Extended abstract submitted at the European Population Conference
Edinburgh, 12-15 June 2024

Cause-specific Mortality Forecasting. A Constrained Penalized Regression Model

Carlo G. Camarda[1] and María Durbán[2]

[1]Institut national d'études démographiques, Aubervilliers, France
[2]Department of Statistics, Universidad Carlos III de Madrid, Spain

October 23, 2023

Abstract

Cause of death data provides additional insight on the future trends of mortality, as well as provide valuable information for governments and insurance companies. Models that fit and forecast by cause of death come across several methodological problems, one of them being the inconsistency between individual estimation and forecast of mortality per cause of death and an all-cause scenario. We propose a clear-cut and fast method to obtain coherent cause-specific mortality trajectories based on Lagrange multipliers. We apply the method proposed to fit and forecast mortality of males in USA for the most five leading causes of death.

Keywords: Cause of death; Constraints; Forecasting; Mortality; Penalized Likelihood

1 Introduction

Overall mortality trends are the summation of cause-specific mortality experiences. Consequently modelling and forecasting changes in cause of death patterns allows us to recognize the drivers of all-cause mortality and identify emerging health challenges. On the one hand, early literature has argued that all-cause mortality projections based on cause-specific mortality present serious drawbacks (Wilmoth, 1995). On the other, some approaches for forecasting cause-specific mortality has been recently proposed, though either based on the Lee-Carter model and for specific cause (Kjærgaard et al., 2019) or on a Bayesian hierarchical model aiming to forecast cause-specific death rates for geographic subunits (Foreman et al., 2017).

When dealing with cause-specific mortality, we need to ensure that cause-specific deaths must sum to the total number of deaths. In the following, we model log-mortality in a Poisson setting for each cause assuming only smoothness over age and time. The summation constraint thus become non linear with respect to the estimated coefficients. We embed the whole approach in a Generalized Linear Array Model framework (GLAM, Currie et al., 2006) in which Lagrangian multipliers are iteratively updated to enforce constraints.

2 The model

We have deaths, and exposures to the risk of death, arranged in two three-dimensional arrays, $\boldsymbol{Y} = (y_{ijk})$ and $\boldsymbol{E} = (e_{ijk})$, each $m \times n \times k$, whose rows, columns and layers are classified by age

at death ($\boldsymbol{a}$), year of death ($\boldsymbol{t}$) and cause of death ($\boldsymbol{c}$). For ensuring coherence in the model, the final layer of $\boldsymbol{Y}$ contains total number of deaths (that is, k =number of causes of death+1). Note that each layer in $\boldsymbol{E}$ includes the same age-year matrix: we are in a competing risk setting. We assume that the number of deaths y_{ijk} is Poisson distributed with mean $\mu_{ijk}e_{ijk}$. The value of μ_{ijk}, commonly named force of mortality, is the object of all mortality models.

In the following we will illustrate the method for United States, males by age-groups 30-34, ..., 95-99, 100+, years 1978-2018 and the following five coherent groups of causes of deaths: Cardiovascular diseases, Neoplasms, External causes, Diseases of the respiratory system and Other diseases (Human Cause-of-Death Database, 2023). We forecast total and cause-specific mortality up to 2040.

With $k-1$ causes of death, we deal with a three-dimensional setting. The vectorized linear predictor is given by $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\alpha}$ where the design matrix is $\boldsymbol{B} = \boldsymbol{I}_k \otimes \boldsymbol{B}_t \otimes \boldsymbol{B}_a$. We use a rich basis of B-splines for age and year, and smooth surfaces are then obtained by marginal penalization. With no summation constraint the model simply reduces to a series of two-dimensional GLAMs.

To hinder singularity issues in the resulting scoring algorithm, we enforce our constraint for a large number of equally-spaced data-points. Smoothness will guarantee the remaining coherence between cause-specific and overall mortality. The penalized 3D GLAM is then subject to

$$\boldsymbol{C}\exp(\boldsymbol{\eta}) = \boldsymbol{0} \tag{1}$$

where $\boldsymbol{C}$ sums up, for the large selected number of age and years, cause-specific deaths and then subtract the associated total number of deaths. The matrix $\boldsymbol{C}$ can be written as a Kronecker product $\boldsymbol{C} = \boldsymbol{C}_c \otimes \boldsymbol{C}_t \otimes \boldsymbol{C}_a$ and therefore, as for the linear functions and the inner products within the scoring algorithm, it can be included as a sequence of nested matrix operations in a GLAM framework.

