Modelling the lagged registers of COVID-19 confirmed deaths in the Mexican health officials daily reports (working paper)

Humberto G.R.

May 2020

1 Introduction

On a daily basis, the Mexican government publishes a data base that includes the records of COVID-19 confirmed death cases. This data base is cumulative, in the sense that each day new records are included. As a consequence of the process with which the health officials collect and test the suspected COVID-19 cases that are included in the data bases, there is a lag between the date of occurrence of the death and the date in which it is registered in the data base. The new confirmed deaths that are included between two dates t_1 and t_2 ($t_1 < t_2$) do not correspond necessarily to the death cases that occurred between these two dates, but they also correspond to deaths that occurred before, i.e., at dates $t \leq t_1$. An implication is that the shape of the curve of cumulative registered death counts at a given date does not correspond to the *real* count, which may lead to wrong interpretations of the state of the pandemic in the country. Here, a model is proposed in order to estimate the lagged death cases and thus provide a better approximation of the state of the pandemic.

NOTE: The models proposed here are models for the process in which the deaths are logged into the data bases, not an epidemiological models.

2 Models

Let N_t be the number of missing records at time t, i.e., the number of unregistered deaths with date of occurrence $t' \leq t$ in the data base at time t. This quantity is latent, as it cannot be directly observed from data. Let $N_{t,l}$ be the number of *new* deceased records with date of occurrence $t' \leq t$ that are included in the data base at time $t + l$. This is, $N_{t,1}$ is the number of deaths before time t that were registered in the data base in the following day, $N_{t,2}$ two days after, etc. The number of missing cases for the data base at time t is given by

$$N_t = N_{t,1} + N_{t,2} + \dots \,. \tag{1}$$

Note that if t^* is the date of the latest data base available, then N_t is partially observed for $t < t^*$. Assuming that $t^* = t + L_t$, then Eq. (1) can be written as

$$N_t = N_{t,1} + N_{t,2} + \cdots + N_{t,L_t} + \dots \tag{2}$$

where $N_{t,l}$ for $l = 1, \dots, L_t$ are observed values.

The main assumption that is made in order to estimate the model is that the proportion of missing cases depends only on the lag l and not on the date of the data base t. This is, $N_{t,l} = \lambda_l N_t$ for all t and thus,

$$N_t = \lambda_1 N_t + \lambda_2 N_t + \cdots + \lambda_{L_t} N_t + \dots \tag{3}$$

with $\sum_l \lambda_l = 1$. The values λ_l can be interpreted as the rate at which the missing registers arrive in the following days. These values need to be estimated. Although the assumption made here may be strong, it allows us to infer the total number of missing cases N_t through the relationship

$$\sum_{l=1}^{L_t} N_{t,l} = N_t \sum_{l=1}^{L_t} \lambda_l \,, \tag{4}$$

with the left-hand side of the equation being fully observable.

Two models are proposed in order to estimate the parameters λ and the latent variables N_t for all t. Both models are based on the multinomial and Poisson distributions; the difference is the assumptions made over the parameters λ.

Model 1 The first model assumes the *logit* functional form for the parameters λ.

$$\begin{aligned}
N_{t,1}, N_{t_2}, \ldots, N_{t,L_t} &\sim Multinomial(\sum_{l=1}^{L_t} N_{t,l} \,, \{\lambda_l\}_{l=1}^{L_t}) \;\; \forall t \\
\sum_{l=1}^{L_t} N_{t,l} &\sim Poisson(N_t \sum_{l=1}^{L_t} \lambda_l) \;\; \forall t \\
\lambda_l &= \frac{e^{\beta l}}{\sum_l e^{\beta l}} \;\; \forall l \\
\text{Priors:} & \\
\beta &\sim \mathcal{N}(\mu, \sigma^2) \\
N_t &\sim Gamma(a, b) \;\; \forall t
\end{aligned}$$

Model 2 This model does not assumes any functional form for the parameters λ, rather they are considered as the outcomes of a Dirichlet distribution.

$$\begin{aligned}
N_{t,1}, N_{t_2}, \ldots, N_{t,L_t} &\sim Multinomial(\sum_{l=1}^{L_t} N_{t,l} \,, \{\lambda_l\}_{l=1}^{L_t}) \;\; \forall t \\
\sum_{l=1}^{L_t} N_{t,l} &\sim Poisson(N_t \sum_{l=1}^{L_t} \lambda_l) \;\; \forall t \\
\text{Priors:} & \\
\{\lambda_l\}_{l=1}^{L_{max}} &\sim Dirichlet(\alpha) \\
N_t &\sim Gamma(a, b) \;\; \forall t
\end{aligned}$$

3 Results

With the available data, it is impossible to know what is the number of missing records for today's data base. However, the data with different cut-off dates allow us to observe and calculate the delay with which new records arrive, and to estimate the number of missing records for previous dates, in this case, for the cut of three days ago (May 24). According to the model, as of May 24, there would be 11,148 deaths from COVID-19 (Figure 1). This number contrasts with the 7,394 total cases registered in the base with that date, and with the 8,392 records accumulated until May 24 on the cut-off date of May 27. Therefore, as of May 27, $2,756 = 11,148 - 8,392$ deaths accumulated until May 24 would be missing, which we would expect to be counted in the coming weeks. Also, it can be seen that the model predicts few missing records for distant dates in the past, but that this number increases as we get closer to today's date.

The expected value of the distribution of λ is presented in Figure 2, where it can be seen that half of the missing cases are expected to arrive around one week after, 75% after 15 days, and 95% 25 days latter.

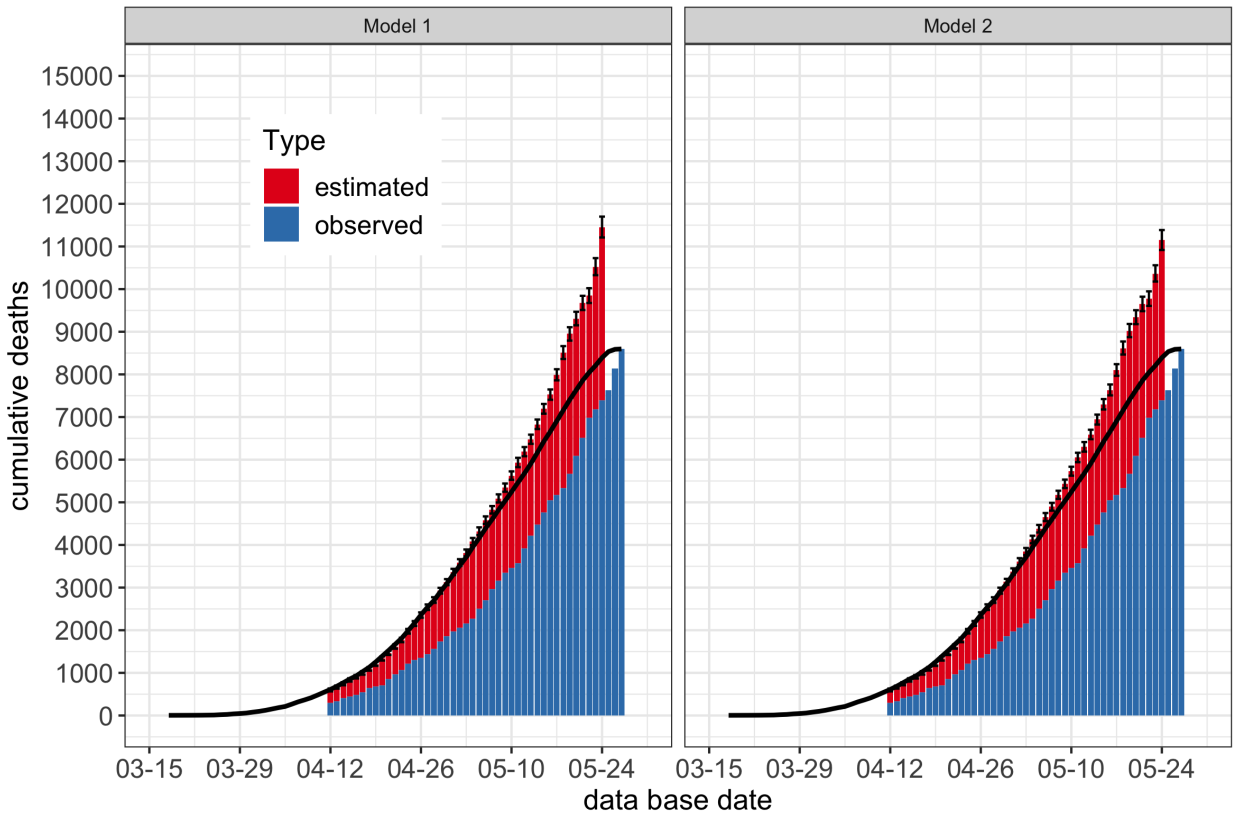

Figure 1: Observed and estimated missing death counts. The blue bars are the cumulative number of deaths as of the cutoff date. In red the number of subregisters estimated by the model with respect to the publication date of each data base. The black line is the number of accumulated deaths on the last cut-off date (May 27). The bars that go beyond the black line are the missing records as of the last cutoff date. For example, by May 7, 2,961 deaths had been reported (in blue), by May 27 that number had been updated to 4,600 (the black line), but according to the model, more deaths will continue to appear prior to that date, until reaching approximately 4,896.

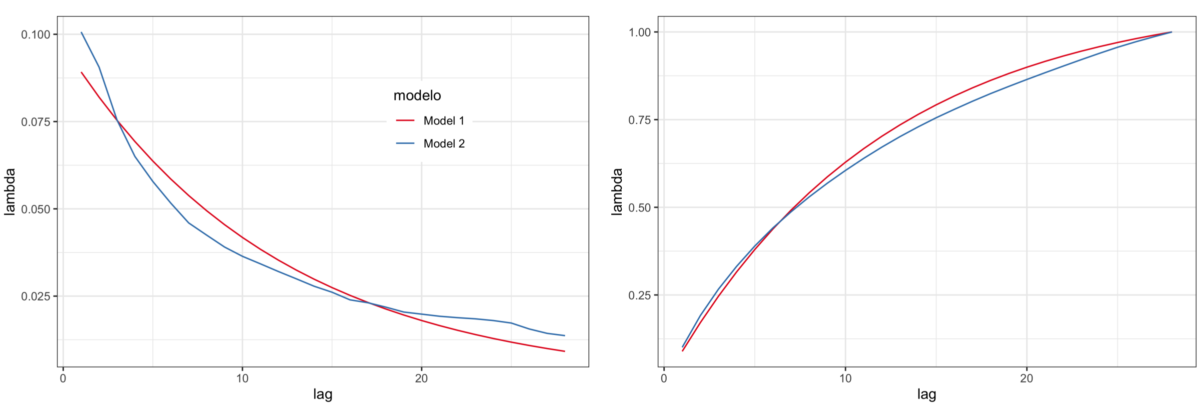

Figure 2: Distribution of λ_l (left) and cumulative distribution (right).

3.1 Validation

In order to see how the models predict the lagged death cases, both models are estimated using the information available in past dates. Then, the predictions are compared against more recent information (May

27th). The results are shown in Figure 3, where it can be seen that the predictions of the models are a better approximation of the *real* counts. There is one issue, however, that the model does not take into account: the number of registers included in the databases changes considerably depending on the day of the week. The new registers for Mondays and Tuesdays are considerably lower than in the rest of the week and, thus, the model subestimates the counts when these days lie in the three day interval between the last observed data for the estimation and the prediction. This may explain part of the variance in the predictions shown in the figure.

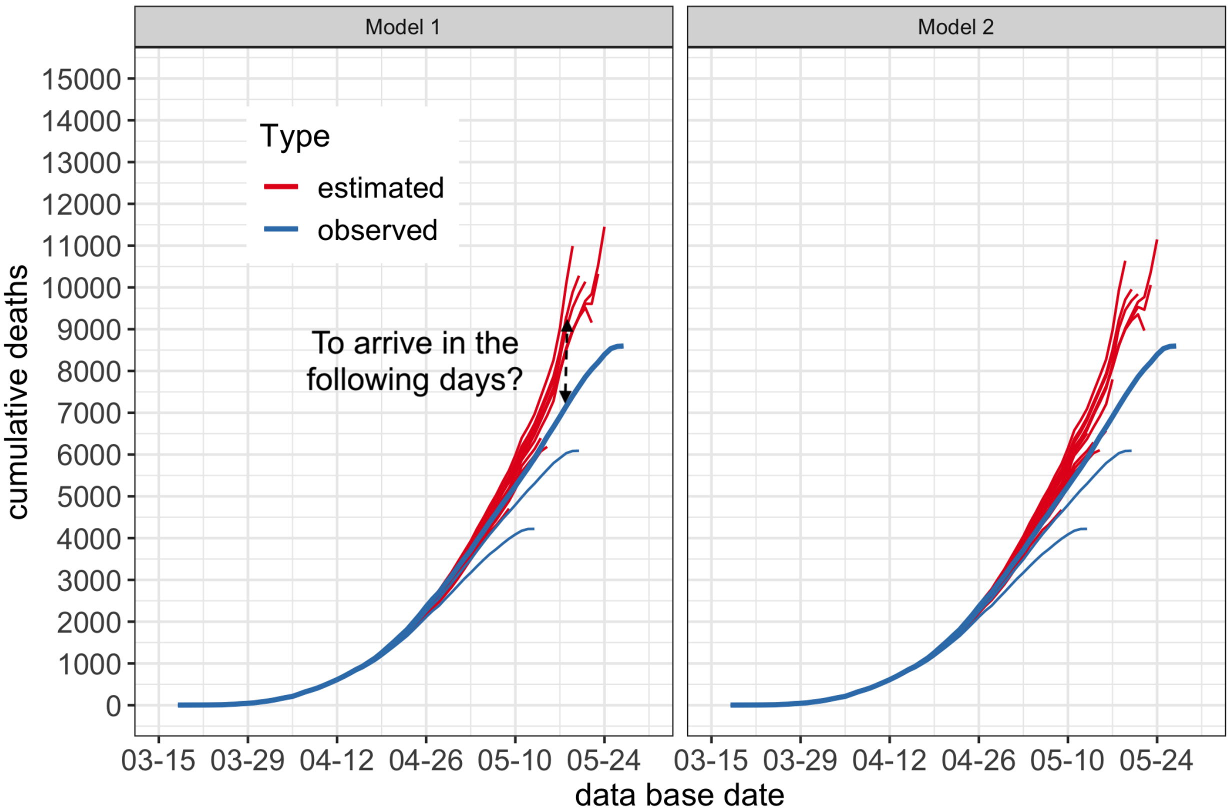

Figure 3: Predicted number of cumulative deaths against the observed cumulative count. It can be seen that that the observed and predicted counts are similar for dates before May 10th. For more recent dates, new data need to be observed in order to evaluate the predictions.