The effects of public health interventions on covid-19 transmission using R the importance of such public health interventions in flattening the curve

SUBMITTED BY: SWATI LATHWAL
UCF ID: 5069251

Table of Content

Abstract

The project is based on the actual and current dataset for the COVID-19 pandemic.

Specifcation :

- The dataset is from the John Hopkins University Center for Systems Science and Engineering (JHU CCSE) and is available on the link: https://github.com/CSSEGISandData/COVID-19.

- Visualization created for the project are through the R programming in the R studio.

- The packages used for the project are tidyverse, Magrittr, lubridate, stringr, tibble, broom, ggplot2, ggthemes, gt, knitr, devtools, tidyverse, magrittr, lubridate, stringr, tibble, broom, ggplot2, devtools, DiagrammeR, Parallel, foreach, tictoc, EpiModel, Incidence, earlyR.

Introduction

In this project we will evaluate how contagious COVID-19 is and the effects of public health interventions on covid-19 transmission in the communities. As we are aware about the invisible enemy that's present in our surrounding almost everywhere the Novel Corona virus. The Corona virus has turned out to be the more lethal than the SARS epidemic. Through this project we will be able to figure out how contagious this virus is. For this purpose, the R language platform has been used in order to calculate its contagiousness of the COVID-19 in the most affected country Unites States. The SIR model has turned out the best fit among all of the epidemiological models around.

Body

ABOUT data

- The dataset has 3340 rows and 195 columns. Two are character variables and two numeric variables rest of the variables are integer only.

```
> dim(US.confirmed.cases)
[1] 3340  195
> str(US.confirmed.cases)
'data.frame':  3340 obs. of  195 variables:
 $ Country_Region: chr  "US" "US" "US" "US" ...
 $ Province_State: chr  "American Samoa" "Guam" "Northern Mariana Islands" "Puerto Rico" ...
 $ Lat           : num  -14.3 13.4 15.1 18.2 18.4 ...
 $ Long_         : num  -170.1 144.8 145.7 -66.8 -67.2 ...
 $ 2020-01-22    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-23    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-24    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-25    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-26    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-27    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-28    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-29    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-30    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ 2020-01-31    : int  0 0 0 0 0 0 0 0 0 0 ...
```

- There was no missing value in the dataset

Methods of analysis

Let's get acquainted with the current scenario that exist in the most affected country i.e. United states of America.

<u>Report summary</u>

```
**********************************************************************************
****************************** OVERALL SUMMARY*********************************
**********************************************************************************
   ****  Time Series  US TOTS ****
          ts-confirmed    ts-deaths        ts-recovered
          4356206         149310  1355363
                          3.43%            31.11%
   ****  Time Series  US AVGS ****
          ts-confirmed    ts-deaths        ts-recovered
          4356206         149310  1355363
                          3.43%            31.11%
   ****  Time Series  US SDS ****
          ts-confirmed    ts-deaths        ts-recovered
          NA      NA      NA
                          NA%              NA%


   * Statistical estimators computed considering 1/1/1 independent reported entries per case-type
   **********************************************************************************
```

Fig 1.1

From the image above we can visualize the percentage of population that is affected by the Corona virus pandemic in the United States.
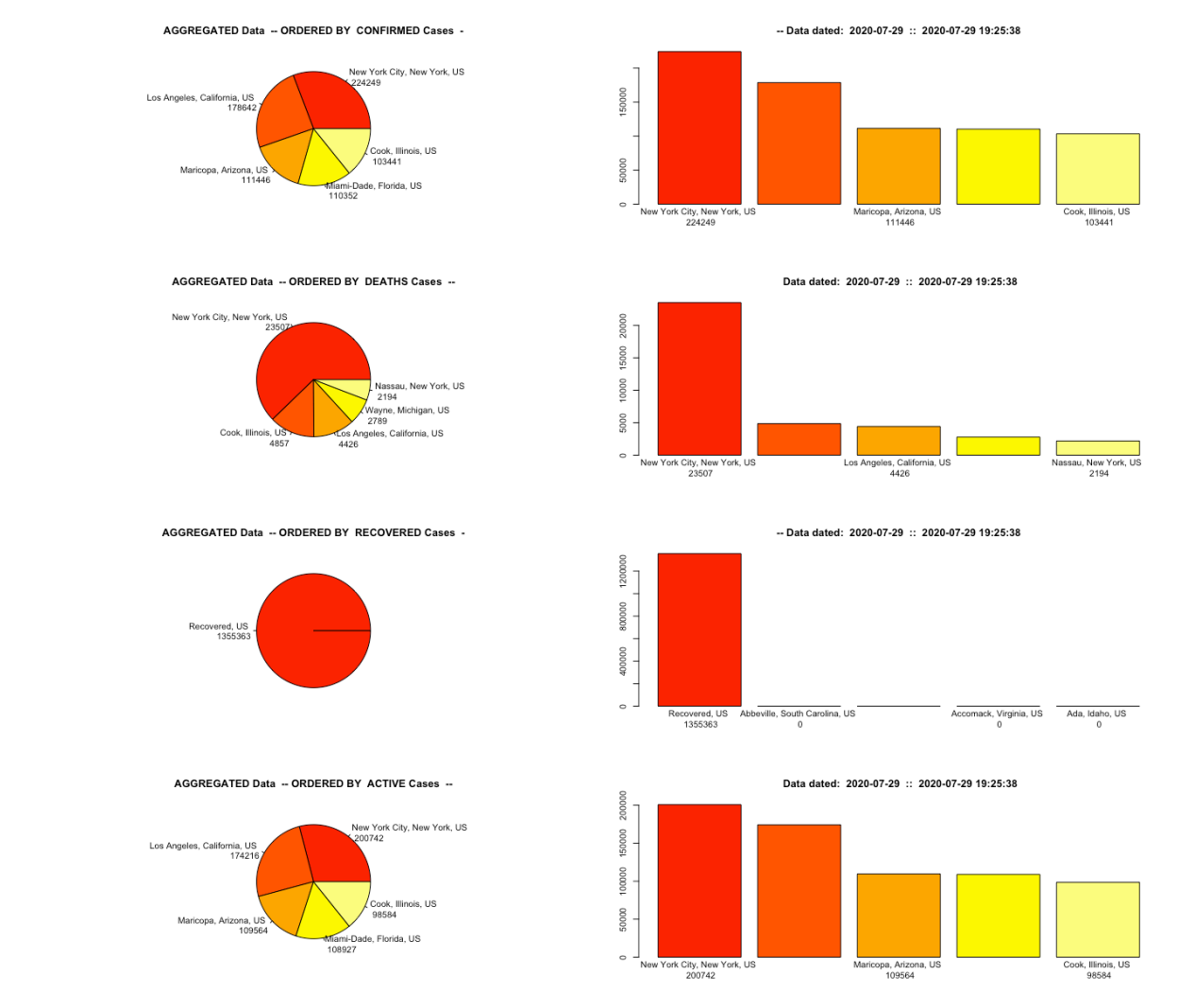
Fig1.2 The insights from each category of the cases from the USA.

The graph above depict the statistics of the most affected states in the United States i.e. New York, California, Arizona, Illinois and Florida. The data which has been taken is the recent dataset available on the John Hopkins University. The color ranges from "Red" to "Yellow" depicting as the highest affected states to the comparatively lowest affected states respectively. It can be observed that the New York has turned out to be the one where the virus was spread rapidly and affected the most.

<u>Confirmed Cases</u>
The stats in the figure depict three different fits that adjust with the data.
The exponential model has been used that uses the Linear Regression method.
The Poisson model uses the General Linear Regression method. The same methods have been used to plot the results in the later part of the topic.

```
------------------------------------------------------------------------
  Linear Regression (lm):

Call:
lm(formula = y.var ~ x.var)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3486 -2.0610 -0.2867  2.1820  3.5309

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.080877   0.330097   9.333   <2e-16 ***
x.var       0.081604   0.003013  27.083   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.26 on 187 degrees of freedom
Multiple R-squared:  0.7968,    Adjusted R-squared:  0.7958
F-statistic: 733.5 on 1 and 187 DF,  p-value: < 2.2e-16
------------------------------------------------------------------------
  GLM using Family [1] "poisson" :

Call:
glm(formula = y.var ~ x.var, family = family)

Deviance Residuals:
   Min     1Q  Median      3Q     Max
-662.5 -453.0  -144.0   279.8   482.9

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.113e+01  2.695e-04   41275   <2e-16 ***
x.var       2.303e-02  1.759e-06   13095   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 277048078  on 188  degrees of freedom
Residual deviance:  29760901  on 187  degrees of freedom
AIC: 29763296

Number of Fisher Scoring iterations: 5
------------------------------------------------------------------------
  GLM using Family Family: Gamma Link function: log  :

Call:
glm(formula = y.var ~ x.var, family = family)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.5171 -2.2963 -1.0519  0.5515  1.8111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.067800   0.188944   37.41   <2e-16 ***
x.var       0.058197   0.001725   33.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 1.67347)

    Null deviance: 1210.50  on 188  degrees of freedom
Residual deviance:  677.65  on 187  degrees of freedom
AIC: 4962.1

Number of Fisher Scoring iterations: 25
------------------------------------------------------------------------
```

