Recovering Indirect Information in Demographic Applications

Jutta Gampe

Abstract In many demographic applications the information of interest can only be estimated indirectly. Modelling events and rates is typical for demographic analyses so that statistical models based on counts are a natural starting point. We will demonstrate that the Penalized Composite Link model is a versatile and valuable tool to solve such indirect estimation problems in demography.

Key words: Demography, Indirect observations, Penalized Composite Link Model

1 Introduction

Demography deals with the analysis of populations and population dynamics, that is, the change of populations over time with respect to their size and composition. More specifically the discipline analyzes the processes that drive population change: Mortality, fertility and migration, as well as related processes such as marriage or divorce.

The models used to describe and to analyze these processes are mostly based on age-specific rates (of fertility, mortality, ...) and the underlying data are the number of events (births, deaths,...) and the number of individuals at-risk of the event. With its emphasis on rates and events, many methods in demography are based on distributional models for counts (and corresponding exposures), so that extensions of Poisson models are a natural starting point.

In many demographic applications the information of interest cannot be estimated directly though, but indirect estimation procedures have to be employed. For many applications the indirect estimation problem can be phrased as a Penalized Composite Link Model (PCLM). This model combines a remarkable versatility of

Jutta Gampe
Max Planck Institute of Demographic Research, 18057 Rostock, Germany,
e-mail: gampe@demogr.mpg.de

model formulation with modest additional assumptions, such as smoothness of rates across age or across time. The resulting procedures are computationally efficient which allows to handle large data sets that are common in demography.

In the next section we give a short summary of the PCLM, followed by several demographic examples in Section 3.

2 The Penalized Composite Link Model

Based on the Composite Link Model suggested in [1] the model was extended by an additional smoothness penalty in [2]. The model considers a series of Poisson counts $y_1, \ldots, y_n$ with $E(y_i) = \mu_i$. In contrast to a simple Poisson regression the means μ_i are supposed to be linearly combined from a series of $\gamma_1, \ldots, \gamma_m$ so that

$$\mu_i = \sum_{j=1}^{m} c_{ij} \gamma_j \,. \tag{1}$$

For the γ_j the common specification of generalized linear models (GLM) is retained: $\gamma_j = e^{\eta_j}$ and $\eta_j = \sum_{k=1}^{p} x_{jk} \beta_k$ is the linear predictor. In matrix notation we can write

$$\mu = C\gamma, \quad \gamma = e^{\eta}, \quad \eta = X\beta \tag{2}$$

The matrix C determines how the expected values of the observable y_i are composed from the elements of γ. To estimate the parameters in the CLM, it can be shown that the standard iteratively weighted least-squares algorithm (IWLS) can be modified in the following way:

Define $\Gamma = diag(\gamma)$, $M = diag(\mu)$ and $U = M^{-1} C \Gamma X$. Then the next IWLS iteration solves

$$\tilde{U}' \tilde{M} \tilde{U} \beta = \tilde{U}'(y - \tilde{\mu}) + \tilde{U}' \tilde{M} \tilde{U} \tilde{\beta} \,, \tag{3}$$

where the tilde indicates current values in the iteration.

If the number of observations n is smaller than the number of parameters p, then the problem is ill-conditioned but an additional penalty can help. This penalty will constrain the elements of the parameter vector β and thereby will allow the estimation of the model. A typical assumption is that the elements of γ are a smooth series. In the simplest case of $X = I$ and hence $\gamma = e^{\beta}$ a difference penalty on the elements of β will do the job. Alternatively we can express $\ln \gamma$ as a linear combination of B-splines so that X holds the spline basis, see [3]. Again a difference penalty on the elements of β will ensure the required smoothness of γ.

If D is a matrix that builds differences of order d (typically $d = 1$ or $d = 2$) and λ is the smoothing parameter that balances the effect of the penalty relative to the model deviance, then maximizing the penalized log-likelihood leads to the following modified system of the IWLS for this penalized composite link model (PCLM):

$$(\tilde{U}' \tilde{M} \tilde{U} + \lambda D' D) \beta = \tilde{U}'(y - \tilde{\mu}) + \tilde{U}' \tilde{M} \tilde{U} \tilde{\beta} \,. \tag{4}$$

For detailed derivations see [2]. The choice of the optimal smoothing parameter λ is usually done by minimizing an information criterion, AIC or BIC, over a grid of λ-values.

