A new model life table with an optimal number of age groups

Ibraheem Ahmed- European Doctoral School of Demography
Julio Romero Prieto- London School of Hygiene and Tropical Medicine

Abstract (150 words)
For countries with inadequate vital registration systems accurate measurements of mortality are difficult. In the absence of reliable mortality data, we rely on model life tables to simulate the real mortality experience of these populations. We propose a new model which simplifies and improves upon earlier approaches. This model requires only a small number of key optimised ages to recover the entire mortality schedule and utilise multiple eigenvectors to explain the error term of the model. Variations in these two features provide models that are suitable for in and out of sample predictions. This model is able to correctly synthesise the general pattern of mortality across a range of scenarios and outperforms the existing Log-Quad and SVD-Comp models by a significant margin. Therefore, this new model's vastly superior veracity in all settings makes it ideal for application in data deficient settings compared with the existing alternatives.

Background and Introduction

For the majority of developing countries, accurate collection and reporting on vital statistics data is still incomplete or simply unavailable. The resulting life tables that can be estimated from imperfect data cannot provide reliable estimates of the true level and mortality profile of these populations. Hence, model life tables have been developed in order to estimate the mortality regimes schedule and the resulting life expectancy of countries with unreliable data.

Current models are over-parameterised and include at least 2 coefficients for each age group creating redundancy which resulted in a number of computationally heavy regression models that are not practical in many applications. These models utilise data on all age groups which results in them being subject to more data issues, increasing variability and decreasing confidence in their predictions. Moreover, existing models are not very flexible and do a poor job at explaining the full variability in the data. This makes them insufficient for predicting the mortality patterns in settings not used for calibration of the model- typically low-income settings–i.e., out of sample prediction. The flexibility of the model life tables relies on the eigenvectors resulting from a Singular Value Decomposition (SVD) of the covariance matrix of errors. Heuristically, current approaches are limited to the first vector (Wilmoth et al., 2012; Clark, 2019) and do not have a satisfactory fitting to some populations with very particular age patterns of mortality. Hence the development of a model that utilises multiple eigen-vectors from the SVD should do a better job at modelling mortality in these settings which can lead to more accurate understanding of the true patterns of mortality.

In this paper we propose a new simpler method for countering the aforementioned issues. We identify the set of ages that generate the most efficient model life table–i.e., optimisation of the age groups through minimising the Root Mean Squared Error (RMSE). Optimisation will enable the identification of the age points of inflexion within the mortality schedule. From these points we can generate the full empirical model life table provided the model efficiently makes use of the identified age components. From this perspective, model life tables can be more concentrated in a few key age groups which would provide more parsimonious representations of the human mortality. Similarly, the efficient choice of multiple eigenvectors is increasing the accuracy of the model's predictions.

Data Description

The model has been fitted exclusively using life tables obtained from the Human Mortality Database (HMD, www.mortality.org). Period life tables containing year by year death and exposure-time (person-year) counts by single year age groups were extracted from the HMD and used to calculate mortality rates from which ${}_nq_x$ could be inferred.

Recent model comparisons have assessed efficacy against out-of-sample life tables from Mexico, 1983-1985 and South Africa, 2005 (Clark, 2019). The model presented here was also assessed using these lifetables for

comparison with existing contemporary models. These can be found in the Human Life-Table Database (HLD, www.lifetable.de). The comparison year for South Africa is 2005 but as this year is not found in the database, the next closest year, 2006, was selected to give a good comparison to models presented in previous literatures.

Methods

The model developed for modelling the relationship between age and probability of mortality ${}_nq_x$, follows the basic form:

$$y_{ages \times 1} = a_{ages \times 1} + (u \cdot k)_{ages \times 1}$$

Where **y** = $\log({}_nq_x)$, **a** is a vector of mean responses from $\log({}_nq_x)$ across the HMD training dataset, **ages** is the length of the vector of fixed age points to be selected from the optimisation or those pre-fixed, **u** is the matrix (of dimension ages x ages) of orthonormal eigenvectors resulting from the SVD of the error between y and a, and **k** is the vector of constants (of dimension ages x 1) calculated from fitting the model life table- mathematically this is calculated as the product of the transpose of u and the error y-a.

This can be expanded to:

$$y_{ages \times 1} = a_{ages \times 1} + u_x^1 \cdot k_1 + u_x^2 \cdot k_2 + \cdots + u_x^{ages} \cdot k_{ages}$$

The model can be limited to the first **n** eigenvectors by only including the $u \cdot k$ terms up to and including $u_x^{\boldsymbol{n}} \cdot k_{\boldsymbol{n}}$ such that $\boldsymbol{n} \in [1, ages]$

For selecting the optimal age groups, combinatorial methods were employed. The optimal ages were evaluated as the set of ages that minimised the mean RMSE for all life tables in the training dataset calculated as the mean response of $\sqrt{\sum_1^{100} \frac{e^2}{n}}$ Where **e** is the error of prediction between the real and predicted central death rates in logs, **n** is the number of identifiable usable ages within each life table in the training dataset. Prior to optimisation, the ages 1, 5, 80, 90 and 95 were prefixed and the 'free' ages to be optimised were selected from the interval from [5,80] under the conditions that they were: a) a multiple of 5 (in keeping with 5-year age convention), b) no closer than 10 years from any other free age selected and from the boundaries of the interval.

Once the 'free' ages have been selected and y predicted at these points using the model, the predictions at the remaining ages were obtained via monotonic interpolation by a piecewise cubic function between each of the free ages following the approach found in 'A simple method for monotonic interpolation in one dimension' (Steffan, 1990). The predictions for each life table were compared to the true values in the training dataset and the RMSE computed. By exhausting each possible combination of free ages under the specified constraints, we identified the optimised ages to fix in our model when making out of sample predictions.

One addition is the use of the Gompertz curve to smooth mortality at older ages through the assumption that mortality at older ages follows a double exponential function. Here the Gompertz transformation is applied to values between 80 and 100 years using the values of q(x)* obtained from fitting the model. This is to ensure no disjointedness between the Gompertz curve and the interpolated values. The Gompertz function for the training dataset was programmed using the ages 80, 90 and 95 as inputs. The Gompertz adjusted ages were used to replace the interpolated ages giving the final model.

The model can be varied in two areas to ensure the best fit: In the number of eigenvectors from the matrix U resulting from the SVD or, in the number of key ages. The model can be refitted to any number of eigenvectors up-to the total number of fixed and free ages, k, and for any number of 'free' ages. For the number of 'free' ages, a selection of models containing 2 to 6 free ages was made to make comparisons across a number of fitted models and this proved sufficient for this analysis. Hence the upper limit for the number of eigenvectors generated for any one model was 11 for the model with 6 free ages.

The final model can be used in the following manner:

1) Convert data into the form ${}_nq_x$ – this is the input for the model.
2) If the life table you wish to model concludes before the age of 100, use the Gompertz function to extend the life table until age 100. Here we used age inputs of 80, 90 and 95 to inform the Gompertz function but in the absence of these ages in the life table, these ages can be changed as needed to fit the function for the given life table
3) Identify the number of free ages and eigenvectors for which you would like to run the model. The corresponding set of optimised ages for each combination of free ages and eigenvectors by sex is located in appendix 2
4) Use the set of Optimised ages identified in step 3 to calculate the MLT coefficients using the model on the training dataset
5) Use the MLT coefficients obtained to generate predictions of y using the life table you wish to model using the model
6) Convert the predictions back to q(x). Interpolate the intermediary coefficients and fit the Gompertz to the prediction at ages [80, 90, 95] to obtain the full Model Life table

To test the efficacy of the model developed and enable comparisons across models, life table data from Mexico (1983-1985) and South Africa (2006-2008) was used. This is in keeping with Sam Clark's 2012 paper, where his 'SVD Comp' model was tested against the 'Log-Quad' model (Wilmoth et al, 2012).

Preliminary Results

		Number of Eigenvectors						
		5	6	7	8	9	10	11
Number of Free Ages	2	0.21267	0.20774	0.20670	NA	NA	NA	NA
	3	0.16451	0.15635	0.15332	0.15226	NA	NA	NA
	4	0.14605	0.13595	0.12962	0.12650	0.12525	NA	NA
	5	0.14116	0.13206	0.12587	0.12284	0.12081	0.11981	NA
	6	0.13786	0.12930	0.12371	0.12100	0.11835	0.11762	0.11668

Table 1: Mean RMSE from the training dataset for each combination of free ages, optimised via combinatorial optimisation, and eigenvectors in the model life table for both sexes combined

For the Combinatorial optimisation, the number of eigenvectors were also varied introducing measurement of the error of MLT prediction. Table 1 details the minimised mean RMSE arising from the optimal model for each combination of eigenvectors and free ages. The results generated show that the greater the parameterisation of the model, through increasing the number of eigenvectors used, and/or the greater the number of optimised/free ages, the greater the minimisation off the error. It was also observed that for a given number of free ages, the selection of free ages themselves remain largely unchanged as additional eigenvectors are added.

$\text{Log}({}_nm_x)$ is the most difficult measure for models to predict. In measuring so we are able to test the efficacy of the model. To observe the extent to which these models can correctly predict variation within different populations, extreme scenarios from outside of the sample can be used to illustrate the models predictive power. Figure 1 details a mortality profile during the height of the HIV epidemic in South Africa, where the mortality profile differs significantly from those observed in the training dataset.

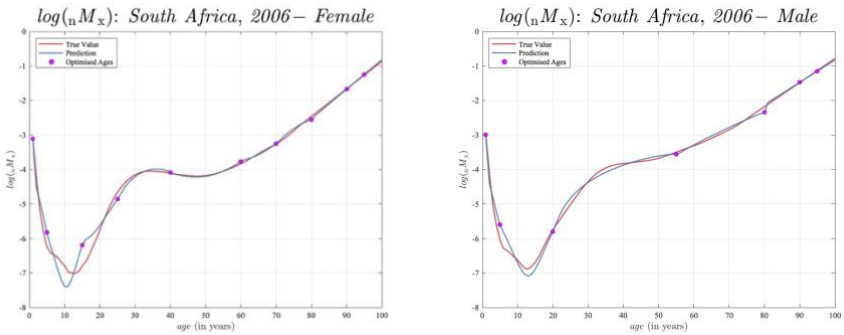

Figure 9: Best fitted Model Predictions for South African Life Tables, 2006-2008, by Sex

Female		Number of Eigenvectors						Male		Number of Eigenvectors					
		5	6	7	8	9	10			5	6	7	8	9	10
Number of Free Ages	2	0.21520	0.20062	NA	NA	NA	NA	Number of Free Ages	2	0.12921	0.12499	NA	NA	NA	NA
	3	0.28465	0.27819	0.27743	NA	NA	NA		3	0.32704	0.26944	0.26792	NA	NA	NA
	4	0.28841	0.29245	0.22939	0.20508	NA	NA		4	0.26641	0.14593	0.14562	0.14417	NA	NA
	5	0.37378	0.33091	0.19982	0.19211	0.18182	NA		5	0.20076	0.14884	0.14800	0.14354	0.14314	NA
	6	0.33438	0.33564	0.21349	0.19145	0.18284	0.18183		6	0.19475	0.18591	0.18351	0.17458	0.17381	0.17341

Table 4: Table enumerating RMSE resulting from model predictions for each model variation on South African life table data, 2006-2008, by Sex

In this example we see fit improves as number of eigenvectors are increased but can sometimes deteriorate when additional 'free' ages are added. Interestingly, the models that minimise RMSE for these life tables are not the most complex. Whilst for any number of free ages the models with the most eigenvectors are always minimising, it is not necessary to have 6 free ages to produce the best fit. For females, the 5 free age model is as good as the 6 free age model. For males the favoured model is in-fact the 2 free age model.

Despite the lower presence of young adult mortality bulges in the female training dataset, the optimal model is able to accurately reproduce this feature as well as mortality at subsequent ages. The only shortcoming is the overprediction of the minimum in child mortality and a subsequent slight overprediction in mortality shortly after until the age of 20. In Males, the 2 free age model performs optimally, benefitting from the lack of flexibility to almost regenerate the true mortality profile completely. The bulge in mortality in young adults is smoothed slightly as a consequence but still closely captured. Alternative models with 4 or 5 free ages provide a better fit at capturing this bulge but lose accuracy in overpredicting the minimum in child mortality.

Further Analysis

Further research can be conducted to explore the possibilities of this model. These may include alternative choices of splines, different approaches to optimisation and different choices for fixed ages. For practical usage, on account of there being multiple variations of the model presented here, these variations can be tested against different populations with reliable datasets to identify the best fitted model in particular settings for estimation of like scenarios.

References

1. Wilmoth J, S Zureick, V Canudas-Romo, M Inoue, and C Sawyer (2012). "A flexible two-dimensional mortality
model for use in indirect estimation." Population Studies 66(1):1-28.
2. Clark SJ (2019). "A general age-specific mortality model with an example indexed by child mortality or both child
and adult mortality." Demography 56(3):1131-1159.
3. Max Planck Institute for Demographic Research, University of California, Berkeley, & Institut d'études demographiques (INED). (n.d.) Human life table database [Data set]. Retrieved from https://www.lifetable.de/data/hld.zip
4. Steffen M (1990). A simple method for monotonic interpolation in one dimension. Astronomy & Astrophysics 239:443-450.