The image below depicts the current scenario of the number of "confirmed cases" in the USA. The insight from the plot depicts the number of cases as the time function for the USA in two different plots.
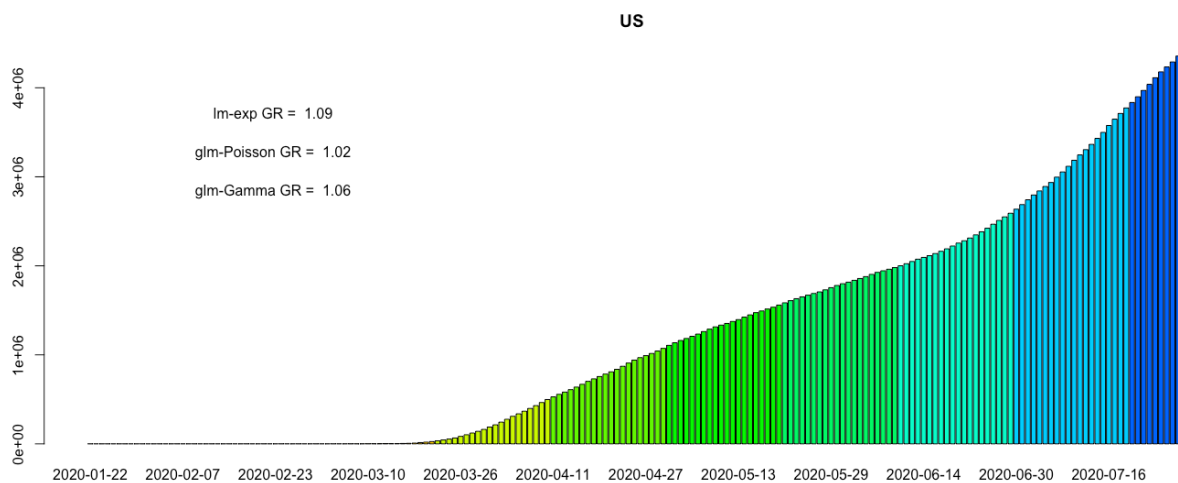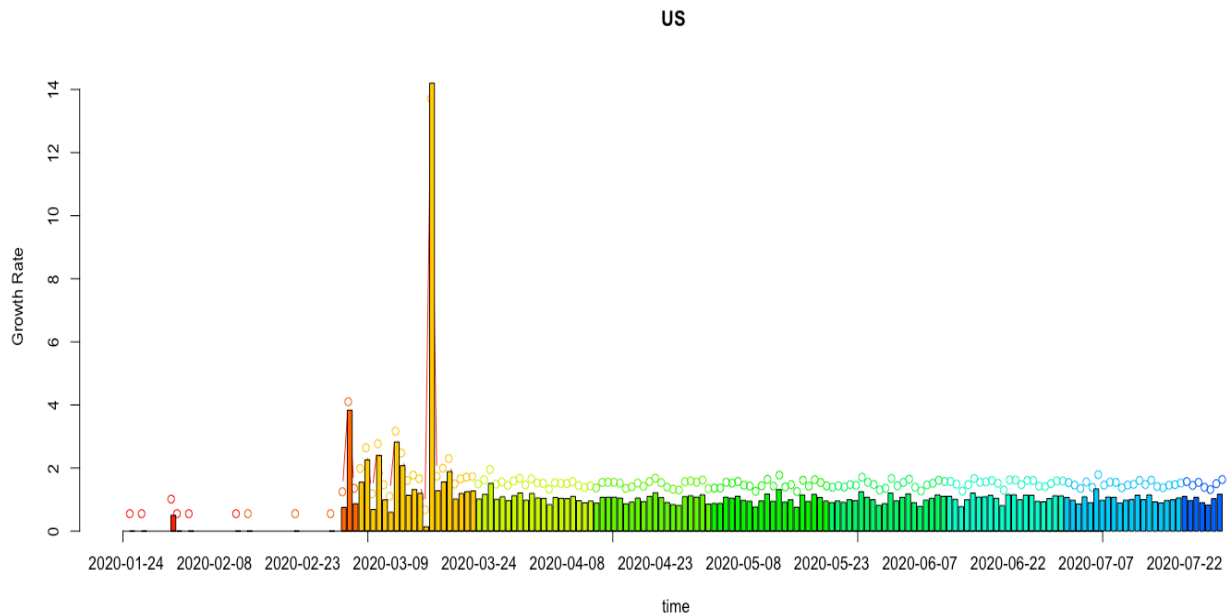

Figure 2.2 confirmed cases spread in the linear scale bar plot.

Growth Rate

Growth Rate is the changes that are occurring between the period of two consecutive dates. The plot below depicts the monthly growth rate occurring in the US. It is clearly observed that the maximum growth rate has occurred on March 24,2020. After March 2020


The plot below depicts the number of changes that took place from the period of Feb 2020 till the end of July 2020.
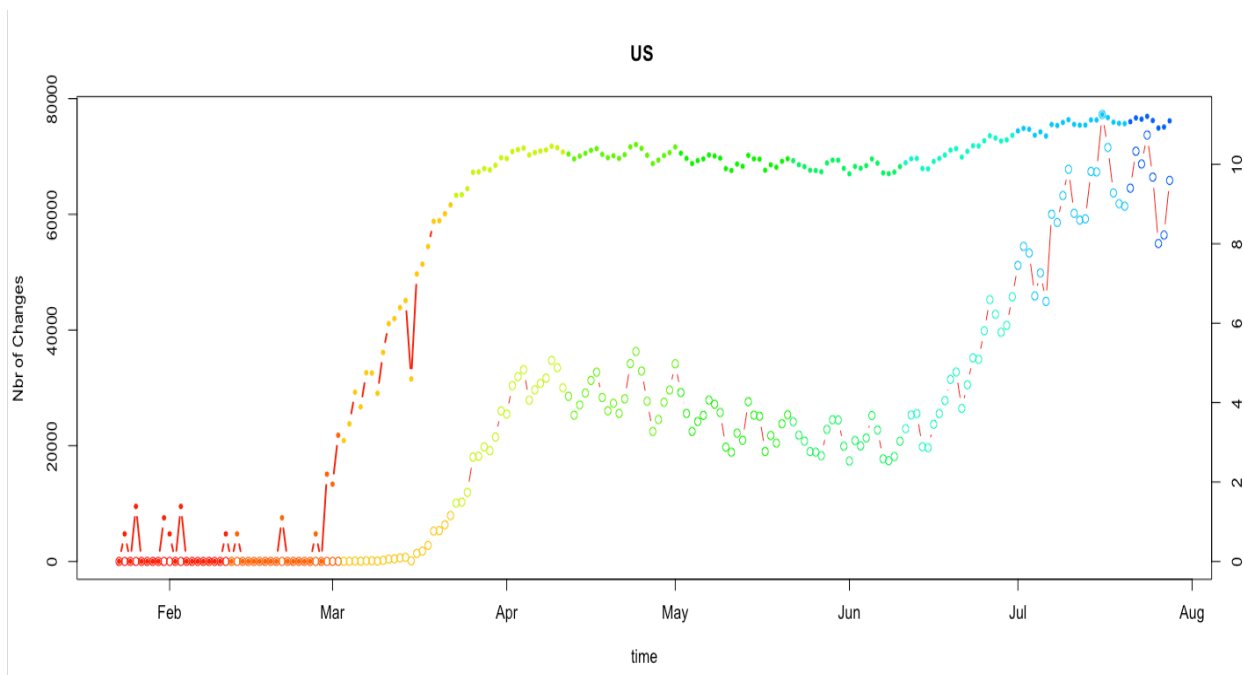
SIR modelling

The line chart below is the comparison of the global Covid cases with the US cases. The chart is static and interactive both. It is observed that the types of cases are divided into four categories i.e. the number of "confirmed", "recovered", "deaths" and "active cases" . The confirmed cases have crossed the four million bar by the end of July 2020. 2.9 million cases are still active. The cases recovered are very less as compared to the confirmed cases. The chart also depicts the number of deaths that happened in US so far. However, the number of death count cannot be evaluated properly through this chart. SIR model is the best fit to depict the fatality rate in the US.
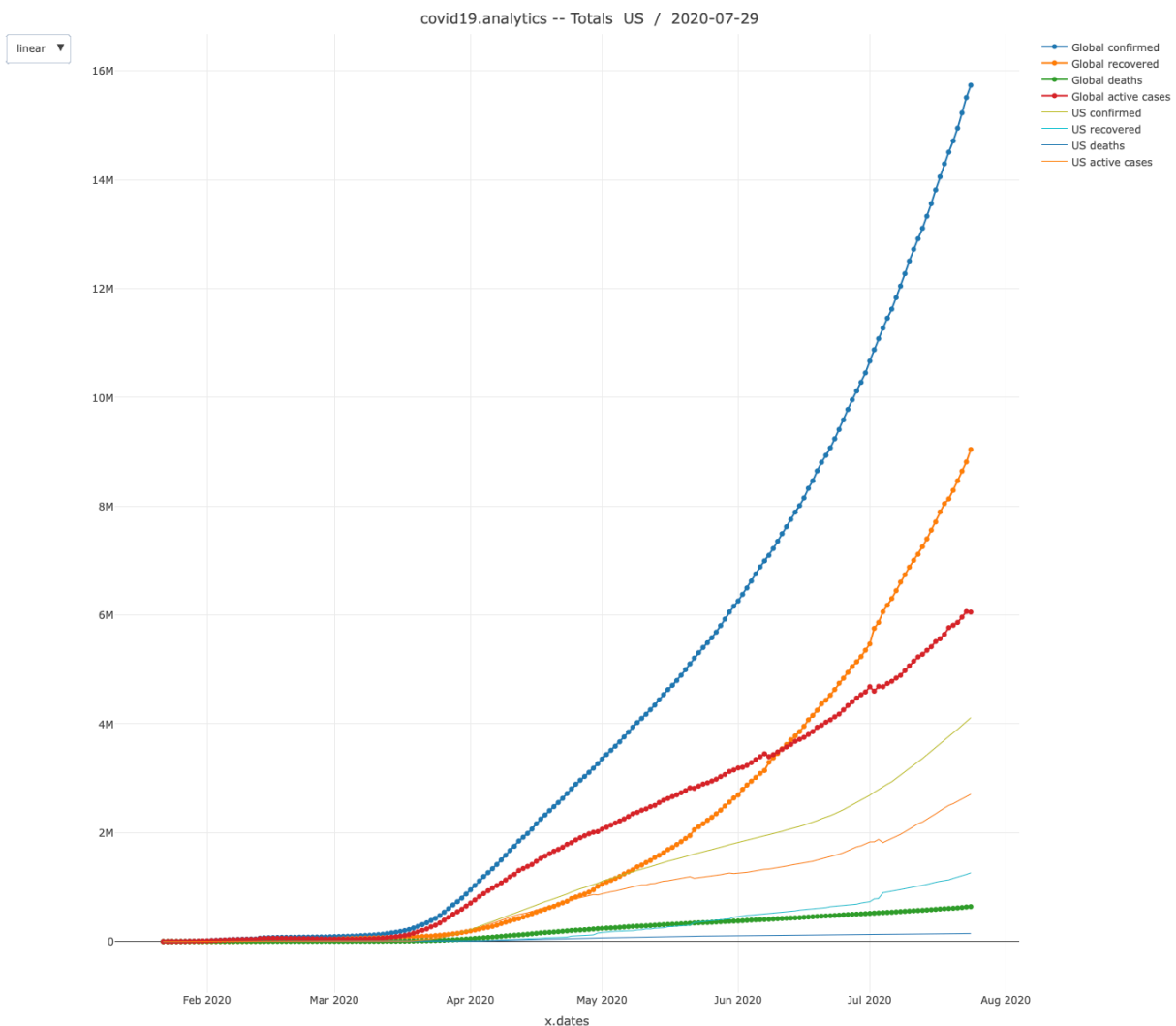
The statistics below predict all the information that is needed. SIR model depicts this information on the basis of R_0 which is the estimation number. R_0 will be discussed in detail in the later part of the report. R_0 uses the branching process for the estimation that requires a serial interval distribution from the onset of the Epidemic.

```
####################################################################################
[1] "US"
Processing...  US
 [1]  1  1  2  2  5  5  5  5  5  7  8  8 11 11 11
----------------------- Parameters used to create model -----------------------
        Region: US
        Time interval to consider: t0=1 - t1=15 ; tfinal=90
                t0: 2020-01-23 -- t1: 2020-02-06
        Number of days considered for initial guess: 15
        Fatality rate: 0.02
        Population of the region: 7.8e+09
-------------------------------------------------------------------------------
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
     beta      gamma
0.5931729 0.4068271
  R0 = 1.45804644580241
  Max nbr of infected: 15622122.05  ( 0.2 %)
  Max nbr of casualties, assuming  2% fatality rate: 312442.44
  Max reached at day : 90 ==>  2020-04-22
====================================================================================
```

The time-intervals are monthly each month to month period has 15 intervals. The total duration from January to July makes a total of 90 such intervals. The fatality rate through SIR is depicted as 0.02 percent.

The SIR model is a compartmental model for the COVID-19 time series data for US. SIR modelling is simply understanding the rate of transition between three categories of the contagious disease's outbreaks i.e. the healthy people who are likely to get infected are Susceptible (Ss) to the COVID-19, the Infectious (Ii) and the section of people who have Recovered (Rr).
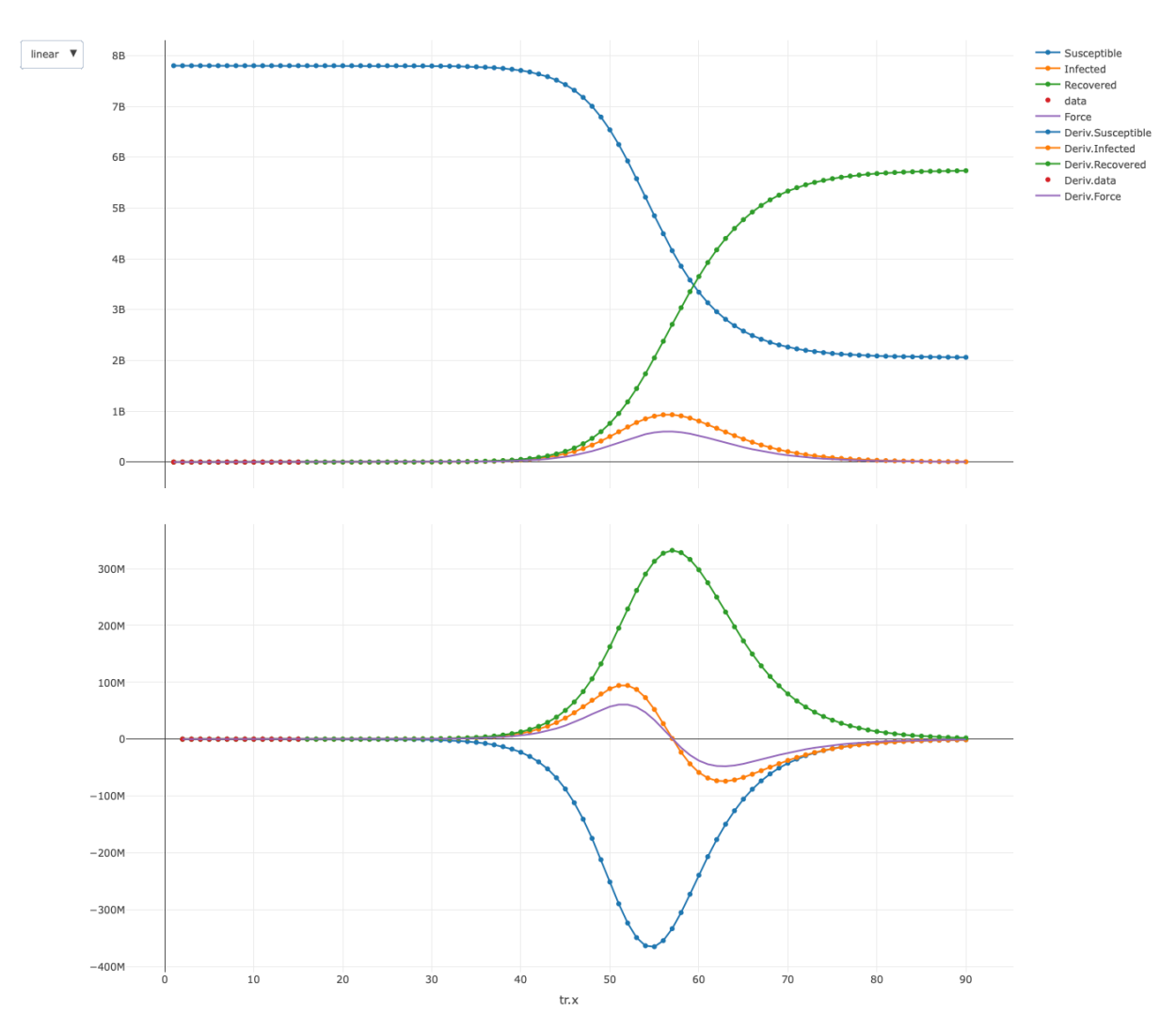
The population is broken into 3 categories those who are susceptible to the disease, those who are infectious and those who have recovered from the infection.

The outbreak dynamics are modelled into the following three differential equations:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

The equations mentioned above provides the insight about the rate of change from one group to the another. $\beta\beta$ is the coefficient the controls the flow between the Susceptible (Ss) and the Infected (Ii). The coefficient $\gamma\gamma$ controls the flow occurring between II and RR.

The calculus and linear algebra are used to discover the parallel result to the differential equations in the certain time intervals. The two main methods like residual sum of square (RSS) and MLE i.e., maximum likelihood can be further used to process the values of the beta function and gamma function. Various extensions can be made by including some more categories like SEIR model. SEIR model has four categories to introduce such categories are "Susceptible", "Exposed", "Infectious", "Recovered". Addition to the compartments will result in better output and invites more complexity. Such models generally are not precise to simulate the coughs or touch, but they can be advanced enough to model the realistic scenarios. These types of models are typically a good choice to inspect the consequence of various kinds of interventions through the what-if analysis.

The ICM SIR model

The simple (stochastic, discrete-time, individual contact models) ICM models are not very quick but are fast enough for the type of illustration that is expected from this model.

SIR (susceptible->infected->recovered) are by default models.

It is very important to give importance to the granularity for the time dimensions in the ICM models since they themselves are discrete-time models. In the ICM model the whole day is taken as the step size. 100 such time steps are simulated. After each simulation are run the average of each simulations are calculated as the final result.

It is also very important to specify the number of individuals from each category during the initial simulations. The total population in the current SIR ICM model were as follows:

Out of the total population of 1000 individuals the 997 were susceptible, 3 of them were infected and zero in the recovered category.

The package Epimodel in R provides an easy way out for default plotting methods. The plot by default the number of people in each S, R and I categories at every day in the simulation below:
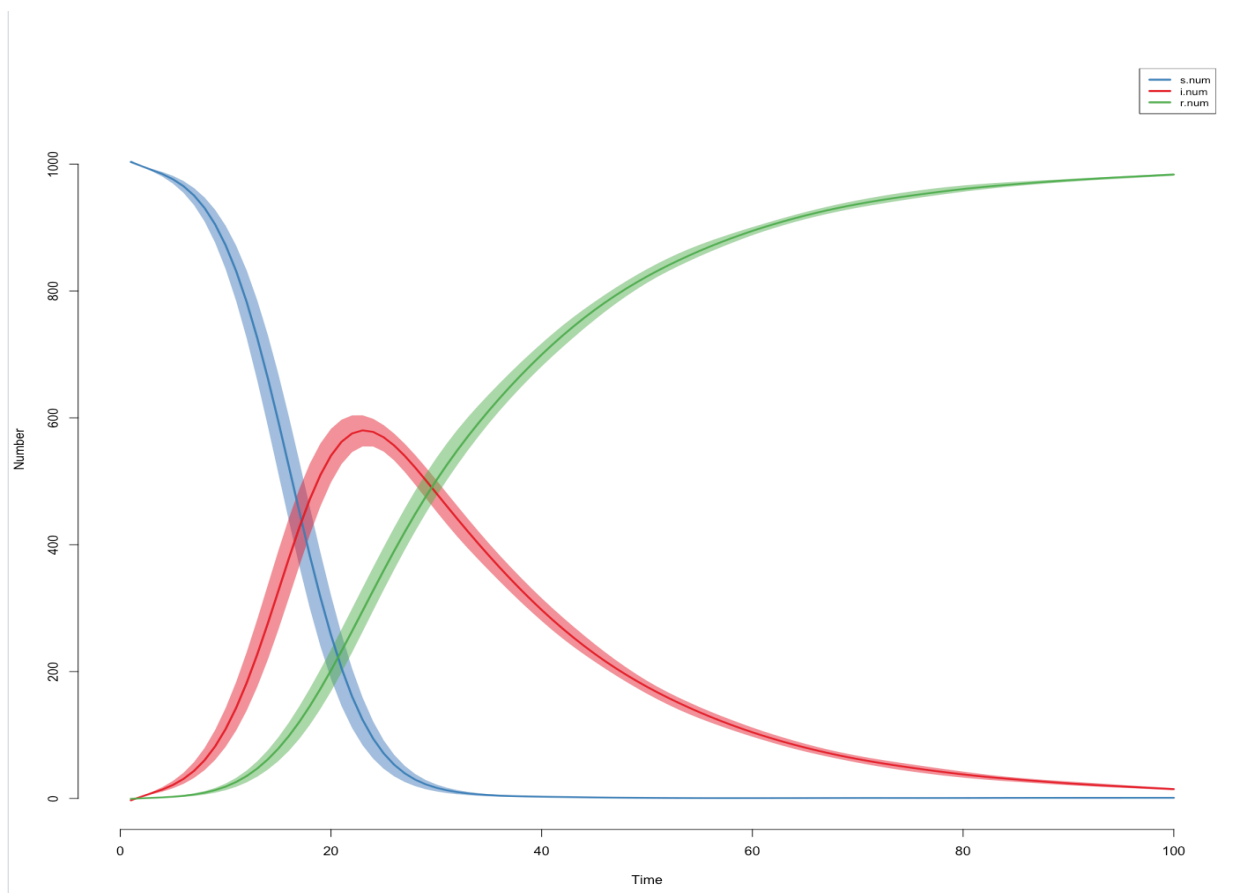
It is very important to mention the rate of transition from Susceptible to infections and infectious to recovered. After calculating all the transition rates the mean probability of the result is evaluated for each person in the respective categories.

Let's say the exposure to infection rate is 10 times each day, the probability of infection is 0.05 from each exposure of the individual. The higher risky exposes are considered as those who are exposed to the coughs and sneezes. The lower exposure risks are those where the individual touches the infected place. The mean recovery time is set as 20 days and median for the same is 14 days. The estimated mean and median are calculated from the random draws from the binomial distribution where the mean is set to the recovery rate.

