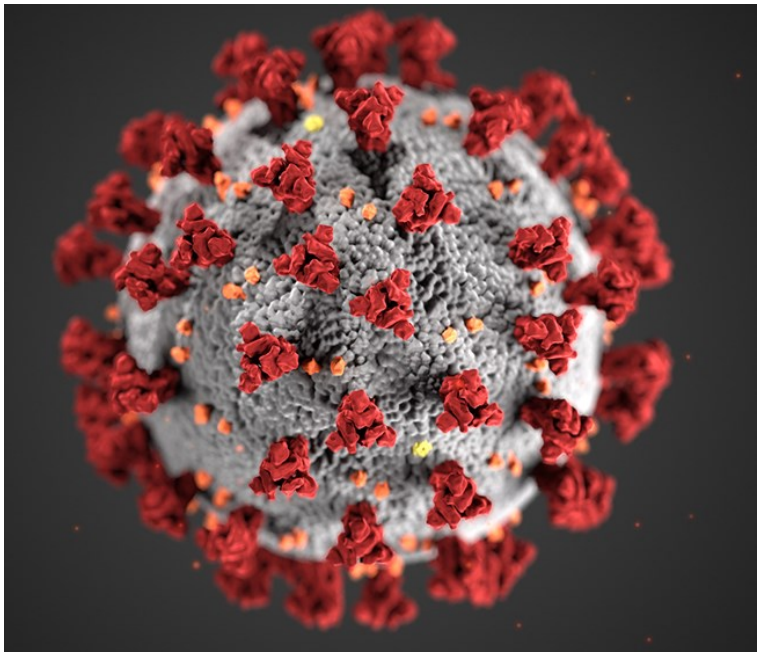
Fitting the SIRD Model Using Real Covid-19 Data with R and RStudio

SIMIODE EXPO 2021

Boyan Kostadinov, New York City College of Technology, CUNY

02/13/2021

Introduction



We use real COVID-19 data from January 22, 2020 until February 11, 2021, provided by the Johns Hopkins University Center for Systems Science and Engineering.

First, we do some exploratory data analysis, but the main goal is to use the data to fit the SIRD model parameters via least squares, and then visualize and compare the solution to the fitted model with the actual data.

In this talk, we focus on the **computational** aspects of the project. We use the **R** programming environment with **RStudio**, where we can create an interactive **R Markdown** notebook and generate a PDF report with the project narrative, LaTeX expressions, code, numerical and graphical results, all seamlessly knitted into the final report.

The COVID-19 Data

The **Johns Hopkins University Center for Systems Science and Engineering** (JHU CSSE) created a publicly available GitHub data repository to consolidate data from various sources, such as the World Health Organization (WHO), the Center for Disease Control and Prevention (CDC) in the US and many other sources.

We use the **coronavirus** R package, which provides a convenient access to the JHU data without having to download any data sources. The raw data is pulled from the JHU CSSE COVID-19 data repository.

Once the **coronavirus** R package is installed and loaded, we can start working with the data inside an interactive R Markdown document in RStudio.

Getting the COVID-19 data

The **coronavirus** package contains the daily summary of Coronavirus cases (confirmed, death, and recovered), by state/province, from all reporting countries around the world.

Once the R package **coronavirus** is installed in RStudio, we can load it inside an **R Markdown** (Rmd) document, by running the chunk of code below:

```
library(coronavirus)
data("coronavirus")
```

Now, the COVID-19 data are ready to be used.

Getting the COVID-19 data

There are 7 variables in the data with more than 300,000 observations, between January 22, 2020 and February 11, 2021. We can peek inside the data using the `head()` function:

```
##         date province             country       lat      long      type cases
## 1 2020-01-22              Afghanistan  33.93911  67.70995 confirmed     0
## 2 2020-01-22                  Albania  41.15330  20.16830 confirmed     0
## 3 2020-01-22                  Algeria  28.03390   1.65960 confirmed     0
## 4 2020-01-22                  Andorra  42.50630   1.52180 confirmed     0
## 5 2020-01-22                   Angola -11.20270  17.87390 confirmed     0
## 6 2020-01-22 Antigua and Barbuda  17.06080 -61.79640 confirmed     0
```

The `type` variable has 3 possible values: `confirmed` (infected), `recovered` and `death`. We use the `date`, `country`, `type` and `cases` data variables.

