Extract from "Estimating Country-Specific Excess Mortality During the COVID-19 Pandemic"

Victoria Knutson[1], Serge Aleshin-Guendel[1], Ariel Karlinsky[2], William Msemburi[3], Jon Wakefield[1,4]

[1]Department of Biostatistics, University of Washington, Seattle, USA

[2]Hebrew University, Jerusalem, Israel

[3]World Health Organisation, Geneva, Switzerland

[4]Department of Biostatistics, University of Washington, Seattle, USA

The following is an edited extract from a paper that describes the model that was developed to estimate excess mortality for WHO. The full paper (which has been submitted for publication) will be available shortly.

For a small number of countries for which national all-cause mortality (ACM) data are not available (Argentina, China, India, Indonesia and Turkey) we instead have ACM data from subregions, with the number of regions with data potentially changing over time. For other countries we obtain national annual ACM data only, while for China we have subnational monthly and national annual data. In this section we describe the models we use in these situations. For the subnational scenario we construct a statistical model building on, and expanding, a method previously proposed by Karlinsky (2022) that is based on a proportionality assumption.

For Turkey we have subnational monthly data over the complete two years of the pandemic, while for Indonesia we have monthly subnational data for 2020 and for the first six month of 2021. Argentina has observed data for 2020 and subnational monthly data for 2021. For India we have data from up to 17 states and union territories (from now on, states) over the pandemic period (out of 36), but this number varies by month.

We consider the most complex subnational scenario in which the number of regions with monthly data varies by month, using India as an example. For India, we use a variety of sources for registered number of deaths at the state and union-territory level. The information was either reported directly by the states through official reports and automatic vital registration, or by journalists who obtained death registration information through Right To Information requests (the Supplementary Materials of the full paper, contains full details). We stress that for India the global predictive covariate model is *not used* and so the estimates of excess mortality are based on data from India only.

We assume in total that there are up to K regions that contribute data at any time. We develop the model for a generic country. For the historic data in month t we have total death counts along with counts over regions, which we denote as $Y_{t,k}$, $k \in K_t$, so that in period t, $|K_t|$ is the number of regions that provide data with $k \in K_t$ being the indices of these areas from $1, \dots, K$. We let region 0 denote all other regions, which are not observed in pandemic times, and $S_t = \{0, K_t\}$ at time t. To motivate our model, we assume, in month t:

$$Y_{t,k}|\lambda_{t,k} \sim \text{Poisson}(N_{t,k}\lambda_{t,k}), \qquad k \in S_t,$$

where $N_{t,k}$ is the population size, and $\lambda_{t,k}$ is the rate of mortality. Hence,

$$Y_{t,+}|\lambda_{t,k}, k \in S_t \sim \text{Poisson}\left(\sum_{k \in S_t} N_{t,k}\lambda_{t,k}\right).$$

If we condition on the total deaths, we obtain,

$$\boldsymbol{Y}_t|\boldsymbol{p}_t \sim \text{Multinomial}_{|S_t|}(Y_{t,+}, \boldsymbol{p}_t),$$

with $\boldsymbol{p}_t = \{p_{t,k}, k \in S_t\}$, with

$$p_{t,k} = \Pr(\text{ death in region } k \mid \text{month } t, \text{ death }) = \frac{N_{t,k}\lambda_{t,k}}{N_{t,+}\lambda_{t,+}},$$

Our method hinges on this ratio being approximately constant over time. If, over all regions, there are significant changes in the proportions of deaths in the regions as compared to the national total, or large changes in the populations within the regions over time, then the approach will be imprecise. We stress that it is overall deviations for the totality of states that are important – some states may have a greater proportion of deaths during pandemic months, but others may have a smaller proportion. Of course, in practice we do not know for sure whether the assumption remains reasonable over the pandemic. To address this, we carry out extensive sensitivity (for example, we remove different subsets of states and run the model) and cross-validation analyses – these are fully reported in the Supplementary Materials.

We model the monthly probabilities as,

$$\log\left(\frac{p_{t,k}}{p_{t,K_t+1}}\right) = \alpha_k + e_t, \qquad k \in S_t, \tag{1}$$

where the α_k parameters are unrestricted and $e_t \sim \mathrm{N}(0, \sigma^2_\epsilon)$, and we examine the size and temporal structure of the error terms e_t, to assess the proportionality assumption, at least over the available pre-pandemic period.

To specify the model, we take a multinomial with a total number of categories that corresponds to all regions that appear in the data, K, and specify the likelihood over all months by exploiting the property that a multinomial collapsed over a subset of cells is also multinomial. Hence, in year t we have a multinomial with $|K_t|+1$ categories with constituent probabilities constructed from the full set of $K+1$ probabilities.

To derive the predictive distribution for the total deaths in the pandemic, we abuse notation and let $Y_{t,1}$ denote the total number of observed subnational deaths at time t, and $Y_{t,2}$ the total number of unobserved subnational deaths at time t, with $Y_{t,+} = Y_{t,1} + Y_{t,2}$ being the total (national) number of deaths at time t. Hence, at time t,

$$Y_{t,1} | p_t, Y_{t,+} \sim \mathrm{Binomial}(Y_{t,+}, p_t),$$

where $p_t = \sum_{k \in K_t} p_{t,k}$. In order to fit the multinomial model in a Bayesian framework and predict the total number of deaths in 2020–2021, we need to specify a prior for $Y_{t,2}$ or, equivalently, for $Y_{t,+}$, where t indexes months in this period. We will use the prior $p(Y_{t,+}) \propto 1/Y_{t,+}$, which is a common non-informative prior for a binomial sample size (Link, 2013), and has the desirable property that the posterior mean for $Y_{t,2}$, conditional on p_t, is $\mathrm{E}[Y_{t,2}|p_t] = Y_{t,1}(1-p_t)/p_t$, i.e., of the same form as the simple frequentist "obvious" estimator, which leads to the naive estimate of the ACM, $Y_{t,1} + \widehat{Y}_{t,2} = Y_{t,1}/p_t$.

To give more details for implementation we will use a general result. Suppose

$$\begin{aligned} Y_{t,1} | Y_{t,+}, p_t &\sim \mathrm{Binomial}(Y_{t,+}, p_t) \\ p(Y_{t,+}) &\propto 1/Y_{t,+}, \end{aligned}$$

so that, in particular, the marginal distribution of Y_{+t} does not depend on p_t. Then, the posterior for the missing ACM count, conditional on p_t, is

$$Y_{t,+}|Y_{t,1}, p_t \sim Y_{t,1} + \mathrm{NegBin}(Y_{t,+}, 1 - p_t),$$

or, equivalently,

$$Y_{t,+} - Y_{t,1}|Y_{t,1}, p_t \sim \mathrm{NegBin}(Y_{t,1}, 1 - p_t).$$

This links to one of the usual motivations for a negative binomial (the number of trials until we observe a certain fixed number of events) — making inference for the number of total deaths it takes to produce $Y_{t,1}$ deaths in the sub-regions. We implement this model in `Stan`. In the Supplementary Materials we detail a simulation study that validates the method in the situation in which the missing data follow the assumed form.

References

Karlinsky, A. (2022). Estimating national excess mortality from subnational data: application to Argentina. *Revista Panamericana de Salud Publica.*

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial N. *Ecology*, 94:2173–2179.