The use of Lagrange multipliers $\boldsymbol{\omega}$ for each age-year ensures that the constraint is enforced, yielding the following constrained penalized Poisson log-likelihood:

$$\ell_P = \boldsymbol{y}'\boldsymbol{B}\boldsymbol{\alpha} - \boldsymbol{e}'\exp(\boldsymbol{B}\boldsymbol{\alpha}) - \frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{P}\boldsymbol{\alpha} - \boldsymbol{\omega}'\boldsymbol{C}\boldsymbol{E}\exp(\boldsymbol{B}\boldsymbol{\alpha}). \tag{2}$$

We compute the derivatives of (2), and by means of Newton-Raphson we find the following scoring algorithm:

$$\begin{bmatrix} \boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{B} + \boldsymbol{P} + \boldsymbol{B}'\mathtt{diag}(\boldsymbol{C}'\tilde{\boldsymbol{\omega}})\tilde{\boldsymbol{V}}\boldsymbol{B} & \boldsymbol{B}'\tilde{\boldsymbol{V}}\boldsymbol{C}' \\ \boldsymbol{C}\tilde{\boldsymbol{V}}\boldsymbol{B} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\alpha}} \\ \tilde{\boldsymbol{\omega}} \end{bmatrix} =$$

$$\begin{bmatrix} \boldsymbol{B}'\tilde{\boldsymbol{W}}\boldsymbol{z} + \boldsymbol{B}'\mathtt{diag}(\boldsymbol{C}'\tilde{\boldsymbol{\omega}})\tilde{\boldsymbol{V}}\boldsymbol{B}\tilde{\boldsymbol{\alpha}} \\ \boldsymbol{C}\tilde{\boldsymbol{V}}\boldsymbol{B}\tilde{\boldsymbol{\alpha}} - \boldsymbol{C}\boldsymbol{\gamma} \end{bmatrix} \tag{3}$$

where $\boldsymbol{\gamma} = \exp\boldsymbol{\eta}$, $\boldsymbol{V} = \mathtt{diag}(\boldsymbol{\gamma})$. $\boldsymbol{W}$ and $\boldsymbol{z}$ are the Poisson regression weights and working response, respectively. The penalty $\boldsymbol{P}$ ensures smoothness over age and time for each cause and it has a block-diagonal structure.

In order to handle $k-1$ causes of death across different age groups and years, we face the challenge of optimizing $2 \cdot k$ smoothing parameters. To avoid dealing with such a high-dimensional optimization problem, we decided to utilize the smoothing parameters optimized by BIC when estimating each cause-specific age-time matrix independently.

Confidence intervals for the estimated mortality surface are calculated by stepping into the Bayesian framework, therefore:

$$\mathrm{Var}(\boldsymbol{B}\hat{\boldsymbol{\alpha}}) = \boldsymbol{B}\boldsymbol{R}\boldsymbol{B}',$$

where $\boldsymbol{R}$ is the top left block of the inverse of the matrix on the left hand side of (3). Of course, if there are no constraints, i.e $\boldsymbol{\omega} = \boldsymbol{0}$, we obtain the usual expression of the variance.

As in Currie et al. (2004) we treat forecasting as a missing value problem and we add shape constraints to enforce future cause-specific mortality patterns to lie within a range of valid profiles computed from observed trends (Camarda, 2019).

3 Results

We fitted the proposed model to the US male mortality data described in the previous section. The left panel of Figure 1 shows actual, estimated and forecast log-mortality for a selected age (50) over years along with their 95% confidence intervals. The proposed model is able to well described historical cause-specific patterns as well as to reasonably extrapolate them into the future. The right panel of Figure 1 presents estimated death counts for a specific year (2000). Here one can easily acknowledge the equality between the sum of cause-specific deaths and the total number of deaths which is enforced by the non linear set of constraints in (1). Equally satisfactory outcomes are achieved for all ages, years and causes of deaths.

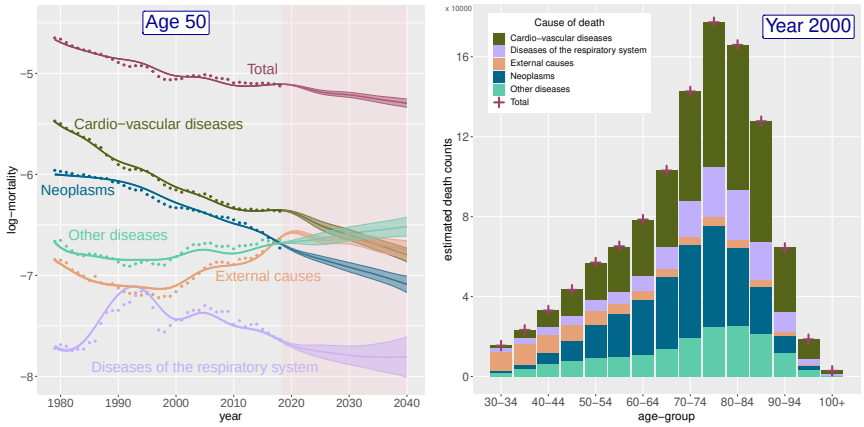

Figure 1: Left panel: observed, estimated and forecast cause-specific mortality rates over years for age 50 (log-scale) along with associated 95% confidence intervals. Right panel: Estimated cause-specific death counts over age-groups for year 2000. In both panels, total mortality and total number of deaths are plotted. USA, males, ages 30-100, years 1978-2018, forecast up to 2040.