Exploratory data analysis of the COVID-19 data

We use the **dplyr** R package to do all the data analysis on the COVID-19 data.

To install the **dplyr** package it is best to install the entire **tidyverse** collection of R packages for modern data analysis and visualizations, developed by RStudio.

Once installed in RStudio, we can load the **tidyverse** collection of packages by running the code below:

```
library(tidyverse)
```

First, we get a summary of total cases by country and type.

Top 8 Total Confirmed Cases as of February 11, 2021

country	type	total_cases
US	confirmed	27,390,465
India	confirmed	10,880,603
Brazil	confirmed	9,713,909
United Kingdom	confirmed	4,010,376
Russia	confirmed	3,983,031
France	confirmed	3,465,964
Spain	confirmed	3,041,454
Italy	confirmed	2,683,403

The R code

```
summary_df <- coronavirus %>%
  group_by(country, type) %>%
  summarise(total_cases = sum(cases)) %>%
  arrange(-total_cases)
# top 8 total confirmed cases
summary_df %>% filter(type=="confirmed") %>% head(8) %>%
  knitr::kable(caption="Top 8 Total Confirmed Cases as of February 11, 2021",
               format.args = list(big.mark = ","))
```

Top 8 Total Death Cases as of February 11, 2021

country	type	total_cases
US	death	475,291
Brazil	death	236,201
Mexico	death	171,234
India	death	155,447
United Kingdom	death	115,748
Italy	death	92,729
France	death	80,951
Russia	death	77,415

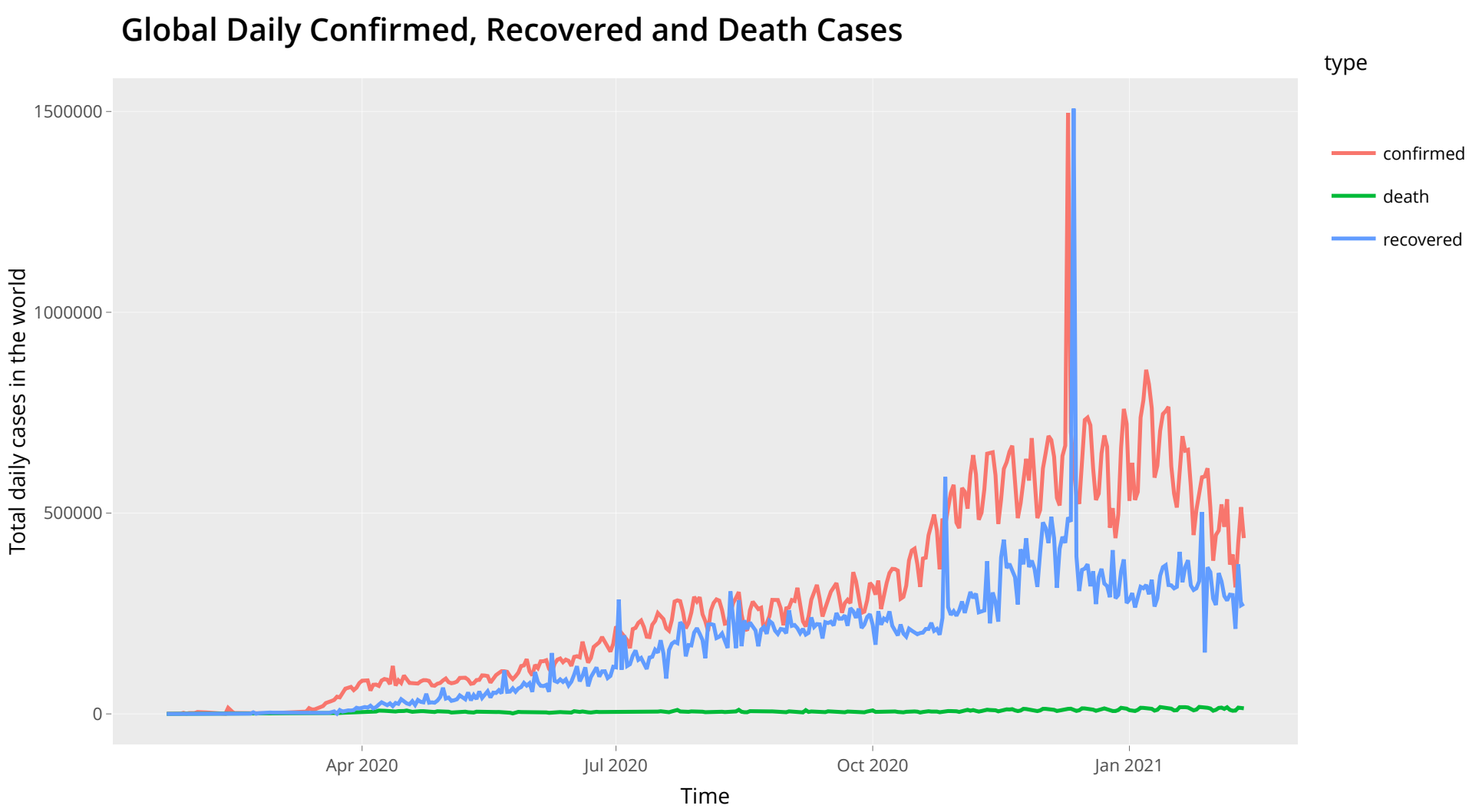
Top 8 Total Recovered Cases as of February 11, 2021

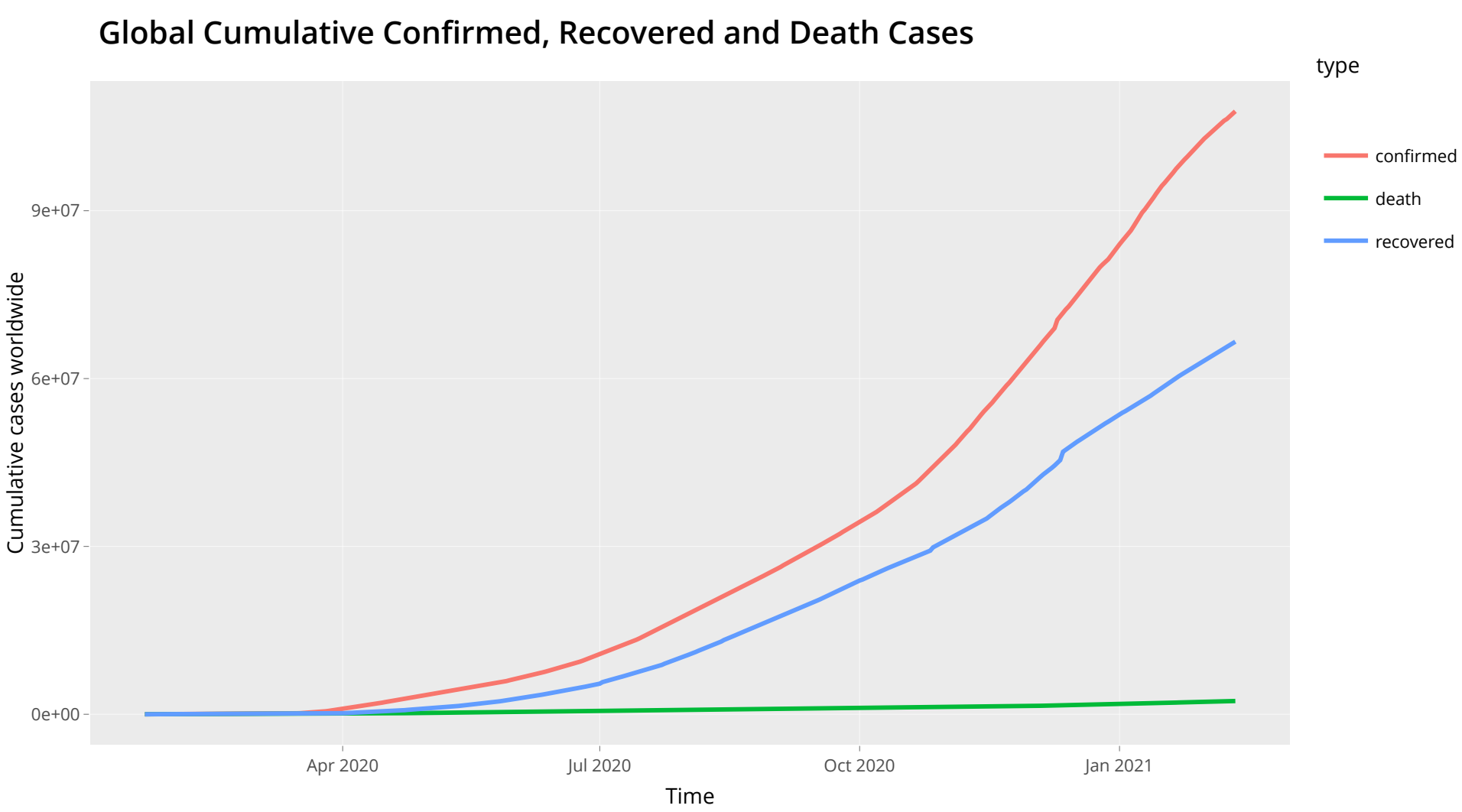
country	type	total_cases
India	recovered	10,589,230
Brazil	recovered	8,637,050
US	recovered	6,298,082
Russia	recovered	3,499,230
Turkey	recovered	2,453,096
Italy	recovered	2,185,655
Germany	recovered	2,109,508
Colombia	recovered	2,065,209

