Explanatory Models of Changes in Life–Table Densities of Deaths for U.S. States 1980–2018.

Proposal for NBER conference 'The Determinants of Mortality'

Jim Oeppen* and Marie–Pier Bergeron–Boucher*

July 29, 2024

1 Introduction

Demographic analysis of the age-specific components of life tables has been almost entirely endogenous. If exogenous variables such as smoking or wealth are considered, it is usually in relation to a single demographic indicator such as period life expectancy at birth, infant mortality, or an index of lifespan inequality. A classic example is the model of life expectancy at birth as a function of time and GDP p.c. presented by Preston (1975).

The innovation in this paper is that we estimate the *age–specific* associations between the distribution of life table deaths, d_x, and possible explanatory variables. The regression equations allow the user to estimate and forecast 'designer life tables' that match any given values for the covariates. The method is illustrated by an analysis of potential explanatory variables associated with the age-specific changes in State life tables from the United States Mortality Database, across the years before the opioid crisis (1980-89) and during it (2010-18).

2 Methodology

The d_x in a life table must sum to its radix (l_0), *i.e.* the assumed number of deaths, equal to births, in the zero–growth synthetic life table population. To ensure that a regression model preserves the constrained structure of the data within this constant–sum constraint, analysis of d_x should consider the principles of Compositional Data Analysis (CoDA) initiated by Aitchison (1986). A 'composition' is any positive, discrete distribution where the 'parts' are constrained to sum to a constant. The CoDA approach allows one to treat life table densities of deaths (d_x) by age as dependent entities in a regression model.

In CoDA, a simple linear regression model can be written, with appropriate compositional operators, as

$$\mathbf{ilr}(\mathbf{Y}_i) = \mathbf{a} \oplus X_i \odot \mathbf{b} \oplus \epsilon_i, \tag{1}$$

where $\mathbf{ilr}(\mathbf{Y}_i)$ is a composition for case i in log–ratio space, X_i is an explanatory variable, and ϵ_i is a residual composition. $\mathbf{a}$ and $\mathbf{b}$ are unknown constant parameter compositions to be estimated. The dependent composition Y can be a d_x density, or a compositional difference between two densities $Y_2 \ominus Y_1$, known as a 'perturbation', allowing models of level and change. Extension of equation 1 to multiple regression follows the expected form.

*Interdisciplinary Centre on Population Dynamics, University of Southern Denmark. e-mail: joeppen@sdu.dk

3 A Univariate Illustration

To provide the simplest illustration of how CoDA regression works, we limit the structure to a simple–regression model of the perturbation, or changes, in the male life–table density of deaths between the life tables for 1980-89 and 2010-2018, as a function of the Foreign–Born percentage in each State. The full paper uses multiple regression. All methods, statistics, and smoothing in the paper are population–weighted.

3.1 Parameter Estimates

In Figure 1, the dashed, black line is the uniform or neutral composition summing to unity, which is analogous to zero in conventional arithmetic.

