Prediction of COVID-19 in India

An Initial Attempt

Let $c_0, c_1, c_2, \ldots, c_n$ be the daily cases on days $0, 1, 2, \ldots, n$. I choose data for the last N days, that is, the counts $c_{n-N}, c_{n-N+1}, c_{n-N+2}, \ldots, c_n$ are only used. The cumulative counts are defined as

$$C_t = c_{n-N} + c_{n-N+1} + c_{n-N+2} + \cdots + c_{n-N+t}$$

for $t = 0, 1, 2, \ldots, N$. In the following algorithm, only these cumulative counts $C_0, C_1, C_2, \ldots, C_N$ are used. In the logistic model, the cumulative counts satisfy the equation

$$C(t) = \frac{K}{1 + A\exp(-rt)},$$

where

$$A = \frac{K - C(0)}{C(0)}.$$

The observed values C_t of $C(t)$ are available. From these, the best estimates of the two unknown quantities are to be computed:

$$\begin{array}{rcl} K &=& \text{The total number of cases} \;=\; C(\infty), \\ r &=& \text{The infection rate.} \end{array}$$

The method of nonlinear regression is used for that. Let us write $C(t)$ as $C(t, K, r)$. Then, we have the following partial derivatives:

$$\begin{aligned} \frac{\partial C(t,K,r)}{\partial K} &= \frac{1 - \exp(-2t)}{\left[1 + \left(\dfrac{K - C_0}{C_0}\right)\exp(-rt)\right]^2}, \\ \frac{\partial C(t,K,r)}{\partial r} &= \frac{K\left(\dfrac{K - C_0}{C_0}\right)t\exp(-rt)}{\left[1 + \left(\dfrac{K - C_0}{C_0}\right)\exp(-rt)\right]^2}. \end{aligned}$$

I start with the initial guess

$$\begin{aligned} K^{(0)} &= 2C_N, \\ r^{(0)} &= \frac{\ln\left(\dfrac{2C_N - C_0}{C_0}\right)}{N}. \end{aligned}$$

Next, the guesses are iteratively improved as follows. Assume that the i-th approximations $K^{(i)}, r^{(i)}$ are available for some $i \geqslant 0$. For $t = 0, 1, 2, \ldots, N$, compute

$$\begin{aligned} J_{t,K} &= \frac{\partial C(t, K^{(i)}, r^{(i)})}{\partial K}, \\ J_{t,r} &= \frac{\partial C(t, K^{(i)}, r^{(i)})}{\partial r}. \end{aligned}$$

Consider the $(N+1) \times 2$ matrix in which the first column consists of the $J_{t,K}$ values and the second column consists of the $J_{t,r}$ values. Also compute the current errors

$$\Delta C_t^{(i)} = C_t - C(t, K^{(i)}, r^{(i)})$$

for $t = 0, 1, 2, \ldots, N$. Let $\Delta C^{(i)}$ be the column vector consisting of these $N+1$ error values. The least-square regression method gives

$$J^{\mathrm{t}} J \begin{pmatrix} \Delta K^{(i)} \\ \Delta r^{(i)} \end{pmatrix} = J^t \Delta C^{(i)}.$$

If the matrix $J^{\mathrm{t}}J$ is invertible, the refinements $\Delta K^{(i)}, \Delta r^{(i)}$ are available, and the next guesses for K and r are computed as:

$$\begin{aligned} K^{(i+1)} &= K^{(i)} + \Delta K^{(i)}, \\ r^{(i+1)} &= r^{(i)} + \Delta r^{(i)}. \end{aligned}$$

This refinement process stops when $|\Delta K^{(i)}| < 0.5$ and $|\Delta r^{(i)}| < 0.0001$.

Experimental Observations

The estimates of K and r vary widely depending upon N. I take $N = 15, 30, 45, 60$, and plot the different curves. It is obvious from these experiments that no meaningful predictions can be made from these curves. The data seems to be too noisy to fit the model gracefully. Another model is used by other groups for such predictions: the SIR (susceptible-infectious-removed) model. This is not implemented here, but it appears that this model too will fail miserably for prediction purposes.

Best Prediction

The value of N leading to the best fitting of the predicted curve against the available data can be taken as the best possibility for N. Given N, estimate K and r as above. The predicted cumulative sums are defined as

$$P_t = C(t + N - n, K, r)$$

for $t = 0, 1, 2, \ldots, n$. The RMS error is then calculated as

$$E_{rms} = \sqrt{\frac{1}{n+1} \sum_{t=0}^{n} (P_t - C_t)^2}.$$

N is varied in the range $[15, n]$. The value of N which corresponds to the minimum RMS error can be taken as the best choice for N.

A better approach is to look at N-day periods not only at the end, but anywhere on the n-day data. Moreover, one may take $N \geqslant 30$. Looking at the data for less than one month is arguably not a good idea.

Using Moving Averages

Moving averages can smooth out temporal fluctuations. I take a $(2w+1)$-day moving window. The average daily count for each position of the window is assigned to the central day of the window. Using these averages in place of c_t is expected to reduce RMS errors. On the other hand, the predictions made are w days old (although partial contributions are taken from the last w days). One-day moving-average predictions are the same as daily-data predictions. The window size should not be too large. It appears reasonable to take $2w+1 \in \{3, 5, 7\}$.

References

1. Motivation: `https://ddi.sutd.edu.sg/`
2. Logistic model: `https://www.medrxiv.org/content/10.1101/2020.02.16.20023606v5`
3. Logistic model: `https://en.wikipedia.org/wiki/Logistic_function`
4. Nonlinear least-square regression: `https://en.wikipedia.org/wiki/Non-linear_least_squares`
5. Database: `https://www.covid19india.org/`