Forecasting the COVID-19 Epidemic for the U.S. Appendix: Model and Data Details

Paul Ho
Federal Reserve Bank of Richmond

Thomas A. Lubik
Federal Reserve Bank of Richmond

Christian Matthes
Indiana University

May 8, 2020

1 Model for infections

We specify a model in which the growth rate of the cumulative number of cases depends on the current cumulative number of cases. In particular, denoting the cumulative number of cases normalized by population in period t by C_t, we consider:

$$\Delta \log C_t = \frac{\log(1+\gamma)}{\phi(10^{-5}; \beta_0, \beta_1, \zeta, \eta)} \phi(C_{t-1}; \beta_0, \beta_1, \zeta, \eta) \exp(\varepsilon_t^C) \tag{1}$$

$$\phi(C_{t-1}; \beta_0, \beta_1, \zeta, \eta) \equiv \exp[-C_{t-1}^{-\beta_0} - (\zeta - C_{t-1}^{\eta})^{-\beta_1}], \tag{2}$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma_C^2)$.

The model is set up to flexibly match the trajectory of cases. Initially, the rate of growth is approximately exponential. When a fraction 10^{-5} of the population has been infected, the growth rate in the absence of shocks is γ. The parameter β_0 determines how the growth rate of C_t increases or decreases in the early stages of the pandemic, capturing the appearance of large clusters or the effects of social distancing measures. The parameters η, β_1 and ζ determine the long-run number of cumulative cases and the speed at which society converges to that number, which could depend on factors such as demographics or policies. Finally, ε_t is a shock that allows for deviations from the model predictions, arising due to randomness in how the virus spreads.

We estimate the model using Bayesian methods, obtaining 2.5 million Monte Carlo draws from posterior distributions of each of the parameters $\{\gamma, \beta_0, \beta_1, \zeta, \eta, \sigma_C\}$. For our forecasts, we simulate paths for $\{\varepsilon_t^C\}$ for a sub-sample of the Monte Carlo draws.

2 Model for mortality

We assume that the number of deaths is proportional to the average number of cases over some window of n days ending h days ago:

$$D_t = \lambda \frac{C_{t-h} - C_{t-h-n}}{n} + \varepsilon_t^D \tag{3}$$

$$\varepsilon_t^D \sim \mathcal{N}(0, \frac{C_{t-h} - C_{t-h-n}}{n} \sigma_C^2). \tag{4}$$

The constant λ is the death rate. The variance of the shock ε_t^D is proportional to the average number of cases in the window, reflecting a higher variability in the absolute number of deaths from day to day when there are more infections. We take h and n as parameters to be estimated, reflecting uncertainty about the time between testing positive for Covid-19 and dying from the virus.

We estimate the model using Bayesian methods, obtaining Monte Carlo draws from posterior distributions of each of the parameters $\{\lambda, h, n, \sigma_D\}$. For our forecasts, we first simulate paths for the cumulative number of cases C_t. For each path, we draw parameters $\{\lambda, h, n, \sigma_D\}$ from the posterior and a path of shocks $\{\varepsilon_t^D\}$, which we then use to compute the implied number of deaths each day.

3 Data

Data for the United States is taken from the European Centre for Disease Prevention and Control[1], which has data for the daily number of cases in different countries since December 31, 2019. We use data from March 5, 2020 to May 3, 2020.

[1] https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases