Nowcasting COVID-19 infections

Roger Lord[1]

First version: December 29, 2020
This version: December 30, 2020

ABSTRACT

This paper deals with the nowcasting of the number of COVID-19 infections that occurred on a given date, conditional on the number of positive tests that have been reported. Dealing with the prediction of such events that are subject to a reporting delay is not new, with many applications in economics, operational research, insurance and epidemiology. We explain how the setup of Crevecoeur et al. [2019] has been used in the GitHub repository of Lord [2020] to nowcast the number of COVID-19 infections.

Keywords: Occurrence of events, reporting delay, observation delay, COVID-19, nowcasting.

1. Introduction

The headline figures of positive COVID-19 tests we read on a daily basis are not the total amount of positive COVID-19 tests that occurred on that date, or the total amount of COVID-19 infections that occurred on that date. Rather, they are very often the change in the total amount of positive COVID-19 tests from one day to the next. As an example, the RIVM, the Dutch National Institute for Public Health and the Environment, reported 8496 positive COVID-19 tests in the Netherlands on 14 December 2020. Only 333 of these positive tests had 14 December as the date of the statistic – 234 out of these 333 were the date of notification (DON, the date when the positive test result has been reported to the public health service), whereas only 99 were actually positively tested on 14 December (DPL, date of first positive lab result), see Veldhuizen and Van Zelst [2020a]. At the time of writing, 29 December 2020, 12194 people have 14 December as the date of the statistic, of which 10783 have 14 December down as the date of the first symptoms (DOO, date of disease onset), 1061 have 14 December as the date of the positive test result (DPL, date of first positive lab result) and 350 have this date as the date of notification, see Veldhuizen and Van Zelst [2020b].

The above is a natural consequence of how a disease and the testing process works. After the first symptoms it takes some time for a person to get tested (the person may not request a test immediately and even if they do, there may be a waiting time), thereafter it takes some time for the sample to be processed, and finally, it can take some time before the positive lab result ends up in the central database containing all positive test results. What we set out to do in this paper is to nowcast the amount of infections on a given date, based on the actual numbers that have been reported on that date. To illustrate this based on the example above, we aim to estimate the amount of infections that occurred on the 14th of December (combining DOO, DPL and DON) based on the 333 positive tests that have been reported on that date.

The term nowcasting is often used in economics to update economic variables that are only reported after a long delay and subject to revision, such as gross domestic product (GDP). Examples also exist in operational research (warranties), insurance (insurance claims) and of course epidemiology (outbreak

[1] Quantitative Analytics, Cardano.

of diseases). In this paper we will explain how the techniques of Crevecoeur, Antonio and Verbelen [2019], who apply their model to insurance claims, are used to nowcast the amount of COVID-19 infections in the GitHub repository of Lord [2020]. The interested reader is also referred to Bastos, Economou, Gomes, Villela, Coelho, Cruz, Stoner, Bailey and Codeço [2019] and Van de Kassteele, Eilers and Wallinga [2019] for two papers that deal with nowcasting in an epidemiological setting. We may include these two models in future research.

The paper is organised as follows. In Section 2 we describe the setup of Crevecoeur et al. [2019]. In Section 3 we finally describe the specific parametrisation that has been chosen in Lord [2020].

2. Setup of Crevecoeur et al.

Let $N_{t,s}$ denote the number of positive COVID-19 tests that occurred at date t and are reported at time s. The total number of positive COVID-19 tests that have then occurred at date t can be found as:

$$N_t = \sum_{s \geq t} N_{t,s}$$

We assume that:

- The process for N_t follows an inhomogeneous Poisson distribution with intensity λ_t
- The reporting delay is independent and identically distributed for events occurring on t

Let us introduce some additional notation. The number of positive COVID-19 tests that can be observed at time τ are denoted by:

$$N_t^{Obs}(\tau) = \sum_{s=t}^{\tau} N_{t,s}$$

Moreover we introduce $p_{t,s}$, which is the probability of observing an event that occurred on date t on date s. Of course these probabilities have to satisfy the following constraints:

$$p_{t,s} \geq 0$$

$$\sum_{s \geq t} p_{t,s} = 1$$

By $p_t^{Obs}(\tau)$ we denote the probability that an event from date t is observed at date τ, $\tau \geq t$. This probability is equal to:

$$p_t^{Obs}(\tau) = \sum_{s=t}^{\tau} p_{t,s}$$

By the assumptions made above it follows that all $N_{t,s}$ are independent and follow a Poisson distribution:

$$N_{t,s} \sim \text{Poisson}\left(\lambda_t \cdot p_{t,s}\right)$$

When $p_{t,s}$ only depends on the reporting delay, the above is referred to as a chain ladder, see Hachemeister and Standard [1975], Renshaw and Verrall [1998] and Mack [1991, 1993]. Crevecoeur et al. [2019] impose more structure on $p_{t,s}$.

To specify the log-likelihood we will denote all observed data at date τ by the vector $\boldsymbol{\chi}$:

$$\boldsymbol{\chi} = \left\{ N_{t,s} \,\middle|\, t \leq s \leq \tau \right\}$$

and combine all relevant probabilities at this date in vector $\boldsymbol{p}$:

$$\boldsymbol{p} = \{p_{t,s} \mid t \leq s \leq \tau\}$$

With this notation Crevecoeur et al. show the joint log-likelihood of $\boldsymbol{p}$ and $\boldsymbol{\chi}$ can be found as:

$$\ell(\boldsymbol{p}, \boldsymbol{\chi}) = \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot ln(p_{t,s}) - \sum_{t=1}^{\tau} N_t^{Obs}(\tau) \cdot ln\left(p_t^{Obs}(\tau)\right) + O(1)$$

The intensity has already been left out of this equation, its maximum likelihood estimate is:

$$\lambda_t = \frac{N_t^{Obs}(\tau)}{p_t^{Obs}(\tau)}$$

The other parameters can be found by maximising the above log-likelihood numerically. The nowcast for N_t on date τ follows as:

$$\mathbb{E}_\tau[N_t] = \lambda_t = \frac{N_t^{Obs}(\tau)}{p_t^{Obs}(\tau)}$$

2.1. Parametrisation of probabilities

In maximising the log-likelihood we have to ensure the constraints on the probabilities $p_{t,s}$ are satisfied. Crevecoeur et al. ensure this is the case by modelling the discrete reporting delay as a continuous random variable under interval censoring. If we denote U_t as the continuous random variable for a positive test at time t, then when $U_t \in [j, j+1)$, the reporting delay is j days. With this representation, the probability $p_{t,s}$ can be viewed as:

$$p_{t,s} = \mathbb{P}(U_t \in [s-t, s-t+1))$$

Crevecoeur et al. assume $\varphi_t(U_t)$ has a distribution which is independent of the reporting date t. The transformation φ_t is chosen such that:

$$\varphi_t(d) = \sum_{i=1}^{d} \alpha_{t,t+i-1}$$

where all $\alpha_{t,s} \geq 0$ and $\lim_{n\to\infty} \sum_{s=t}^{n} \alpha_{t,s} = \infty$. Crevecoeur et al. refer to these parameters $\alpha_{t,s}$ as the observation exposure. If we assume the cumulative distribution function of $\varphi_t(U_t)$ is given by F, the probability $p_{t,s}$ follows as:

$$p_{t,s} = F(\varphi_t(s-t+1)) - F(\varphi_t(s-t))$$

This structure ensures that the previous constraints on $p_{t,s}$ are automatically satisfied, but we do still have constraints on $\alpha_{t,s}$. As a final step Crevecoeur et al. therefore express $\alpha_{t,s}$ as a function of a vector of covariates $\boldsymbol{x}_{t,s}$:

$$ln(\alpha_{t,s}) = \boldsymbol{x}_{t,s}^T \boldsymbol{\gamma}$$

with $\boldsymbol{\gamma}$ equal to a parameter vector. The vector of covariates $\boldsymbol{x}_{t,s}$ could simply contain indicator functions to capture a day-of-week effect, but could also include other information that one expects the probabilities $p_{t,s}$ to depend on. This parametrisation ensures that $\alpha_{t,s} \geq 0$, and we can quite simply ensure the second constraint is met by setting $\alpha_{t,s} = 1$ for $s > t + \nabla$. The parameter ∇ then takes the role of the largest possible reporting delay.

