Survival_rate=df_us_s5$Confirmed/(df_us_s5$Confirmed+df_us_s5$Deaths),
                            Survival_rate_7=df_us_s5$Confirmed/(df_us_s5$Confirmed+df_us_s5$Lead_Deaths
                          Survival_rate_14=df_us_s5$Confirmed/(df_us_s5$Confirmed+df_us_s5$Lead_Deaths_
                            )



colnames(df_us_s6)
```

```
##  [1] "Date"              "Confirmed"         "Deaths"
##  [4] "New_Confirmed"     "New_Deaths"        "Lead_Deaths_7"
##  [7] "Lead_Deaths_14"    "Survival_rate"     "Survival_rate_7"
## [10] "Survival_rate_14"
```

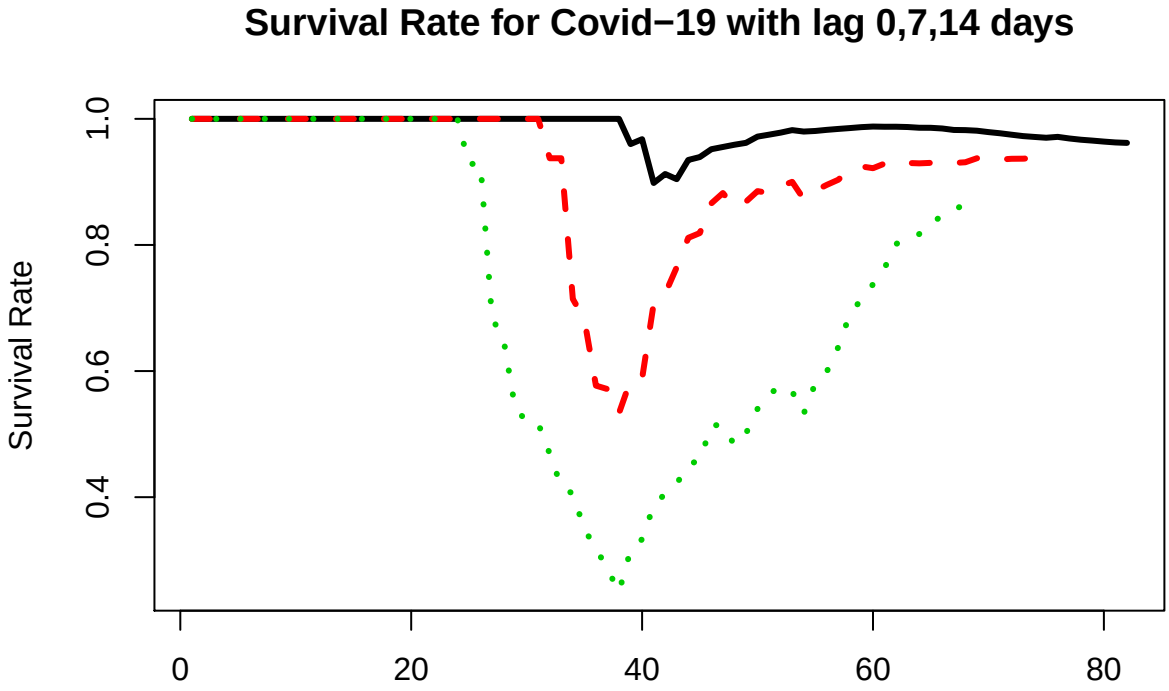
```
matplot(cbind(df_us_s6$Confirmed, df_us_s6$Deaths),type="l", lty = 1:5,pch=c(".","."), lwd=3, ylab="Cumu
```




```
matplot(cbind(df_us_s6$New_Confirmed, df_us_s6$New_Deaths),type="l", lty = 1:5,pch=NULL, lwd=3, ylab="Da
```




```
matplot(cbind(df_us_s6$Survival_rate,df_us_s6$Survival_rate_7,df_us_s6$Survival_rate_14),type="l", lty =
```



Since there is lag from the time of confirmation and death, Kaplan Meyer survival analysis does not make sense in our COVID19 data. We don't have individual level data for COVID-19 patients. To match with confirmed cases and the death toll with the patients, I used lead death data of 7 and 14 days. I also did plot without any lag in the data, but it just doesn't make sense as the patients who are confirmed in last week are still living, but may die in future. Hence, it will do gross overestimation of survival rate. I think it's better to get rid of last 7/14 days confirmed data and not use them in the current analysis. As this code allows for updating the analysis for recent data with a few clicks we can see the change in the survival rate as we will be reciving the death numbers in the coming weeks for recently confirmed cases.
The source of using 7/14 days lag is "https://www.worldometers.info/coronavirus/coronavirus-death-rate/"

"The Wang et al. February 7 study published on JAMA found that the median time from first symptom to

dyspnea was 5.0 days, to hospital admission was 7.0 days, and to ARDS was 8.0 days.[9]

Previously. the China National Health Commission reported the details of the first 17 deaths up to 24 pm 22 Jan 2020. A study of these cases found that the median days from first symptom to death were 14 (range 6-41) days, and tended to be shorter among people of 70 year old or above (11.5 [range 6-19] days) than those with ages below 70 year old (20 [range 10-41] days."

The findings are also consistent with the fact that CDC had very strict guidelines in the beginning on doing the test to only severe cases or patients who are already hospitalized. So, it's not surprising that survival rate is low. However, Testing has been scaled up later to catch up with the demand and we see sharp increase in the survival rate. Also, there are many asymptotic/mild cases which were not reported as people are asked to stay home for self recovery. If we have data regarding the actual penentration of this virus into the populaiton through Rapid antibody test we can see much increase in the survival rate.

Also, this analysis is consistent with the fact that there were underpreparedness of Hospitals with resources, lack of knowledge about therpaies to treat the patients, and degree of social distancing observed by population in the early period. Which explains into sharp drop in the survival rate and subsequent improvement.