Let's say the death rate in the US is 0.07 percent per year. The daily death rate in the susceptible and the recovered will be 7 divided by 365000. But here the death rate is set as the double of the evaluation.

The count of new cases each day is as follows:
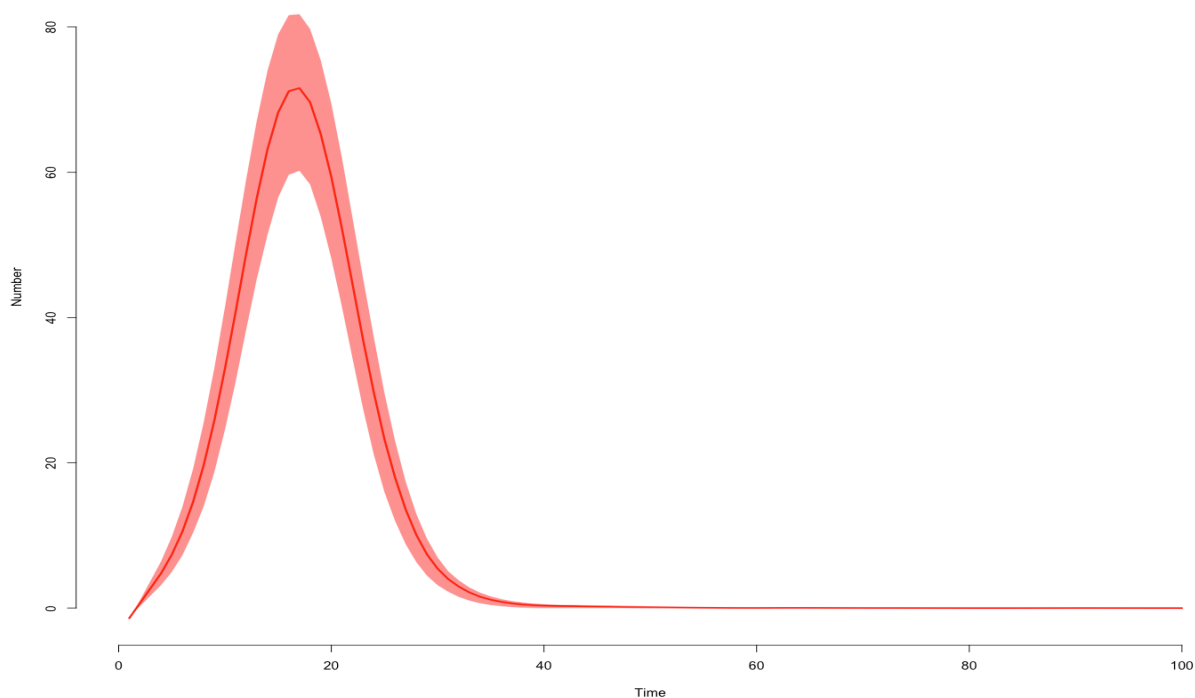
From the figure above it is very clear that the number of cases has plummeted in the initial two weeks. To check if the result is real, we will calculate the Reproduction number, i.e. $R0R0$ for the spread of pandemic curve. In this context the parameters like serial interval distributions are largely quoted in the case of COVID-19.

R_0is the reproductive number that quantify the rate of this spread.

The average for this count across everyone who has been sick is known as the effective reproductive number or R. More commonly discussed number is R_0, which is the value of R in a fully susceptible population, like at the very beginning. We have noticed the label on the simulations, where the way it's calculate is to look at each individual who is currently infectious, count how many they have infected so far. Estimate how many there going to infect in total based on the duration of the illness and average those numbers.

This factor has huge effect on the growth rate.

- When the infection rate is chopped to half, it hovered around the 1.3 to 1.7 range.
- While R is greater than 1, the infection is growing exponentially, and is known as an epidemic when the whole steady around t it called endemic.
- Less than 1 means it's on the decline.

For comparison, R_0 for covid-19 is estimated to be a little above 2, which is also around the mean estimate for the R_0 Spanish flue. In traveling case, as soon as the social distancing is turned on and travel is shut down. The R_0 quickly drops down from 2.
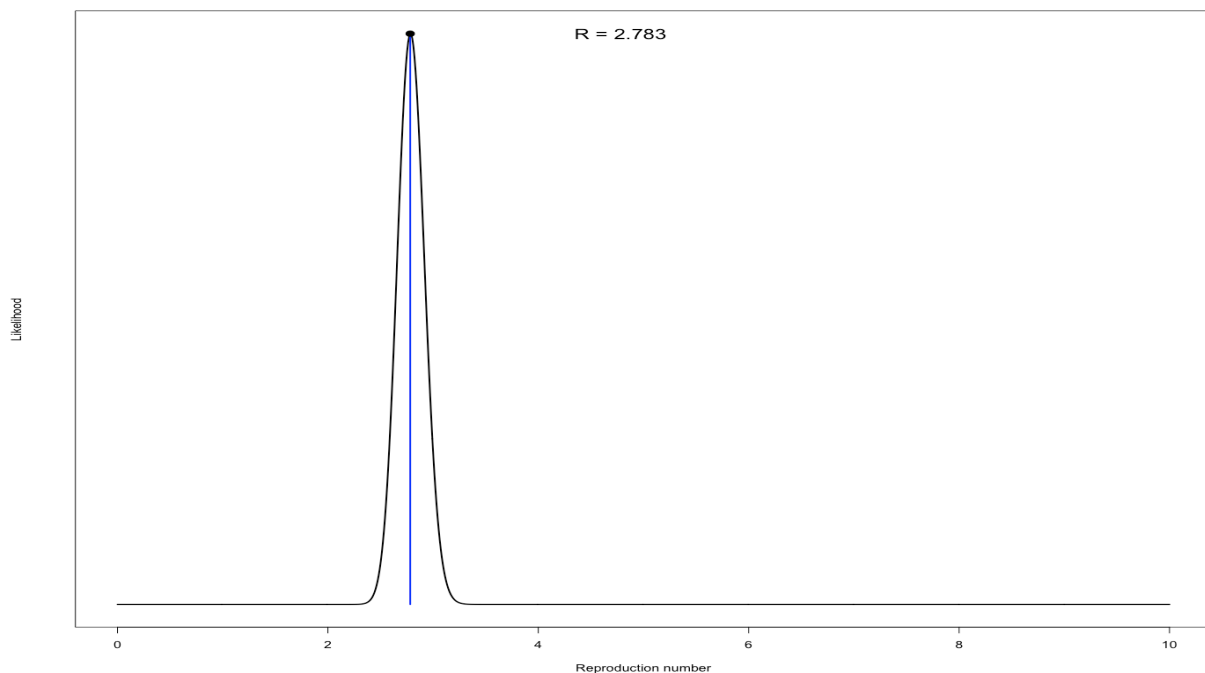
The MLE of R0R0 2.7 is very well orderly with WHO evaluation of the COVID-19 Reproduction Number.

Now we can start to investigate the effects of the public health interventions.

Avoiding the chances of infection by practicing social distancing.
The parameter here, which is how much people try to avoid each other. Let's call it the "social distancing factor". The repulsive force is applied between people.

Initially, everybody meanders about the city and the infection follows the rules that are laid out. After not too long before almost everyone gets infected. So people who are new don't play into it. The way it is written for every unit of time a susceptible person spends within a certain a "infection radius" of someone with the disease, they'll have some probability of contracting themselves. So there is the use of physical proximity as a stand-in for things like shaking hands, touching same surface, kissing, sneezing on each other, all that good stuff. The for each Infectious person, after a given period of time, they will recover and no longer be able to spread the disease or if they die, they also would not be able to spread it anymore, so as a more generic term R-removed. These are toy models with tiny population, inevitably falling far short of the complexities in real people.

What if the radius of infection is doubled. Representing more total interactions between people or more socially engaged society. It will spread more quickly but how much? It's actually very drastic. With a short time span the majority of the population is infected simultaneously. Each unit of time a susceptible person is within the infection radius of an infectious person.
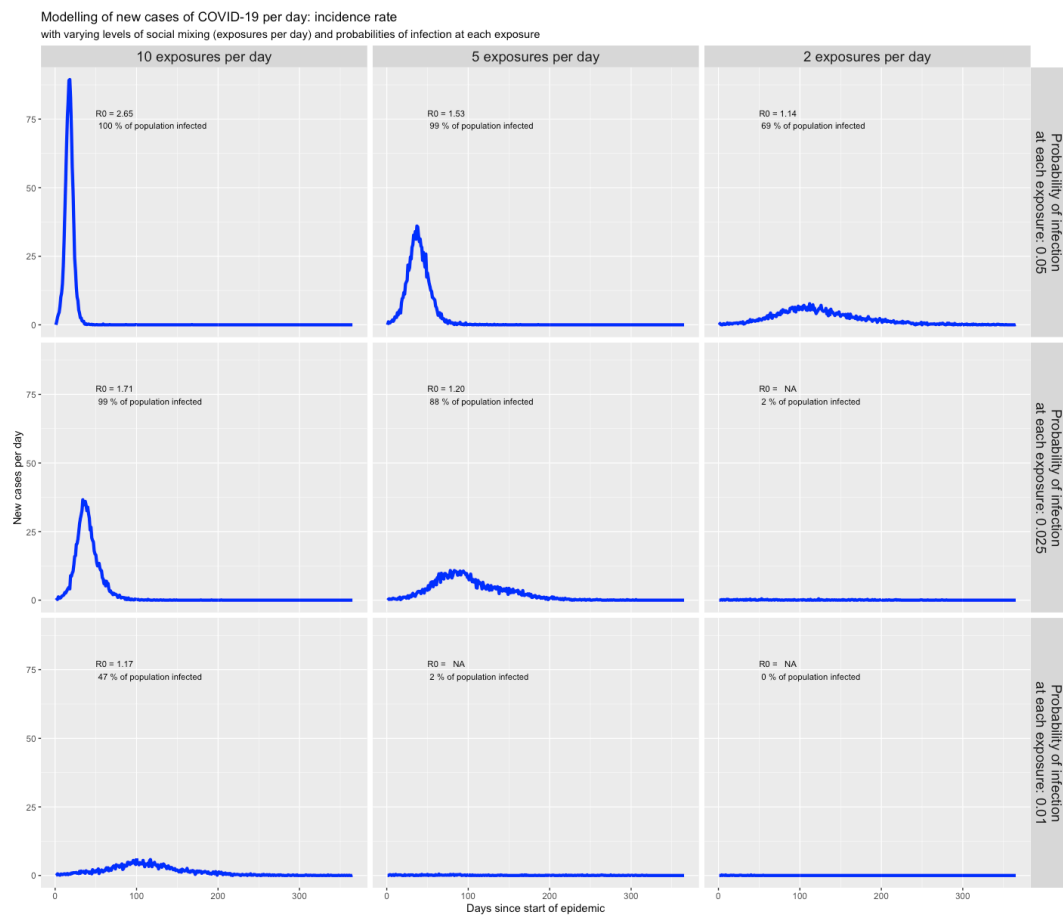
Hygiene measures:
Think of hand washing, better cough protection and less face touching. It spreads out the curve. Infact it does so by quite a lot, really illustrating how changes to hygiene can have very large effects on rate of spreading. The key takeaway to tuck away is just how sensitive the growth is to each parameter in our control.
Parameters are:
1. Daily interactions.
2. Probability of infection.
3. Duration of illness.