3. Chosen parametrisation

The currently available nowcasts in Lord [2020] are constructed using a standard Poisson chain ladder, though a slight modification is made to the log-likelihood function in order to attach more weight to more recent observations. We introduce this modification in Section 3.1, and finally discuss the chosen parametrisation in Section 3.2.

3.1. Introducing weights

The setup as described in Section 2 is quite general and allows for time-varying probabilities. In our setup we will opt for a more parsimonious parametrisation, but will attach larger weight to more recent observations by a modification of the log-likelihood function.

The log-likelihood introduced above assigns an equal weight to all past observations. We opt for a minor modification hereof, and assign a weight w_t for each reporting date t, loosely inspired by Majumder, Biswas, Roy, Bandhari and Basu [2020]. The log-likelihood function then becomes:

$$\ell(\boldsymbol{p}, \boldsymbol{\chi}) = \sum_{t=1}^{\tau} w_t \sum_{s=t}^{\tau} N_{t,s} \cdot ln\left(p_{t,s}\right) - \sum_{t=1}^{\tau} w_t \cdot N_t^{Obs}(\tau) \cdot ln\left(p_t^{Obs}(\tau)\right) + O(1)$$

We will specifically choose a weight that exponentially decays the further t is from the current date τ:

$$w_t = e^{-\beta(\tau - t)}$$

with $\beta \geq 0$ equal to the rate of decay.

3.2. Poisson chain ladder

For nowcasting the amount of positive COVID-19 tests we have opted for a simple parametrisation, where the observation exposure is given by:

$$ln\left(\alpha_{t,s}\right) = \delta_{s-t}$$

Finally, the cumulative distribution function F is chosen to be equal to that of a standard exponential distribution, i.e.:

$$F(x) = 1 - e^{-x}$$

This is equivalent to a standard Poisson chain ladder, as this leads to $p_{t,s}$ only depending on the difference of s and t. Together with the weights we introduced in the log-likelihood function in Section 3.1 this gives us a specification that is reasonably parsimonious, while still allowing for its behaviour to change over time. For countries such as the Netherlands this parameterisation is found to work well. For countries such as Germany, where the amount of positive tests varies strongly depending on the day of the week, more structure is necessary. We leave this for future research.

References

BASTOS, L.S., ECONOMOU, T., GOMES, M.F.C., VILLELA, D.A.M., COELHO, F.C., CRUZ, O.G., STONER, O., BAILEY, T. AND C.T. CODEÇO (2019). "A modelling approach for correcting reporting delays in disease surveillance data", *Statistics in Medicine*, vol. 38, no. 22, pp. 4363-4377, available here.

CREVECOEUR, J., ANTONIO, K. AND R. VERBELEN (2019). "Modeling the number of hidden events subject to observation delay", *European Journal of Operational Research*, vol. 277, no. 3, pp. 930-944, available here.

HACHEMEISTER, C.A. AND J.N. STANARD (1975). "IBNR claims count estimation with static lag functions", in Annals of 12th ASTIN Colloqium, International Actuarial Association, Portimão, Portugal, pp. 1-23.

VAN DE KASSTEELE, J., EILERS, P.H.C. AND J. WALLINGA (2019). "Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing", *Epidemiology*, September, vol. 30, no. 5, pp. 737-745, available here.

LORD, R. (2020). GitHub repository, available here.

MACK, T. (1991). "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves", *ASTIN Bulletin*, vol. 21, no. 1, pp. 93-109, available here.

MACK, T. (1993). "Distribution-free calculation of the standard error of chain ladder reserve estimates", *ASTIN Bulletin*, vol. 23, no. 2, pp. 213-225, available here.

MAJUMDER, S., BISWAS, A., ROY, T., BANDHARI, S.K. AND A. BASU (2020). "Statistical inference based on a new weighted likelihood approach", *Metrika*, available here.

RENSHAW, A. AND R. VERRALL (1998). "A stochastic model underlying the chain-ladder technique", *British Actuarial Journal*, vol. 4, no. 4, pp. 903-923, available here.

VELDHUIZEN, E. AND M. VAN ZELST (2020A). Historical RIVM casus dataset for 14 December 2020, available here.

VELDHUIZEN, E. AND M. VAN ZELST (2020B). Historical RIVM casus dataset for 29 December 2020, available here.