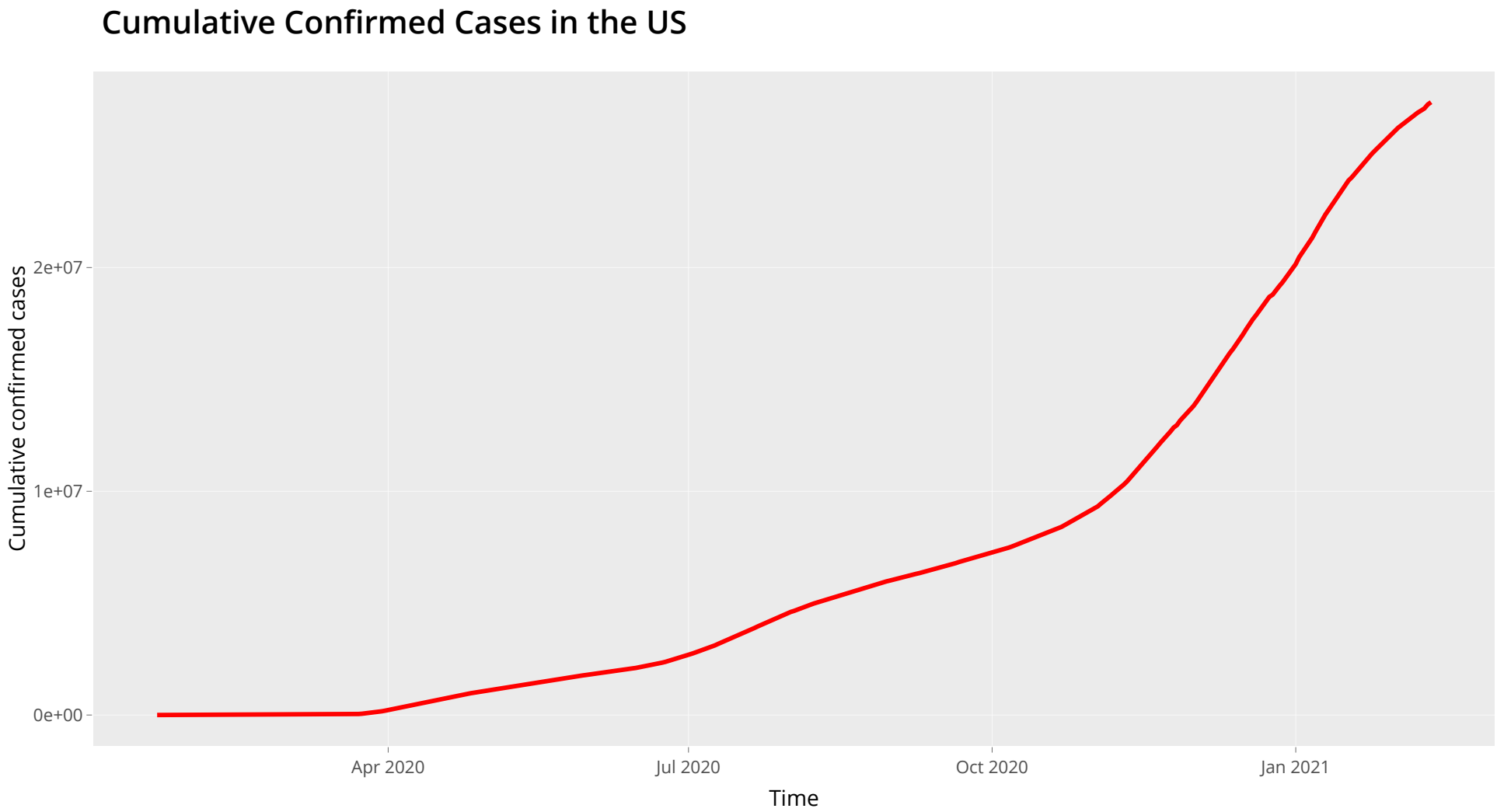
Global Totals as of February 11, 2021

```
summary_df %>%
  group_by(type) %>%
  summarise(total = sum(total_cases)) %>%
  knitr::kable(format.args = list(big.mark = ","))
```