It's not that hard to change the daily habits in ways that multiply the number of people we interact with or that cut the probability of catching an infection in half, but the implications for the pace of this spread are huge The proportion is not constant. If we get to a point where the peak of this infection curve is too high, meaning there is a time when many people are sick at once, that's when available healthcare resources are overwhelmed which are bad and increase the mortality rate.
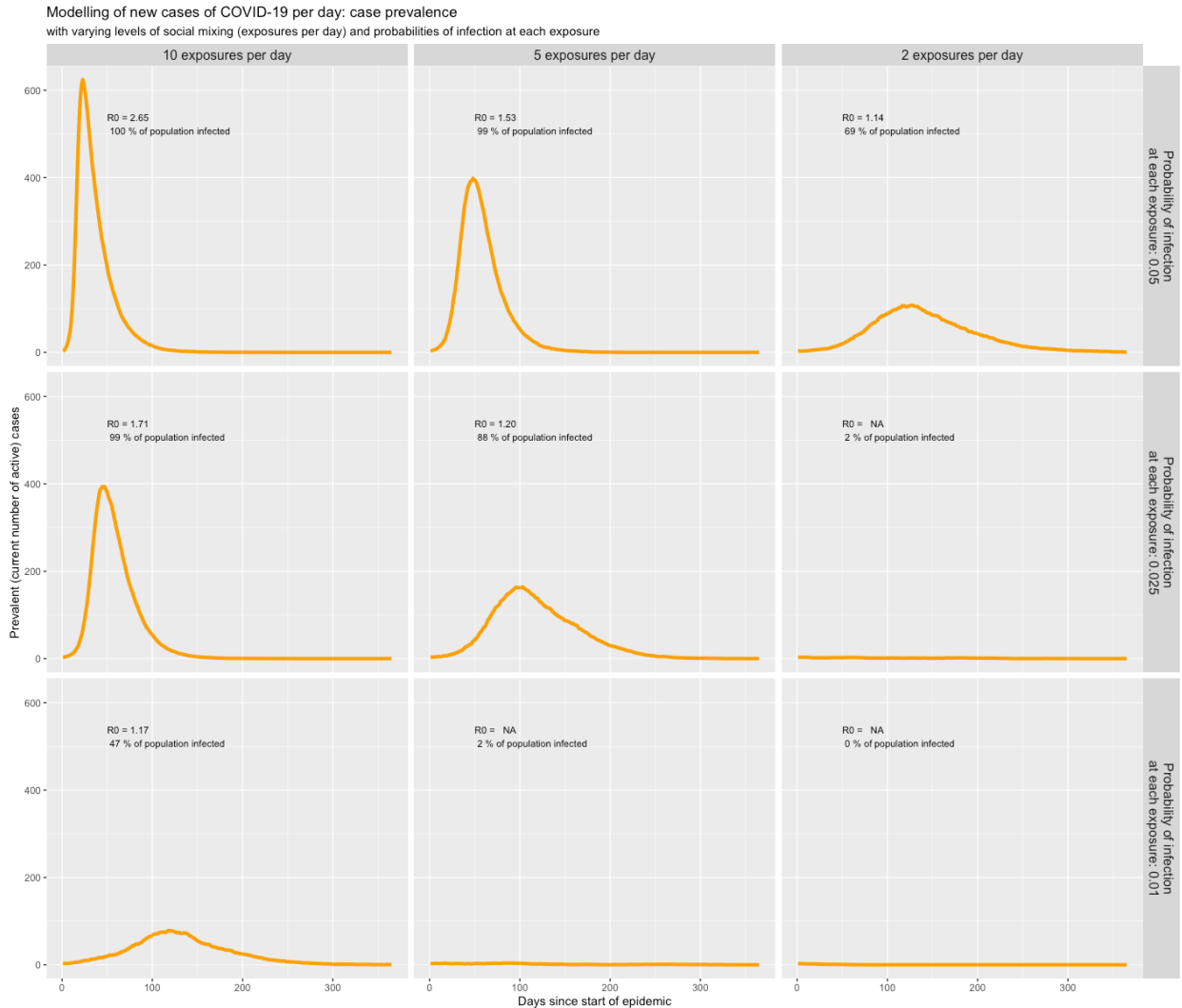
Exploring the Public Interventions using the ICM SIR model

Modelling of new cases of COVID-19 per day: incidence rate
with varying levels of social mixing (exposures per day) and probabilities of infection at each exposure

The results can be visualized in the graphs above. Let's have look at the everyday incidence i.e. the number of new cases per day. It can be observed that the epidemic curve can be flattened if the following steps are done:
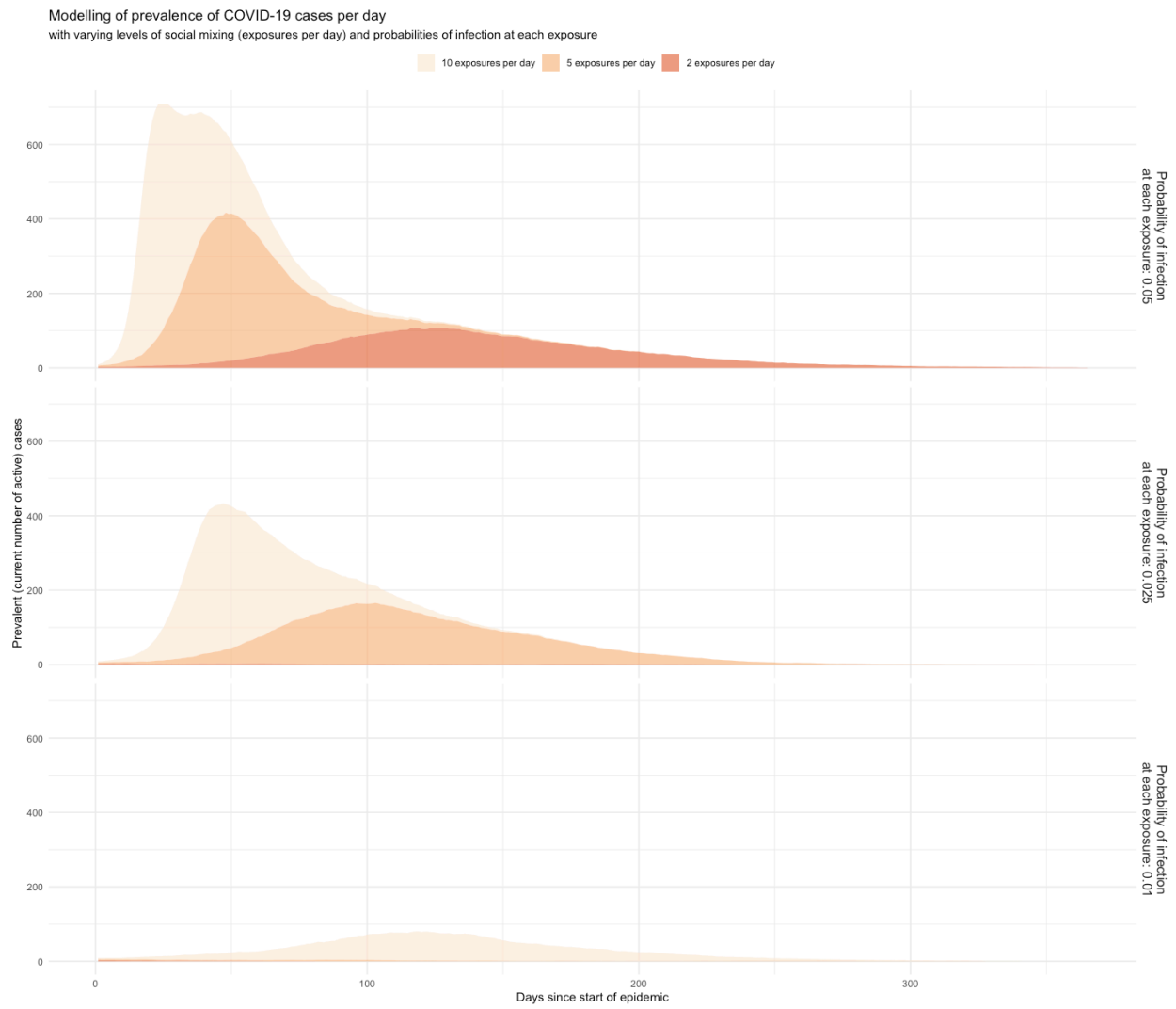
- The average exposures per day to infected individuals are decreased from left to the right.
- Risk of getting the infection at each exposure is decreased from top-to-bottom.
- It can be observed that not only it stretches the time of infection spread but also reduces the number of infections in the population.

- This is how the fatality rate is decreased. The visualization above concludes that social distancing, self-isolation hand washing etc. matter so much to flatten the rising epidemic curve.