3 Examples

3.1 Ungrouping coarse histograms

Many official data (life tables etc.) are reported in grouped form. Age-groups of width 5 or 10 years are common and often the information on the elderly is summarized in a rather wide open age-group such as 80+ or 85+. In aging societies more detailed information on the higher ages is needed. Even if nowadays, in developed countries, the last age-group has been changed into 100+ or 110+ in official statistics, for studying time trends it is required that the coarsely grouped data can be analyzed on a more detailed scale. In [4] it is demonstrated how the PCLM can be employed for this purpose.

Let $a_1, \ldots, a_J$ be the sequence of ages we are interested in (typically $a_1 = 0, a_2 = 1, \ldots, a_J = 110$ or $a_J = 120$) and $\gamma_j, j = 1, \ldots, J$ are the corresponding (but unobserved, because of the grouping) expected counts for these ages. What is observed are the counts $y_1, \ldots, y_n$ in the n age-groups, which are realizations of Poisson variates with means $E(y_i) = \mu_i$. The μ_i are the sum of the γ_j that lie in the age-group represented by the y_i, so that the composition matrix C here is rather simple:

$$C = \begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 1 & \ldots & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & 0 & 0 & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 & 1 & \ldots & 1 \end{pmatrix} \quad (5)$$

The number of rows n is the number of observed age-groups, the number of columns J is the length of the finer age-sequence. Obviously here $n < J$. Assuming that the ungrouped distribution $\gamma_1, \ldots, \gamma_J$ is smooth is not restrictive but adding a smoothness penalty allows to estimate the vector γ. As demonstrated in [4], this approach even works for wide open age-intervals and can also be used to disaggregate exposure numbers if rates are to be estimated on a finer age-grid.

3.2 Additive decomposition of death rates

The trajectory of human mortality over age is rather complex, see Figure 1, and decomposing death rates into additive components has a long tradition. Commonly three components are used, referring to child mortality, a stretch of elevated mortality starting in late puberty – the so called accident-hump – and finally senescent mortality which basically increases exponentially. Particularly the dynamics of the accident hump has recently attracted attention, but this part is rather complex and parametric models rarely fit.

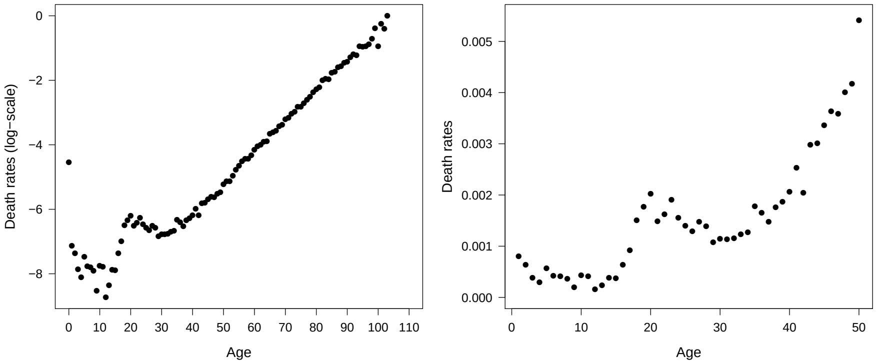

Fig. 1 Age-specific death rates for males in Switzerland in 1980. Data taken from the Human Mortality Database (`www.mortality.org`).

In [5] it was demonstrated that a (partially) nonparametric additive decomposition of death rates can be achieved by a so called sum-of-smooth exponentials (SSE) model. This model is particularly suited to model trajectories with sharp changes.

The observed data here are the death counts y_i and the corresponding exposures e_i across the ages $i = 1, \dots, n$ (where $n = 100$ or even 110). The expected values of the y_i are $\mu_i = \sum_{k=1}^{K} e_i \gamma_{ik}$, where $(\gamma_{1k}, \dots, \gamma_{nk})$ are the death rates for the K components (here $K = 3$). Again the number n is smaller than the number of γ_{ik}.