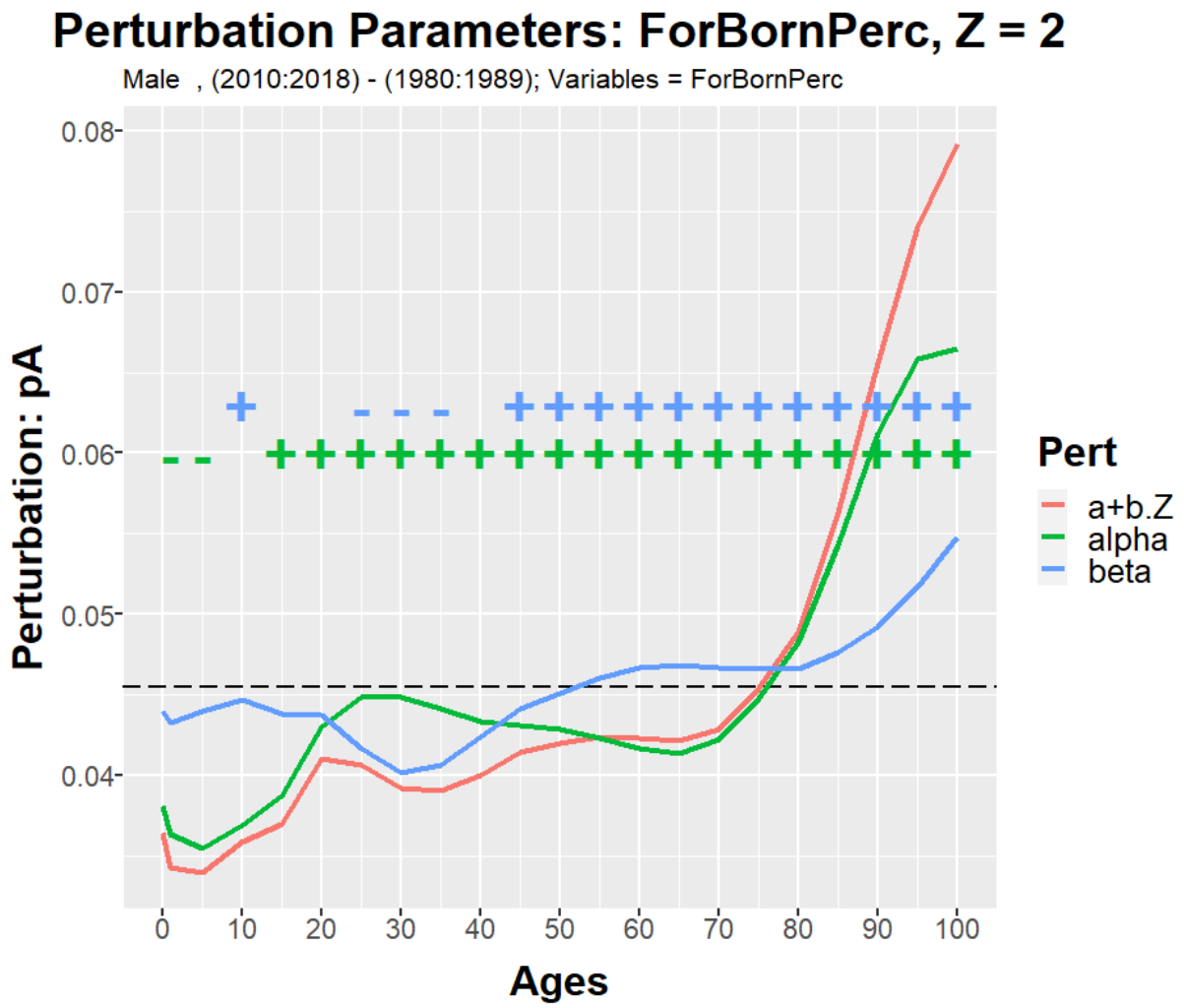


Figure 1: A plus or minus indicates that a parameter with that age and colour is significant at the 5% level and indicates its sign. Neutral perturbation (dashed line).

The green line, alpha, is the estimated perturbation, with unit sum, common to all States, which corresponds to the intercept in a standard regression. When it is below the neutral line, life table deaths have been lost relatively. When above, deaths have been gained. It shows that, overall, there was a transfer of deaths from below age 75, to older ages, between the two periods. However, there was almost no change for young adult males. The interpretation is relative so that, although there are few deaths at ages 5–9 and 100–105, their relative losses and gains were largest.

The blue line is the estimated perturbation linked to the Z–score for the explanatory variable: the percentage Foreign Born. When Z is positive, deaths are shifted from young adult ages to ages 80+.

At other ages, deaths are approximately independent of Z. When Z is negative, the blue line is flipped about the neutral perturbation and the stagnation in life–saving for young adults determined by alpha is converted into an increase in deaths at those ages.

The red line shows the estimated combined effect of the intercept and the explanatory variable when Z equals +2 standard deviations: e.g. the perturbation we might expect for New York State.

3.2 Residual Analysis

The residuals, ϵ_i in equation 1, are compositions. Their norms from the simple–regression model, with the percentage Foreign–Born as the explanatory variable, are mapped in Figure 2. In this cartogram, the background colour shows the error statistic on a linear scale. The abbreviated name of each state is shown in a colour that denotes its position within the population–weighted quartiles. If the text is orange the State falls in the lower error quartile; if red, it is in the upper quartile. States that fall in the inter–quartile range are identified in grey. The range of values covered by these partitions is shown by the horizontal bar on the colour scale.

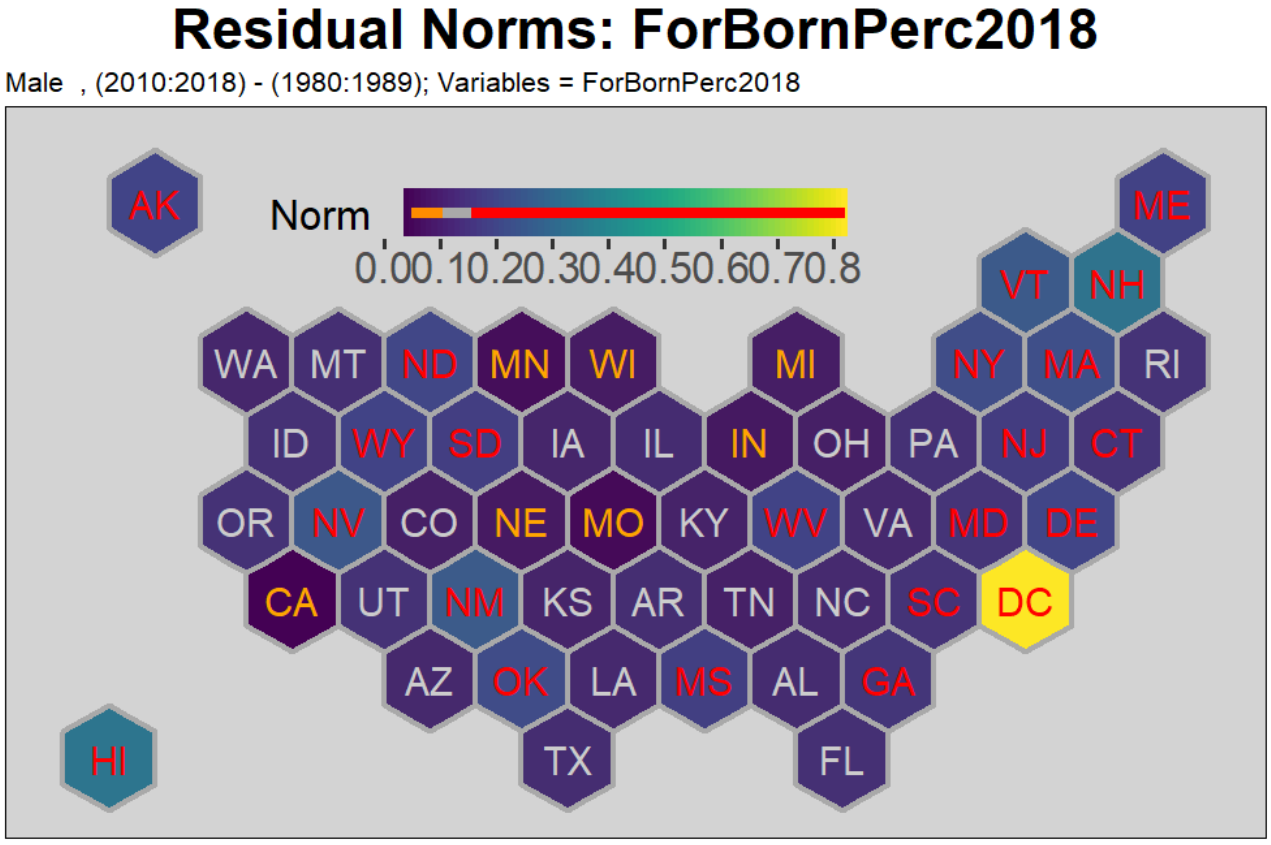


Figure 2: USA State–specific norms of residual perturbations: male (2010:2018 $\ominus$ 1980:1989)

The cartogram shows approximate location and contiguity, but conceals the wide variation in the States' population totals: the biggest contrast is the CA:WY ratio at 66:1.

Washington D.C. has the second–smallest population and therefore little weight in the estimation, but it is a huge outlier. The extraordinary improvements in survival in the District are not captured by this model. The second largest outlier is Hawaii, which ranks 6th highest on Foreign–Born percentage. The model massively overestimates its improvement in survival.

As the residuals for each State are age–specific, they can be explored in further detail.

4 Discussion

As the models estimate the full life table, they can be used to define input–response functions that link covariates to summary outcomes, such as life expectancy or lifespan inequality.

The paper will contain systematic variable selection to extend the analysis to multiple regression. Developing a regression framework that is coherent across all ages of the life table offers access to a very rich array of techniques familiar to economists.

5 References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data.* [Available online.]

Preston, S.H. (1975). 'The Changing Relation between Mortality and Level of Economic Development'. *Population Studies.* 29 (2): 231–248.

United States Mortality DataBase. University of California, Berkeley (USA). Available at usa.mortality.org.