Let us observe the pattern by looking at the number of people who are infected with COVID-19 at point of time:

Modelling of new cases of COVID-19 per day: case prevalence
with varying levels of social mixing (exposures per day) and probabilities of infection at each exposure
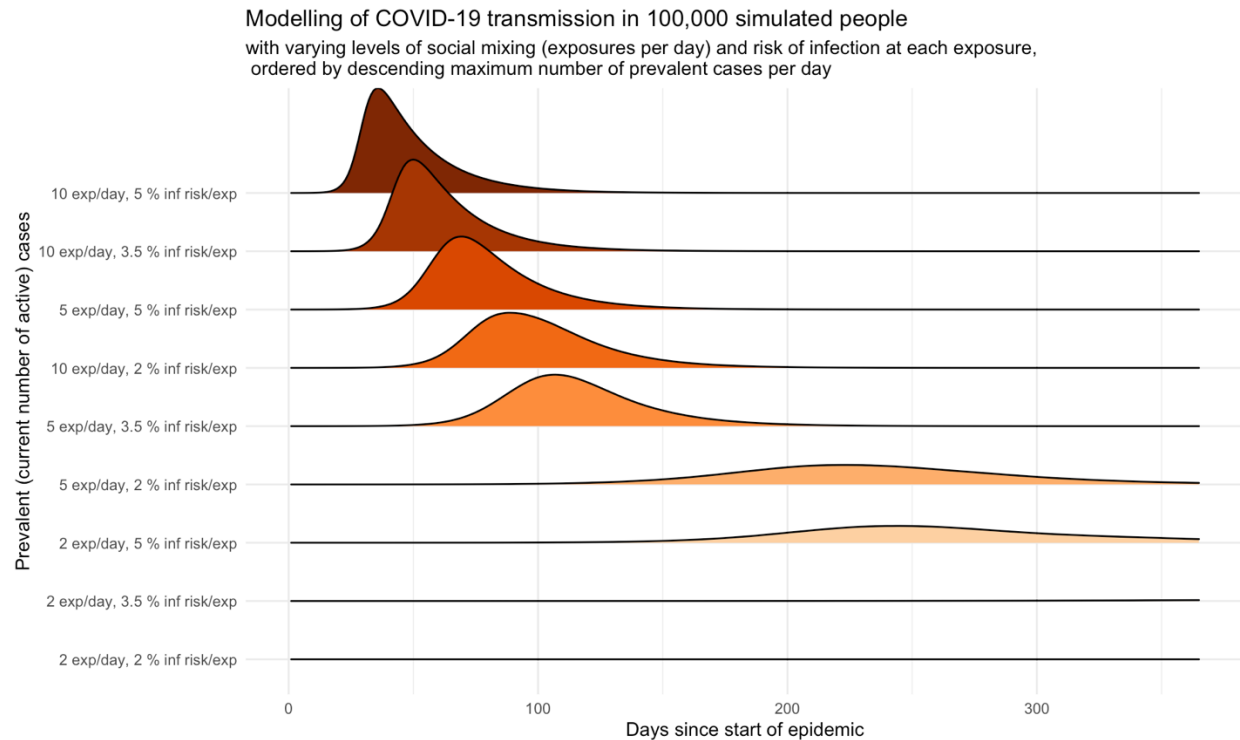
The effect when the little agents are social distancing, they are often end up trembling near the edge of the box. No isolation is perfect though so every now and then even those repulsed by each other will jiggle close enough to get infected. The point is that it's much rarer.

The visualization by putting the incidence curves on a single panel:

Modelling of prevalence of COVID-19 cases per day
with varying levels of social mixing (exposures per day) and probabilities of infection at each exposure

10 exposures per day   5 exposures per day   2 exposures per day

So far, the visualization above have turn out to be the most informative and effective. The different combinations of the levels of practicing social-distancing and hygiene precautions by the maximum number of the infected people at each point of time. Eventually this is exactly the intention where we intend to decrease to the level where the health services can cope.

Modelling of COVID-19 transmission in 100,000 simulated people
with varying levels of social mixing (exposures per day) and risk of infection at each exposure, ordered by descending maximum number of prevalent cases per day

The figure above depicts the modelling of COVID-19 transmission in 1000 simulated people, with varying levels of social mixing i.e. exposure per day.

SIR model with extended compartments

From the results above we come to know that the ICM are easy, simple and very useful for the micro-simulations of each person in the population. The model allows much more flexibility which will be outlined next.

| Compartment	Functional definition
S	Susceptible individuals
E	Exposed **and** infected, not yet symptomatic but potentially infectious
I	Infected, symptomatic **and** infectious
Q	Infectious, but (self–)isolated
H	Requiring hospitalisation (would normally be hospitalised if capacity available)
R	Recovered, immune from further infection
F	Case fatality (death due to COVID–19, not other causes)

Visualization of the extended model:

Prevalence: Prevalence is the number of people in each category on each day

Baseline simulation
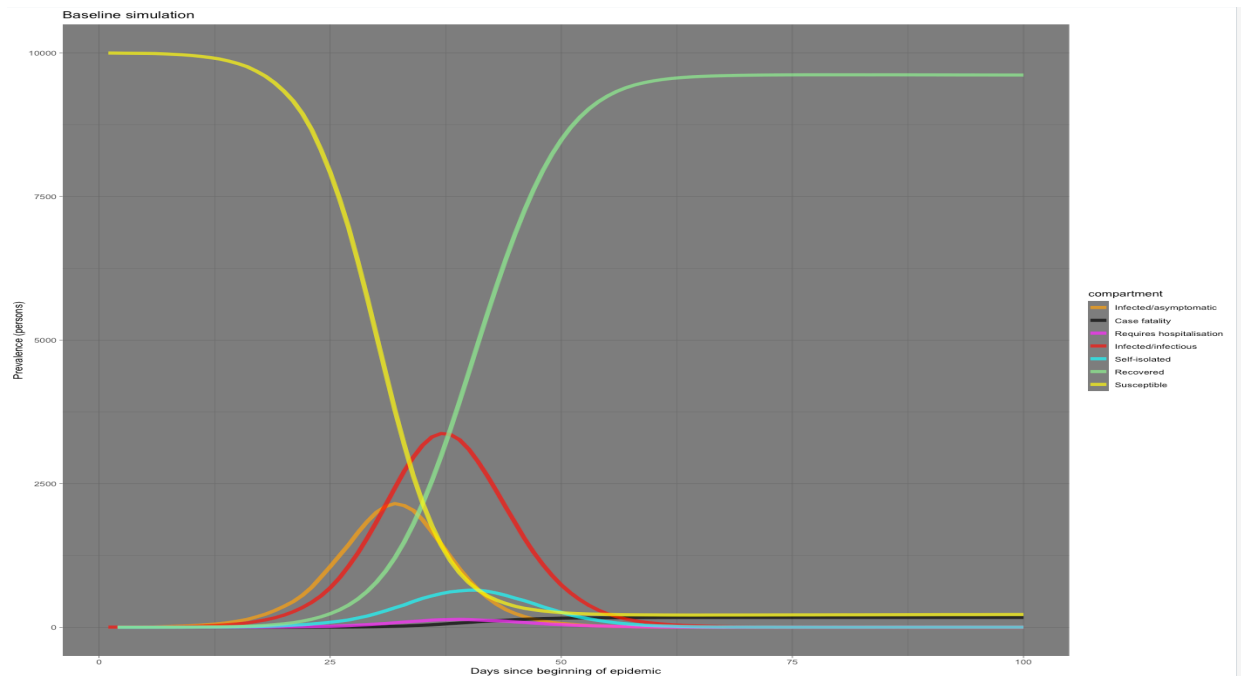
From the figure above It can be predicted that the entire population is infected eventually. It can be observed that the Susceptible (Ss) and the Recovered (Rr) compartments are the one which rules the plot.

Let's remove those compartments for once and visualize the result:

Baseline simulation
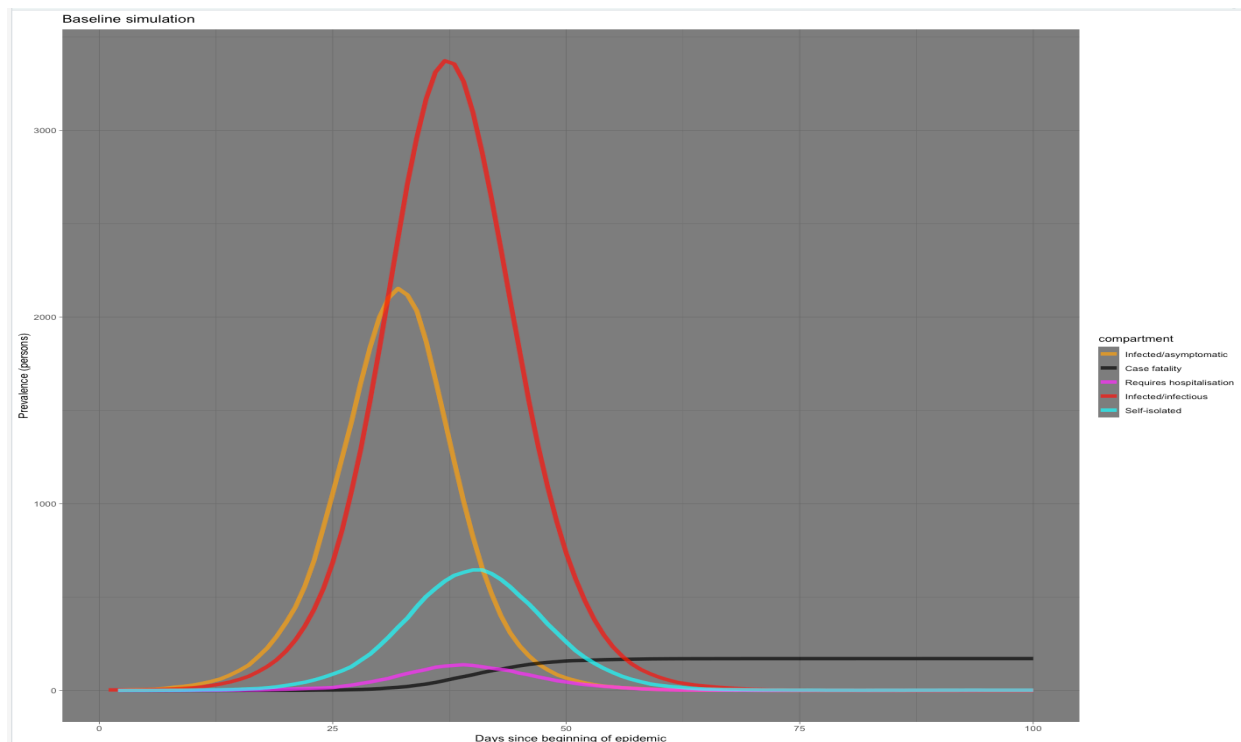
Points to be noted from the above simulation of 10000 people:

1. The pandemic is everywhere in the span of just sixty days.

2. The exponential behavior can be observed in the prevalence which tapers off later. This is good.

3. The prevalence matches with the I and Q categories but not so good in the E compartment.

4. The need of hospitalization is sensible.

5. The fatality rate is yet rising, which is expected.

The intervention experiment in the extended SIR model:

The experiment is at the stage where certain parameters can be altered in order to get better insight. The figure below models the effects of rising the rate of self-quarantine. The time-variant parameters are used here. The self-quarantine rate in the model below is set from 0.033 which begins at the fifteenth day till 0.5 at the thirtieth day.
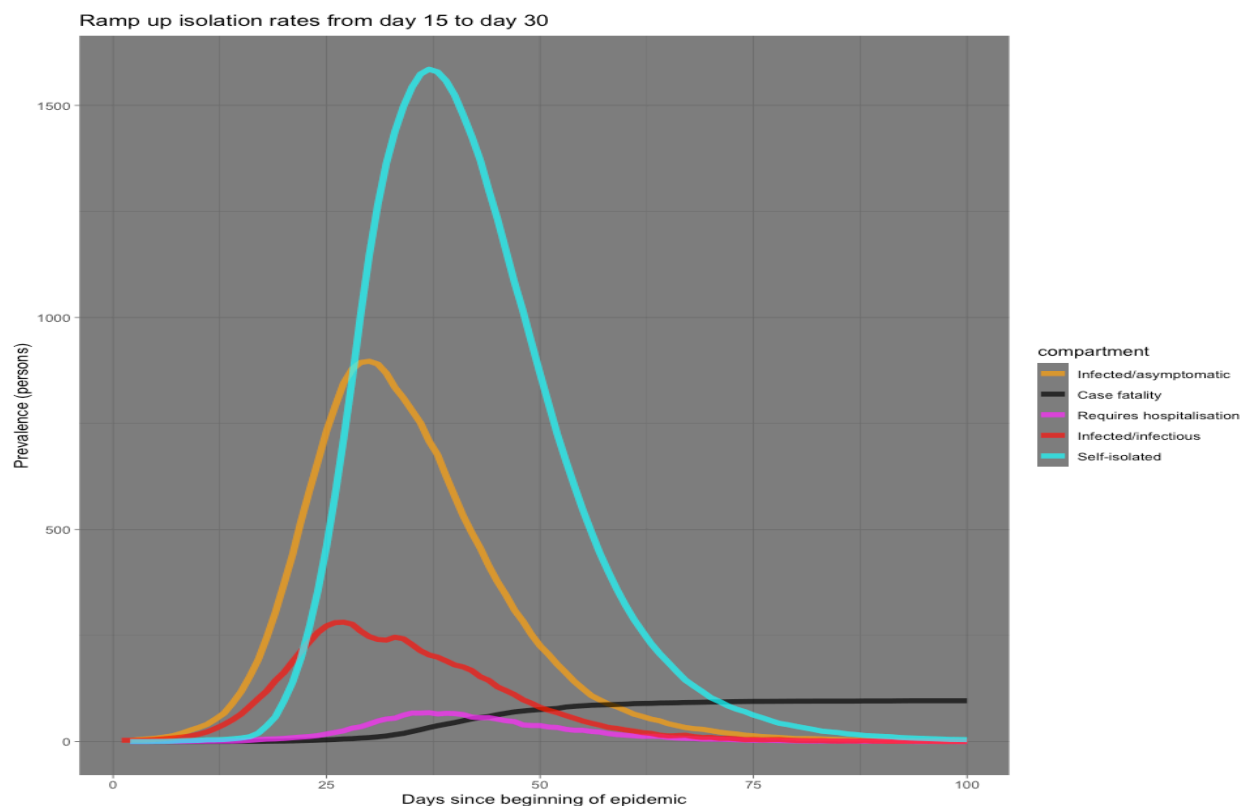

Lets compare the above model with the baseline model:

Baseline vs ramping up isolation simulations
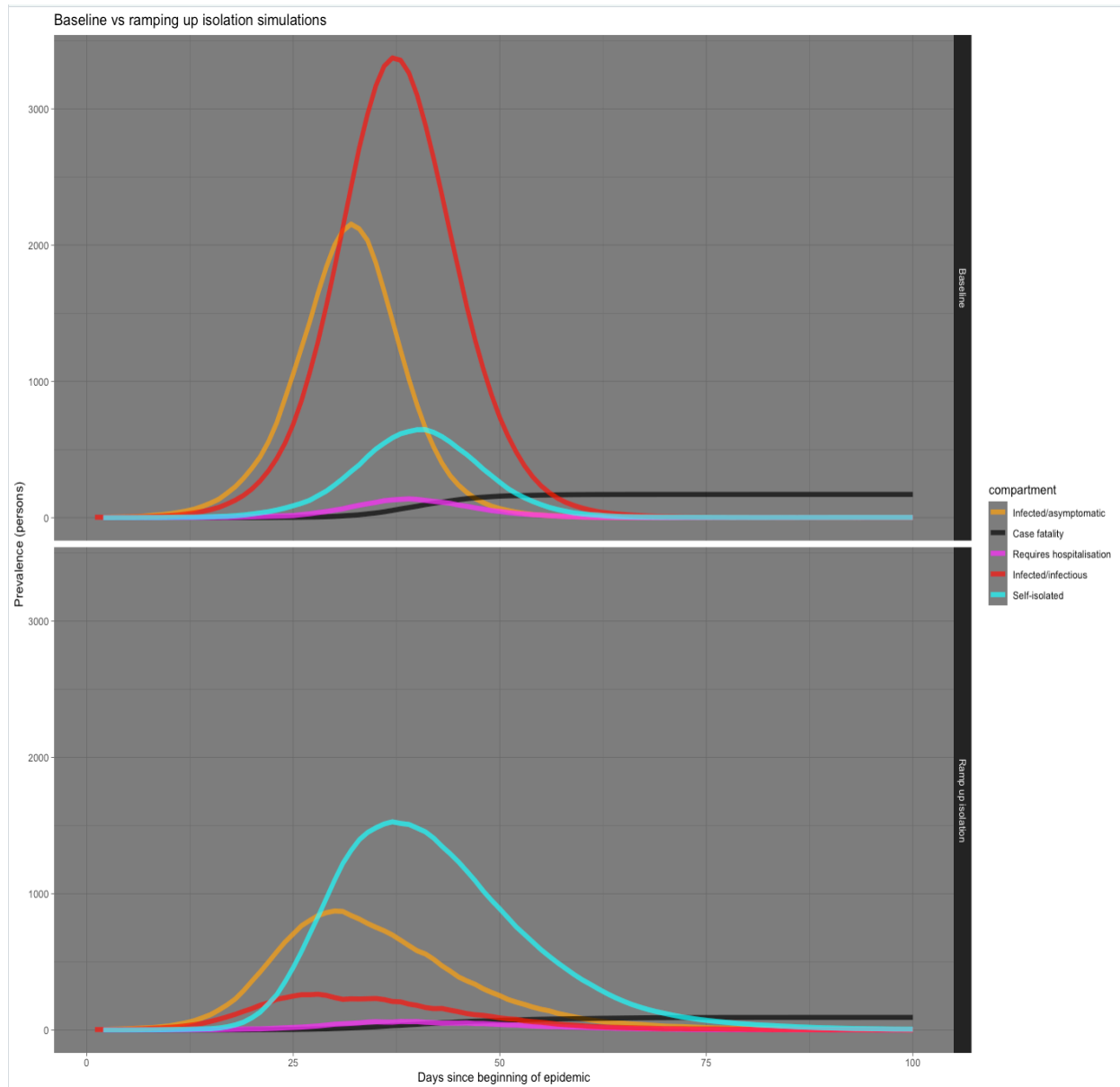
From the figure above it is easy to conclude that one should self-quarantine after getting sick.

Lets compare the same outputs by deleting the parameters like infected/asymptomatic, infected/infectious and self-isolated. Now the only parameters that simulate the model are *requiring hospitalization* and *case fatality* prevalence numbers. The red line which is on the x-axis in the graph depicts the capacity the hospitals have. It is clear from the visualization that *Requires hospitalization* prevalence numbers crosses the line of hospital capacity. Which eventually means that the sick people who need immediate assistance will not be provided medical help and probably end up loosing their lives. This was the whole agenda of simulation of this model.
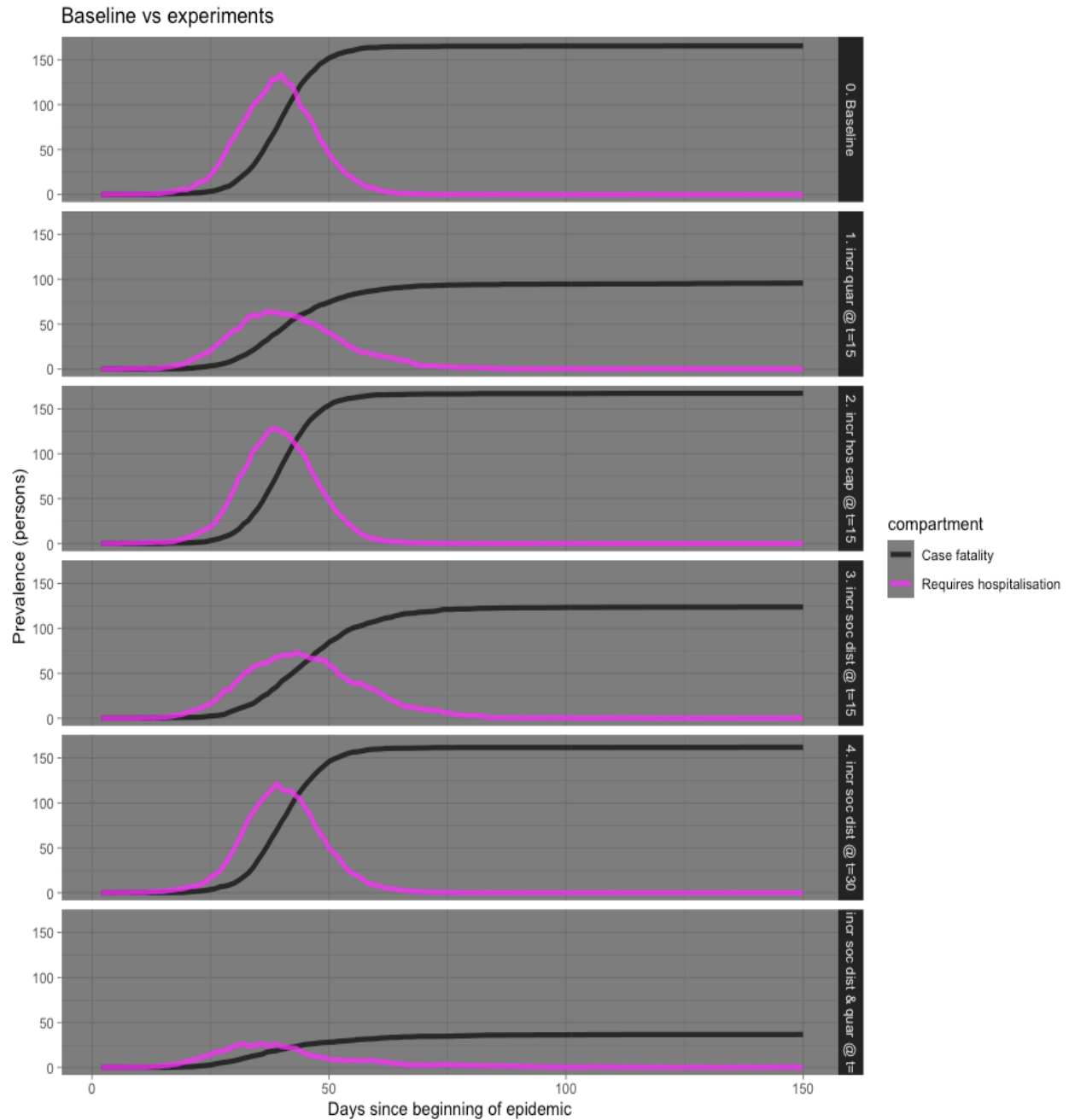
Baseline vs ramping up isolation simulations
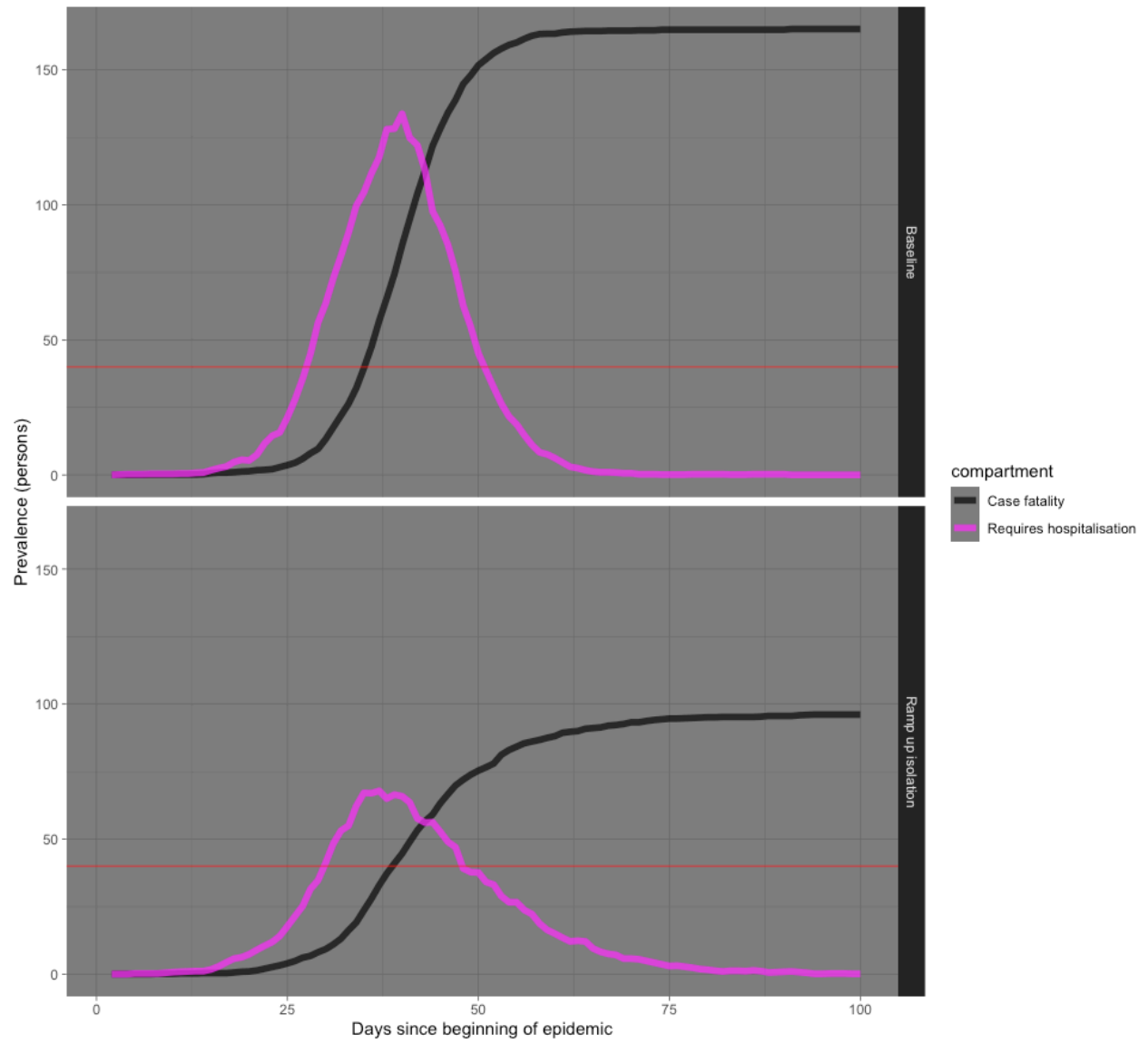
Results and Discussion

