Estimating under-reporting of COVID-19 cases in Indian states

MA4240 - Applied Statistics

IIT Hyderabad

April 26 2022

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Team Members

- Adepu Adarsh Sai - AI20BTECH11001
- Ananthoju Pranav Sai - AI20BTECH11004
- Murarisetty Adhvik - AI20BTECH11015
- Perambuduri Srikaran - AI20BTECH11018
- Arun Siddardha - AI20BTECH11019
- Yashas Tadikamalla - AI20BTECH11027

Contents

Introduction

Introduction

- It is observed that there has been a significant under-reporting in COVID-19 cases in India. Some of the possible reasons for this are:
  - Low testing coverage
  - Asymptomatic cases
- We try to obtain both, an average as well as a time varying estimate for the fraction of COVID-19 cases reported in each state of India.
- The former is calculated using a delay adjusted case fatality ratio (CFR) approach for different states, while the latter is obtained by modelling the problem as a time-varying Poisson process.

Definitions

Log-Normal Distribution

- A log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed.

The PDF of $X \sim \text{Lognormal}(\mu, \sigma^2)$ is given by

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right) \tag{1}$$

- $\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \qquad Var[X] = [e^{(\sigma^2)} - 1]e^{(2\mu+\sigma^2)}$
- The length of chess games tends to follow a log-normal distribution.

Poisson distribution

A discrete random variable X is said to be a Poisson random variable with parameter λ, shown as $X \sim Poisson(\lambda)$, if its range is $R_X = 0, 1, 2, 3, ...$, and its PMF is given by

$$P_X(k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in R_X \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- If $X \sim Poisson(\lambda)$, then $E[X] = \lambda$, and $Var(X) = \lambda$.

Time-varying Poisson process

- Poisson process is most commonly used in cases when we're counting the number of times certain events appear to occur at a specific rate, but, are truly random.
- A time varying Poisson process is a Poisson process whose event rate λ varies with time.
- The probability of an event occuring k times in a time interval with event rate λ_t is

$$P_X(k) = \frac{e^{-\lambda_t} \lambda_t^k}{k!} \tag{3}$$

Poisson Regression

- Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables.
- Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters.
- If $\underline{x}_t \in \mathbb{R}^n$ is a vector of explanatory variables, then we can model λ_t as

$$\log(\lambda_t) = \alpha + \underline{\beta}^\top \underline{x}_t \tag{4}$$

  where $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^n$
- Hence,

$$PMF(Y_t/\underline{x}_t) = \frac{e^{-\lambda_t}\lambda_t^{Y_t}}{Y_t!} \tag{5}$$

Poisson Regression

- The log-likelihood of the data is given by

$$ln(L(\underline{\beta})) = \sum_t (Y_t \underline{x}_t \underline{\beta} - e^{x_t \underline{\beta}} - lnY_t!) \tag{6}$$

- We use Maximum Likelihood Estimation to find the coefficients β of the regression model

$$\sum_t (Y_t - e^{\underline{x}_t \underline{\beta}}) \underline{x}_t = 0 \tag{7}$$

- Solving this, will give the coefficients $\underline{\beta}$.

Data and Key Assumptions

Data and Key Assumptions

- The data is taken from covid19india.org. It consists of the number of cases reported and the number of deaths on each day for every state in India.
- Some of the assumptions we assumed to perform the analysis are:
  - Deaths due to COVID-19 are reported accurately i.e., no under-reporting of deaths.
  - For fatal cases, the distribution of delay from confirmation to death is the same as the distribution of hospitalisation to death. We assume it is a log-normal distribution with mean of 13 days and standard deviation of 12.7 days.
- We will use the baseline CFR value as 0.1%.

Average estimate of fraction of cases reported

Average estimate of fraction of cases reported

- Naive CFR can be calculated as

$$nCFR = \frac{\text{deaths-to-date}}{\text{cases-to-date}} \tag{8}$$

This is clearly inaccurate as the additional deaths that may arise from the cases observed to date will not be counted.