The different components $\gamma_k = (\gamma_{1k}, \dots, \gamma_{nk})$ here are modeled as $\gamma_k = \exp\{B_k \beta_k\}$, where the matrix B_k holds either a B-spline basis for the smooth components or expresses a parametric specification, e.g. for the exponentially increasing senescent component.

The structure of the μ_i again suggests a PCLM approach. In this problem a smoothness penalty is not enough to make the decomposition identifiable, however, adding penalties for shape constraints, such as monotonicity or the accident-hump component being log-concave, solves the problem. Details can be found in [5]. Figure 2 shows the resulting estimates from a three-component SSE model for the Swiss mortality data presented in Figure 2. Here the child and senescent mortality were modeled by parametric functions and the accident-hump component was specified to be smooth and being concave on the log-scale.

The model has also be extended to two dimensions so that the change of the different components over time can be studied.

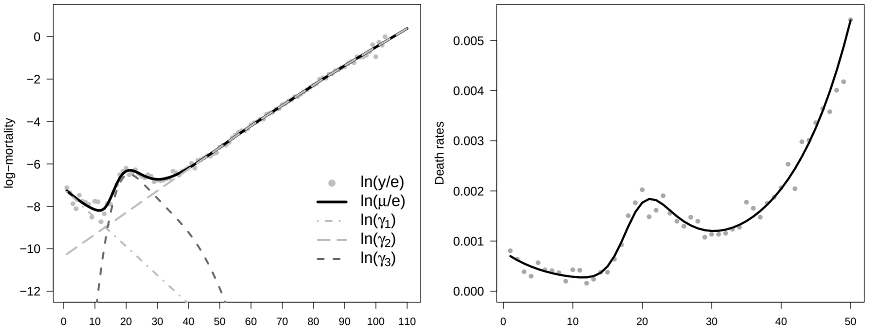

Fig. 2 Age-specific death rates for males in Switzerland in 1980. Estimates of the three-component SSE model.

3.3 Age-at-death distributions in paleodemography

While in modern populations ages-at-death are available from death certificates or, in earlier centuries, from church registers, we do not have such information if we go further back in time. In this case age-estimates from excavated skeletons in historical cemeteries, which can be done by experienced anthropologists, is the only option. Age-estimates are not necessarily unbiased, but they can be calibrated by using so called known-age reference collections.

The overall procedure runs in two steps. If we denote the true age-at-death by a and the estimated skeletal age by s, we can use m skeletons from a reference collection, for whom we have $(a_i, s_i), i = 1, \ldots, m$, to estimate the conditional distribution $f(s|a)$. This is usually done by nonparametric regression to allow for nonlinear and heteroscedastic relations between a and s.

In the so called target population (the excavated burial site) we only observe the sekeletal age on n buried individuals. The age-at-death distribution $f(a)$, which is to be estimated, is related to the distribution $g(s)$ of skeletal ages by

$$g(s) = \int f(s|a) f(a) da \,. \tag{6}$$

Information on $f(s|a)$ is borrowed from the reference collection.

If we model both age-scales on a fine grid, i.e., $s = (s_1, \ldots, s_L)$ and $a = (a_1, \ldots, a_K)$, then the observations are the counts of skeletons in each of the L skeletal-age classes: $y_1, \ldots, y_L$, $\sum_l y_l = n$. These are realizations of Poisson variates with means μ_l, where

$$\mu_l = n \sum_{k=1}^{K} c_{lk} \gamma_k \,. \tag{7}$$

The $c_{lk} = P(s_l|a_k)$ stem from the reference collection and $\gamma_k = P(a_k)$. Again this deconvolution problem has the structure of a PCLM and the unknown age-at-death distribution $\gamma = (\gamma_1, \ldots, \gamma_K)$ can be estimated by using a simple smoothness penalty.

References

1. Thompson R., Baker R.J.: Composite Link Functions in Generalized Linear Models. *Applied Statistics* **30** (2), 125–131 (1981)
2. Eilers, P.H.C: Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling* **7** (3), 239–254 (2007)
3. Eilers, P.H.C, Marx, B.: Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11** (2), 89–121 (1996)
4. Rizzi, S., Gampe, J., Eilers, P.H.C: Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. *American Journal of Epidemiology* **182** (2), 138147 (2015)
5. Camarda, G.C., Eilers, P.H.C., Gampe, J.: Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling* **16** (4), 279–296 (2016)