type	total
confirmed	107,778,443
death	2,368,527
recovered	66,597,854


Global Daily Confirmed, Recovered and Death Cases
type
confirmed
death
recovered
1500000
1000000
500000
0
Total daily cases in the world
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Time


Global Cumulative Confirmed, Recovered and Death Cases
type
confirmed
death
recovered
Cumulative cases worldwide
9e+07
6e+07
3e+07
0e+00
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Time


Cumulative Confirmed Cases in the US
Cumulative confirmed cases
2e+07
1e+07
0e+00
Apr 2020
Jul 2020
Oct 2020
Jan 2021
Time

Modeling COVID-19 in the US with a SIRD model

The SIRD model is based on the SIR model, which was developed in 1927 by W. O. Kermack and A. G. McKendrick. The SIRD model assumes a fixed population:

- **Susceptible**: $S(t)$ is the number of individuals not yet infected with the disease at time t, but susceptible to the disease.
- **Infected**: $I(t)$ is the total number of infected individuals at time t, capable of spreading the disease to the susceptible individuals.
- **Recovered**: $R(t)$ is the total number of individuals at time t who have been infected and then recovered from the disease. This group cannot be infected again and cannot transmit the infection.
- **Deceased**: $D(t)$ is the total number of individuals at time t who have died from the decease.

Modeling COVID-19 in the US with a SIRD model

The Susceptible-Infectious-Recovered-Deceased Model (SIRD) differentiates between recovered, meaning individuals who have survived the disease and now immune, and deceased. The model is based on the following system of ODEs:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta \frac{IS}{N} \\
\frac{dI}{dt} &= \beta \frac{IS}{N} - \gamma I - \mu I \\
\frac{dR}{dt} &= \gamma I \\
\frac{dD}{dt} &= \mu I
\end{aligned}$$

where $N = S + I + R + D$ is the total population, assumed constant, and β, γ and μ are the rates of **infection**, **recovery** and **mortality**, respectively.

Fitting the SIRD parameters to the COVID-19 data

We used the COVID-19 data between March 1 and March 21, 2020 as the training data to calibrate the model to. The model calibration is implemented by fitting the model parameters β, γ and μ to the training data.

Note that $I(t)$ in the model, represents the **cumulative number of infected cases** at time t, and $D(t)$ is the **cumulative number of death cases** at time t.

For this purpose, we compute from the data the vector `Infected` of cumulative infected cases in the US, and the vector `Death` of cumulative US death cases.

The initial values used to initialize the model are as follows: $N = 50\text{M}$, $S(0) = N - \texttt{Infected(0)}$, $I(0) = \texttt{Infected(0)}$, $R(0) = 0$, and $D(0) = 0$.

Numerical solutions of systems of ODEs

We use the R package **deSolve** to obtain numerical solutions to systems of ODEs.

The SIRD model can be specified by the following R function, which is then used by the **deSolve** solver:

```
SIRD <- function(time, state, parameters) {
  par <- as.list(c(time, state, parameters))
  with(par, {
    dS <- -beta*I*S/N
    dI <- beta*I*S/N - gamma*I - mu*I
    dR <- gamma*I
    dD <- mu*I
    list(c(dS, dI, dR, dD))
  })
}
```

Numerical solutions of systems of ODEs

```
library(deSolve)
N <- 50e6 # initial values
init <- c(S = N - 33, I = 33, R = 0, D = 0); days <- 1:30
parameters<-c(0.7, 0.3, 0.01); names(parameters) <- c("beta", "gamma", "mu")
solution <- ode(y = init, times = days, func = SIRD, parms = parameters)
knitr::kable(head(solution,4), digits=2)
```

time	S	I	R	D
1	49999967	33.00	0.00	0.00
2	49999939	48.74	12.11	0.40
3	49999897	71.99	29.99	1.00
4	49999835	106.33	56.40	1.88