From the above experiments there are certain points that can be noted:

Baseline vs experiments
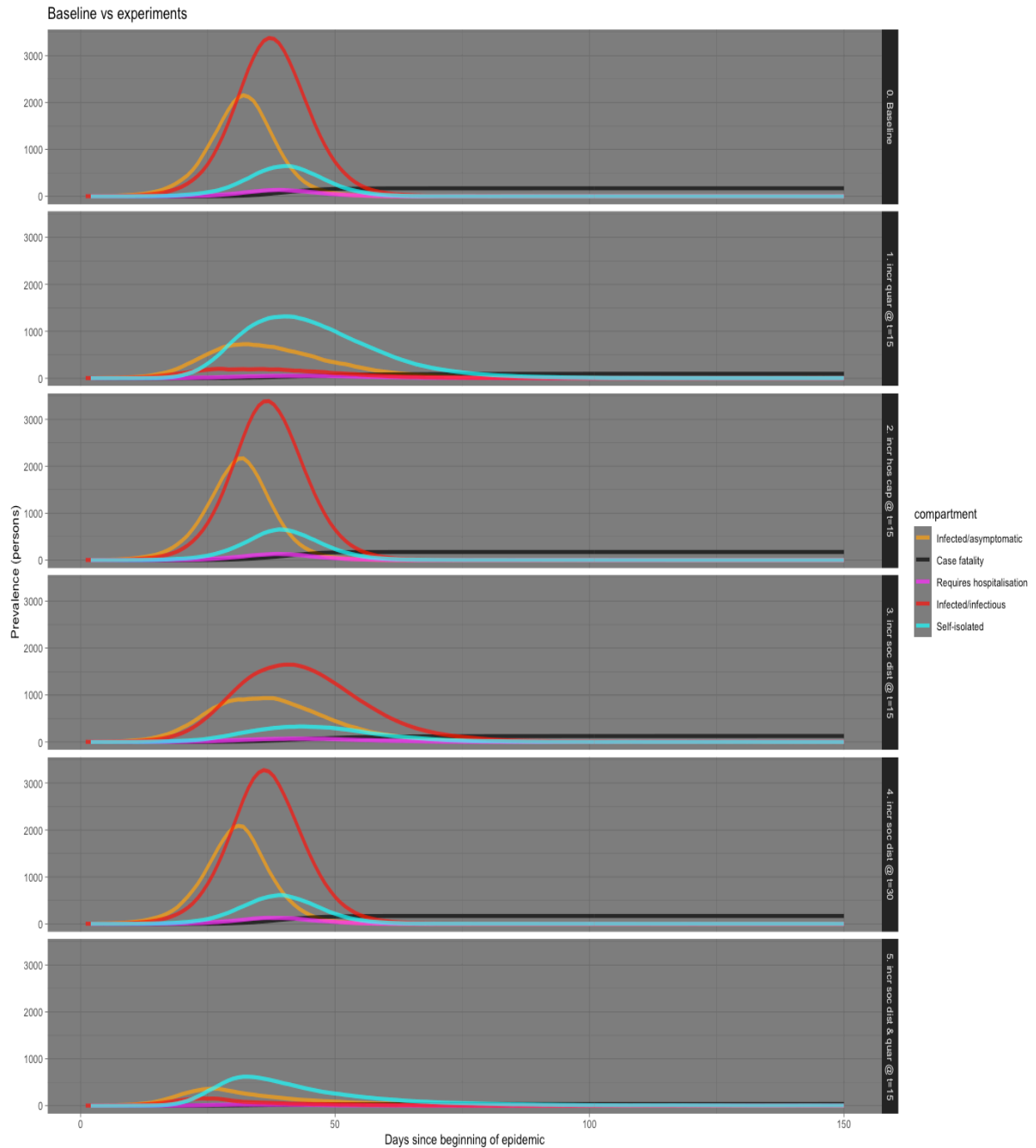
1. It is very sensible to increase the capacity of the hospital which can mitigate the fatality rate to a reasonable level. However it is not an easy job to make large availability in a short span of time.

2. The strict and rigorous self-isolation of the infected individuals is very effective. The sooner an individual practice self-isolation the better it is.

3. The increase in the social distancing is one of the key factor that works in decreasing and flattening the pandemic curve. This practice will be more effective if done very early. Even before leaving a chance of getting suspected for infection.

4. In the end, a combination of both rapid self-isolation and monitoring social- distancing is very effective without the unnecessarily affecting the economy.

The above experiments have been very useful for understanding the effect of public health interventions in a realistic way. The primary insights which are shown by experimenting the above simulations are somehow matching with the ICL report.

Conclusion

By far this model is effective to identify and isolate whoever is infectious, for example with good testing and quick responsiveness.

In the simulations, once we hit some critical threshold of cases, then it starts to send people to a separate location one day after the get the infection. This is the stand-in for whatever isolation would look like in the real world and doesn't need to transport all the infected people n the same box. Well, this totally halts an epidemic in its tracks. These people will be never isolated (who have no symptoms). The peak of the simultaneous cases is only a tiny bit higher, but there is a long tail as it takes much longer to stamp out, resulting in about twice as many total cases. This gets more interesting when the setting with many communities and travel is done between them. The things are set to wait until a certain critical threshold of cases is hit before or little dot society kicks into gear and takes action. It is interesting to note that's several parameters are run that take around three times longer to reach this point than others, before the law of large numbers kicks in, the pace of growth can have as much to do with a role of the dice as it does with anything intrinsic to the disease itself. It does keep the curve flat, but now it is as much thicker tail during the longer time to track all the cases, with over half the population getting affected this time.

If half the infectious people are getting quarantined, it doesn't mean half the total population will end up with the disease. Because there are so many agents out there spreading it, we end up with a situation that's only barely better than if nothing has been done at all. And in all the cases the presence of widespread social distancing absolutely flattens the curve.

However, in terms of the ultimate number of cases that run and even the 99 percent end up with little less than half the population ultimately getting infected. That case with 99% repelling each other takes longer to get there. Nearly 10 % population cheating adds enough instability to keep the fire slowly burning for a long time. These are toy models and it is left on the intelligence of the viewer to judge if the behaviors of the dots accurately reflect what social distancing would mean for you and your life. Someone fully sequestered to their home is not necessarily affected by the random jiggling of the neighbor. But then only few live independently from everyone else. In so far, the models are outlandish.

References

https://cran.rproject.org/web/packages/covid19.analytics/vignettes/covid19.analytics.html

https://stats.stackexchange.com/questions/446712/fitting-sir-model-with-2019-ncov-data-doesnt-conververge

https://rviews.rstudio.com/2020/03/19/simulating-covid-19-interventions-with-r/\

https://rviews.rstudio.com/2020/03/05/covid-19-epidemiology-with-r/

https://www.youtube.com/watch?v=gxAaO2rsdIs&t=626s

https://www.youtube.com/watch?v=Qrp40ck3WpI&t=18s