- Hence, we can use the distribution of delay from hospitalisation to calculate the delay-adjusted CFR as ratio of deaths to the expected number of eventually fatal cases.

$$cCFR = \frac{\sum_{t=0}^{T} d_t}{\sum_{t=0}^{T} \sum_{s<t} p_{t-s} c_s} \tag{9}$$

Here, d_t and c_t are the number of new deaths and the number of new cases reported on day t from a certain region respectively. p_s is the probability that an eventually fatal case will lead to death on the s^{th} day from the day of reporting. T is the last day for which the data is available.

Average estimate of fraction of cases reported

- In regions where the cases have been under reported, the adjusted CFR will be higher than the true CFR. So, calculating the ratio of the true CFR to the adjusted CFR will give an estimate of the fraction of cases reported.

$$f_{avg} = \frac{CFR}{cCFR} \tag{10}$$

- The expected total number of deaths to occur among the reported cases on day t can be written as,

$$e_t = \sum_{s<t} p_{t-s} c_s . CFR \tag{11}$$

Here, CFR is the true CFR which we assumed at the start of the analysis.

Average estimate of fraction of cases reported

- The ratio of $\sum_{t=0}^{T} e_t$ to the number of deaths reported by day T gives the average fraction of true cases that have been reported in a region over the complete time period.

$$f_{avg} = \frac{\sum_t e_t}{\sum_t d_t} \tag{12}$$

Time varying estimate of fraction of cases reported

Time varying estimate of true fraction of cases

- We can further improve the estimate to obtain a time varying estimate of the fraction of cases reported.
- We model the number of daily deaths as a time-varying Poisson process. The number of deaths on day t is a random Poisson variable with mean,

$$\lambda_t = \frac{e_t}{f_t} \tag{13}$$

Here, f_t is the fraction of cases reported on day t

Time varying estimate of fraction of cases reported

- We estimate λ_t by performing Poisson regression on daily death cases. Subsequently we can find f_t using (13).

$$\log(\lambda_t) = b_0 + b_1(t) + b_2(c_t) \tag{14}$$

  where c_t is the number of cases reported on day t.
- Estimate of daily fraction of reported cases is given by

$$f_t = \frac{e_t}{\lambda_t} \tag{15}$$

Confidence Interval bounds

Confidence Interval bounds

- Let $X_1, X_2, .., X_n$ be a random sample from a normal distribution with mean μ and variance σ^2.Then the 95% CI for the mean μ is given by

$$\overline{x} \pm 1.96\sqrt{\frac{\sigma^2}{n}} \tag{16}$$

- Using CLT, we can claim that the 95% CI for λ_t will be

$$\lambda_t \pm 1.96\sqrt{\frac{\lambda_t}{n}} \tag{17}$$

where n is the the total number of days.

$$\text{95\% CI for } \lambda_t = \left(\lambda_t - 1.96\sqrt{\frac{\lambda_t}{n}}, \lambda_t + 1.96\sqrt{\frac{\lambda_t}{n}}\right) \tag{18}$$

Results and Observations

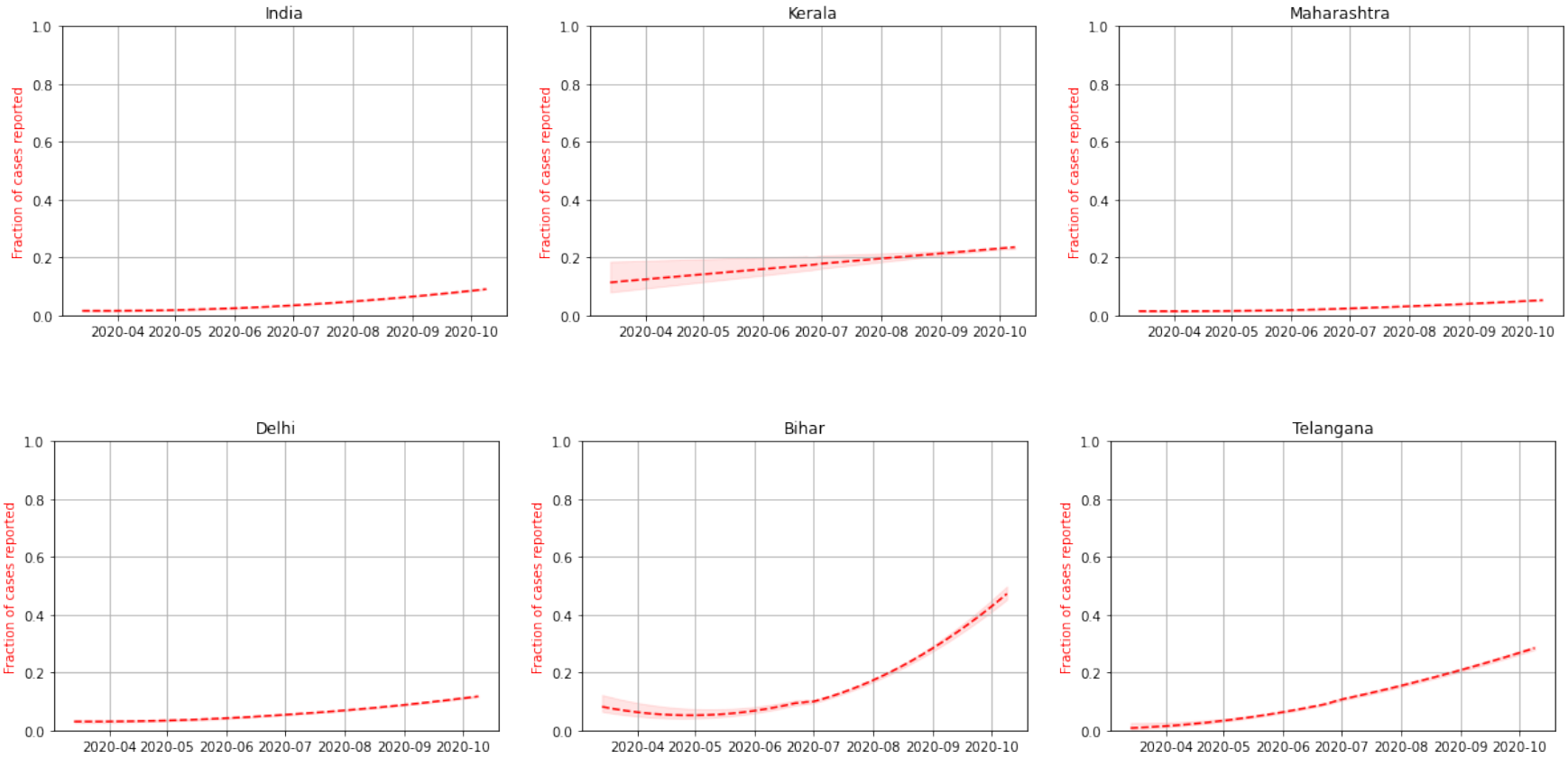
Results

- The codes can be found here.
- Average and time-varying estimate of fraction of cases

State /country	Deaths	Cases	nCFR (%)	cCFR (%)	Percentage Reported (%)
India	107459	6976461	1.54	1.78	5.60
Telangana	1208	208025	0.58	0.65	15.32
Maharashtra	39731	1506018	2.64	3.01	3.32
Bihar	934	193826	0.48	0.52	19.05
Delhi	5692	303693	1.87	2.04	4.91
Kerala	3773	122459	3.08	3.66	2.73

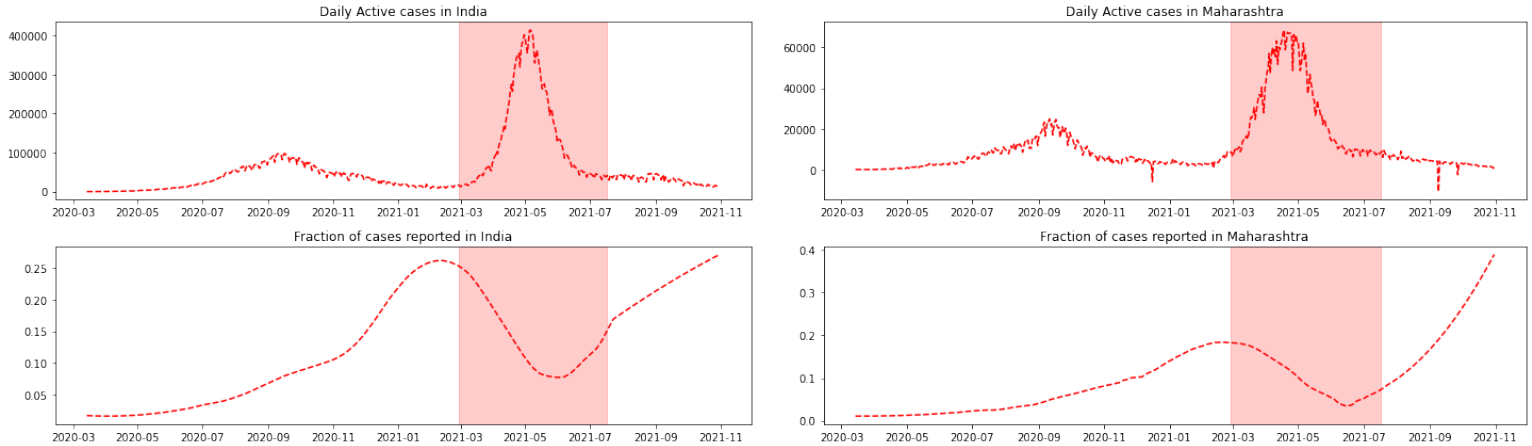
- There has been a major under-reporting, as indicated by the last column.

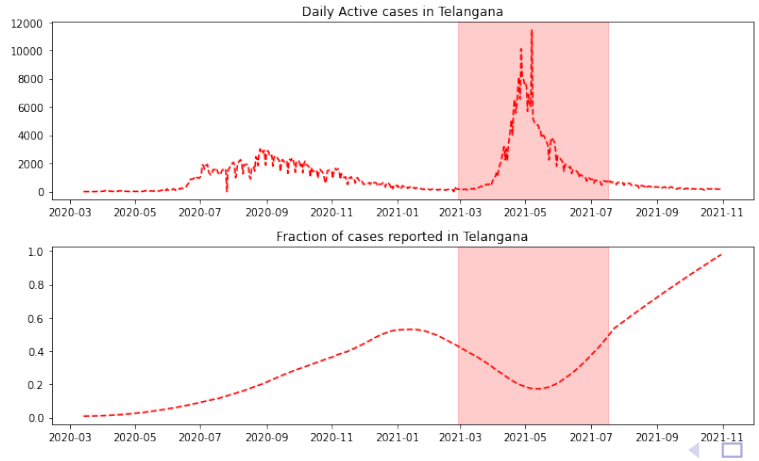
The fraction of cases reported in various regions as a function of time, assuming a baseline CFR of 0.1%



Observations

- It is observed that there has been a significant under-reporting during 2^{nd} wave of COVID-19 across India. The shaded region represents the interval of second wave. Here are a few plots





Strengths and limitations of the study

Strengths:

- In states where extensive testing is infeasible, this study provides a method to quantify the true extent of spread of COVID-19.
- It provides important information, of trends of under-reporting in different states, which could be used for policy making.

Limitations:

- The accuracy of these results depends greatly on the quality of the data and the assumptions being made.

References

1. Jayakrishnan Unnikrishnan, Sujith Mangalathu and Raman V Kutty, "Estimating under-reporting of COVID-19 cases in Indian states: an approach using a delay-adjusted case fatality ratio", 2020.
2. Ioannidis J., "The infection fatality rate of COVID-19 inferred from seroprevalence data", 2020.
3. Linton NM, Kobayashi T, Yang Y, et al. "Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data", 2020.

THANK YOU!