Model fitting with least squares

We want to minimize the sum of squared differences between the data and the model solution with respect to the model parameters β, γ and μ. However, we try to fit the model only to the observed cumulative infected and death cases, and not the recovered cases, which have been considered systematically under-reported. The residual sum of squares (RSS) as a function of the parameters:

$$RSS(\beta, \gamma, \mu) = \sum_{k=1}^{M} (\texttt{Infected}[k] - I_k)^2 + (\texttt{Death}[k] - D_k)^2$$

where M is the number of observations in the training dataset, $\texttt{Infected}[k]$ and $\texttt{Death}[k]$ are the kth components of the observed cumulative infected and death cases, respectively. I_k and D_k are the kth components of the solution vectors (3rd and 5th columns of the ODE solution) produced by the model for the given parameter values.

Model fitting implementation

```r
days <- seq_along(Infected)
N <- 50e6 # initial values
init <- c(S = N - Infected[1], I = Infected[1], R = 0, D = 0)
## RSS as an R function
RSS <- function(parameters){
  names(parameters) <- c("beta", "gamma", "mu") # parameters must be named
  solution <- ode(y = init, times = days, func = SIRD, parms = parameters)
  I <- solution[, 3] # 3rd column of ODE solution
  D <- solution[, 5] # 5th column of ODE solution
  return(sum((Infected - I)^2 + (Death - D)^2))
}
```

Optimization

We use the base R general-purpose optimizer `optim()` to fit the SIRD model to the training data by finding the values of β, γ and μ that **minimize the RSS function**, and thus provide the model with the "best" fit to the training data.

For the optimization, we use the method `L-BFGS-B`, a modification of the `BFGS` quasi-Newton method, which allows constraints, that is each variable can be given a lower and/or upper bound.

In general, the fitting process may not be very stable and slightly different choices for the initial values of the parameters may produce quite different estimates for the optimal parameter values.

Optimal parameter values and R_0

```
 optimal_sol <- optim(c(0.5, 0.5, 0.5), RSS, method = "L-BFGS-B",
                      lower = c(0, 0, 0), upper = c(1, 1, 1))
fitted_pars <- setNames(optimal_sol$par, c("beta", "gamma", "mu"))
print(round(fitted_pars,4))
```

```
##    beta  gamma     mu
## 0.6712 0.3287 0.0064
```

The **basic reproduction** number R_0 represents the expected infected cases generated by one infectious case. If $R_0 > 1$ it predicts an epidemic. R_0 can be computed by analyzing the threshold between a stable and unstable equilibrium of the model, using the eigenvalues of the Jacobian at the free equilibrium.

The SIRD basic reproduction number

The larger the value of R_0, the harder it is to control the epidemic, and the higher the probability of a pandemic. The threshold between a stable and unstable equilibrium of the system of ODEs for the SIRD model is given by:

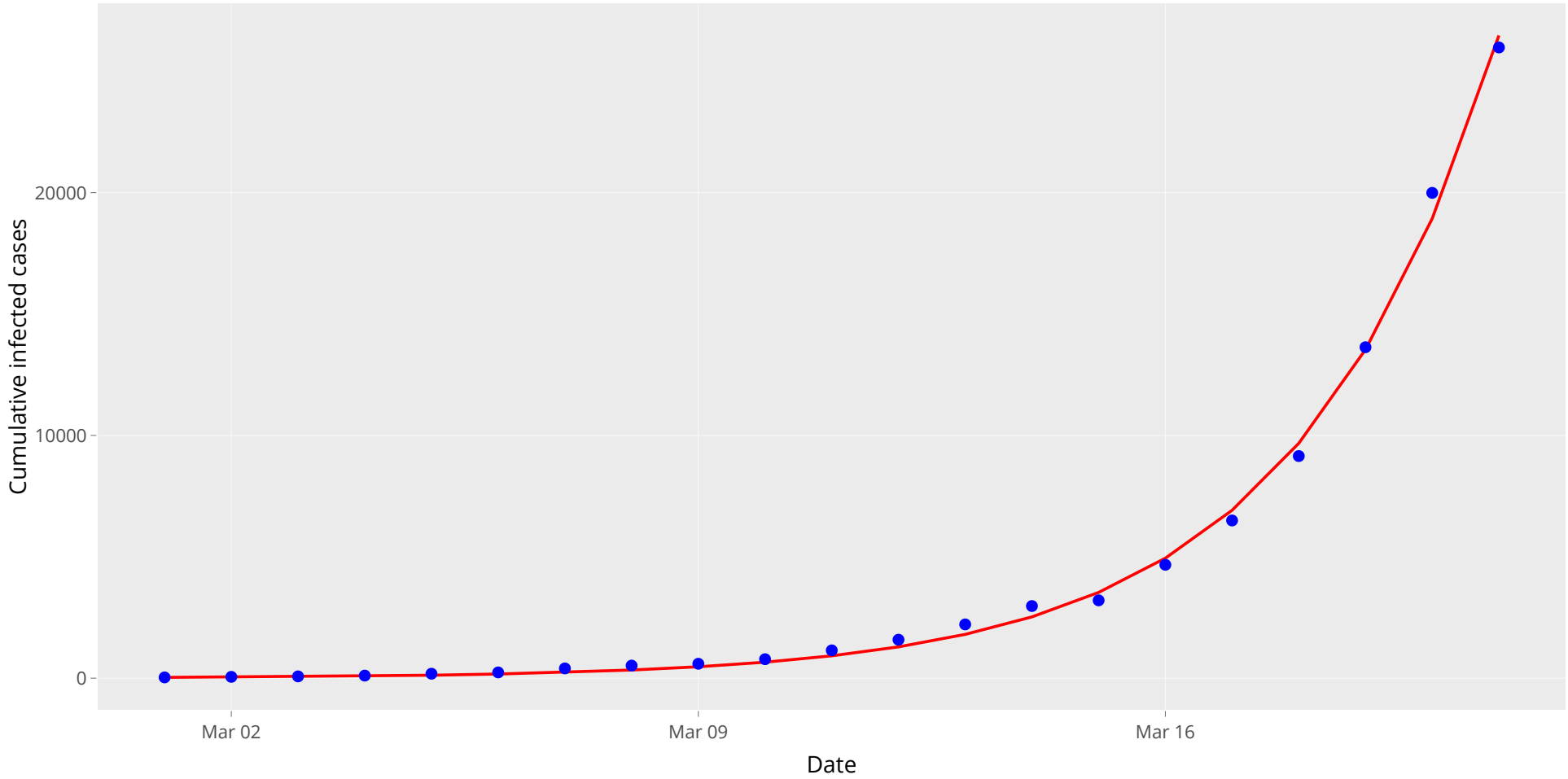
$$R_0 = \frac{\beta}{\gamma + \mu} = 2.0027$$

In the literature, the SIRD basic reproduction number has been found to be in the range $(1, 2.8)$ for different countries, with mild to severe cases of their local epidemics, based on a study from March 2020.