In Figure 2, a commonly used summary indicator is presented: life expectancy, here at the starting age of 30. We compare our model with a estimates obtained on overall mortality using P-splines with shape constraints (Camarda, 2019). While the fitted values for the observed time-window are practically identical, noticeable differences emerge when making forecasts. When accounting for cause-specific patterns and ensuring that cause-specific deaths add up to the total number of deaths, the modeling of mortality yields slightly more pessimistic prospects: remaining life expectancy in 2040 at age 30 is forecast to 50.67 years in our model compared to 51.16 years in the alternative approach.

Of even greater significance is the notable reduction in the confidence intervals, indicating a significant decrease in future uncertainty around overall mortality when cause-specific trends are incorporated into the model. This outcome was expected, given the inclusion of substantial information through the incorporation of what we referred to as the summation constraint.

4 Conclusions

In the presented study, we propose a novel approach to model and forecast cause-specific mortality. By combining penalized likelihood and iteratively computed Lagrangian multipliers we obtain smooth cause-specific mortality surfaces over ages and time, and we simultaneously enforce the necessary constraints in this setting: cause-specific deaths sum to the total number of deaths. Forecasting comes naturally in this setting and constraints are satisfied into future years, too. Additional shape are necessary to demographically informed projected patterns.

In the example analysis of US mortality, focusing on the five leading causes of death, the proposed model yields lower overall future life expectancy compared to similar models. Additionally, it exhibits remarkably narrower future uncertainty.

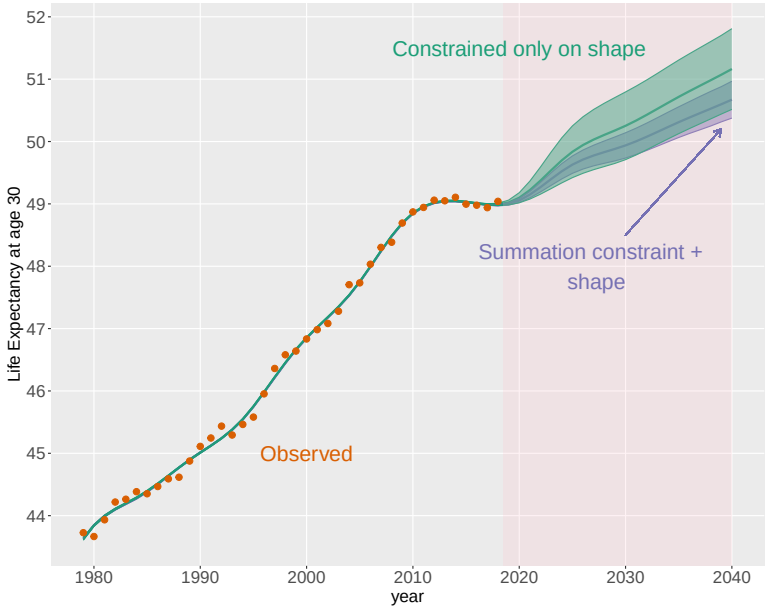

Figure 2: Observed, estimated and forecast life expectancy at age 30 along with associated 95% confidence intervals. Proposed model estimating cause-specific and overall patterns with summation and shape constraints is compared with a simpler approach without summation constraint. USA, males, ages 30-100, years 1978-2018, forecast up to 2040.

Moving forward, we intend to investigate other scenarios where coherence constraints may be required, such as mortality by region, sex, and other factors. Furthermore, we have plans for a comprehensive validation study to evaluate the accuracy of future point estimates and the coverage of associated confidence intervals.

Last, Covid-19 pandemic clearly taught us that forecast future mortality by extrapolating past trends can no longer be tacitly assumed. Accounting for the impact of such short-term shocks will be a clear challenge for further mortality forecasting research as well as for our model.

References

Camarda, C. G. (2019). Smooth Constrained Mortality Forecasting. *Demographic Research 41*(38), 1091–1130.

Currie, I. D., M. Durbán, and P. H. C. Eilers (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling 4*, 279–298.

Currie, I. D., M. Durbán, and P. H. C. Eilers (2006). Generalized Linear Array Models with Applications to Multidimensional Smoothing. *Journal of the Royal Statistical Society. Series B 68*, 259–280.

Foreman, K. J., G. Li, N. Best, and M. Ezzati (2017). Small area forecasts of cause-specific mortality: application of a Bayesian hierarchical model to US vital registration data. *Journal of Royal Statistical Society. Series C 66*, 121–139.

Kjærgaard, S., Y. E. Ergemen, M. Kallestrup-Lamb, J. Oeppen, and R. Lindahl-Jacobsen (2019). Forecasting causes of death by using compositional data analysis: the case of cancer deaths. *Journal of Royal Statistical Society. Series C. 68*(5), 1351–1370.

Wilmoth, J. R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies 5*, 293–319.