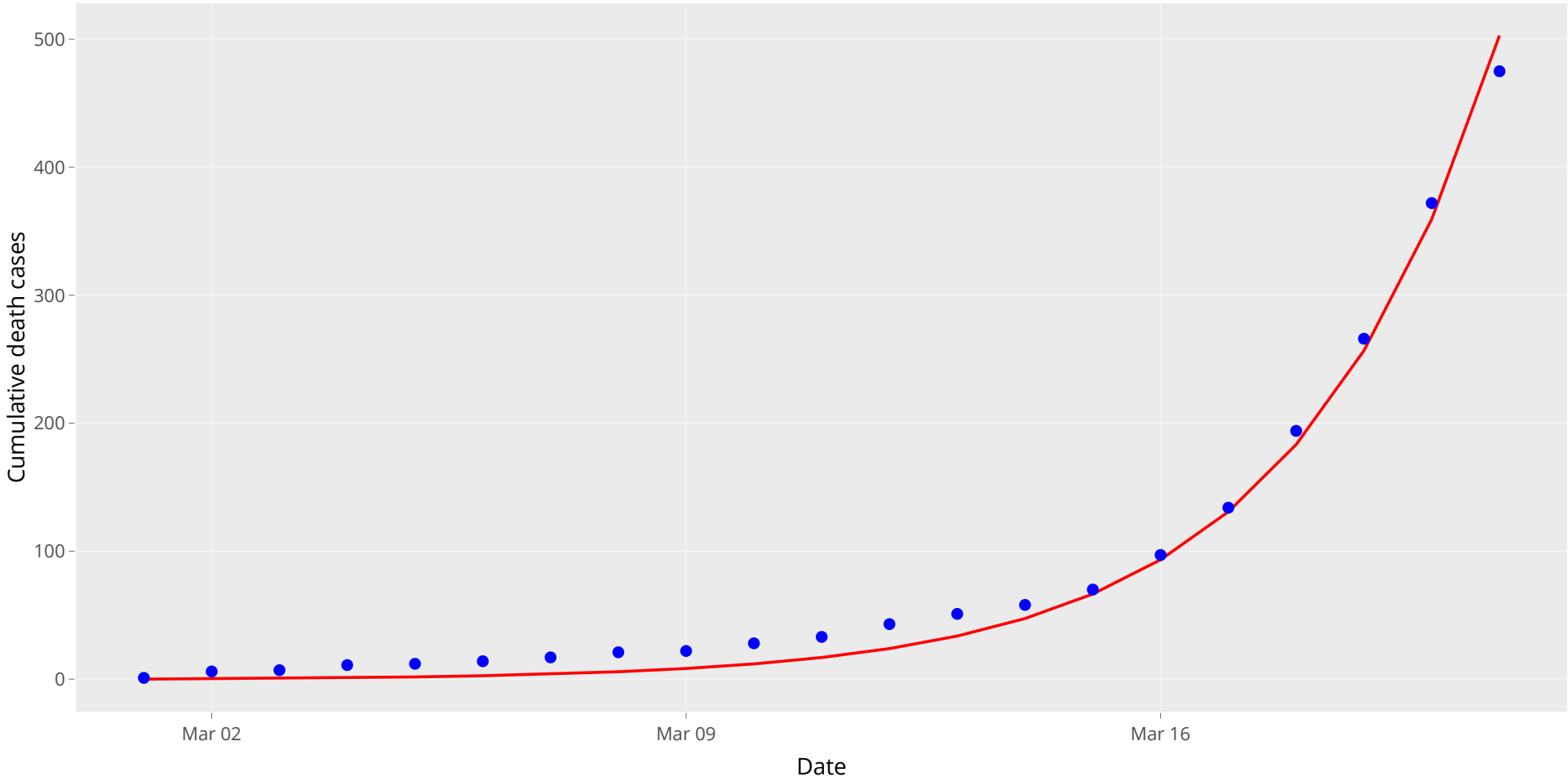
Sources:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7438206/
- https://academic.oup.com/jtm/article/27/2/taaa021/5735319

Fitted SIRD model (red) and observed cumulative infected cases (blue)



Fitted SIRD model (red) and observed cumulative death cases (blue)



Conclusions

For simple models, the fraction of the population that needs to be *effectively immunized* to prevent the spread of the pandemic, known as the **herd immunity threshold**, has been estimated to be larger than:

$$1 - \frac{1}{R_0} \approx 0.5$$

Thus, based on this simple SIRD model, calibrated to COVID-19 data from the first 3 weeks of March 2020, we conclude that more than 50% of the US population should be **effectively immunized** to stop the pandemic.

Conclusions

The current US population is around 330M. As of late-February, around 50M Americans will have been immunized with either the Pfizer or Moderna vaccines (at least the first dose). In addition, around 30M people have been confirmed infected through testing, as of February 2021, and all recovered people will be effectively immune.

With the emergency approval of Johnson & Johnson's single dose vaccine, it is projected that there will be enough supplies to vaccinate about 70M Americans per month. This would be enough to vaccinate around 70% of the adult population by the end of April 2021.

In fact, the real vaccination level reached by then would most likely be higher, given that there must be a significant fraction of the US population that has become immune without being officially confirmed.

Conclusions

This project may serve as an example of:

- solving a real-world problem of great importance that everyone can relate to.
- doing exploratory data analysis and visualizations to explore real data.
- using real COVID-19 data to fit a mathematical model and make predictions.

This project could also serve as the basis for calibrating more sophisticated models to real COVID-19 data. One such model could be the SEIR model.


danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
dank je
thank you
gracias
tapadh leat
bedankt
hvala
mauruuru
mochchakkeram
dziękuję
go raibh maith agat
obrigado
sagolun
sukriya
kop khun krap
arigatō
takk
dakujem
grazie
мерси
terima kasih
감사합니다
merci
ευχαριστώ