Annals of Actuarial Science (2020), **14**, pp. 537–566
doi:10.1017/S1748499520000020

PAPER

Constraints, the identifiability problem and the forecasting of mortality

Iain D. Currie*

Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK
* Correspondence to: Iain D Currie, Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK. Tel: +44 (0)131 451 3203. Fax: +44 (0)131 451 3249. E-mail: I.D.Currie@hw.ac.uk

(Received 28 January 2020; accepted 29 January 2020; first published online 16 March 2020)

Abstract

Models of mortality often require constraints in order that parameters may be estimated uniquely. It is not difficult to find references in the literature to the "identifiability problem", and papers often give arguments to justify the choice of particular constraint systems designed to deal with this problem. Many of these models are generalised linear models, and it is known that the fitted values (of mortality) in such models are identifiable, i.e., invariant with respect to the choice of constraint systems. We show that for a wide class of forecasting models, namely ARIMA(p, δ, q) models with a fitted mean and $\delta = 1$ or 2, identifiability extends to the forecast values of mortality; this extended identifiability continues to hold when some model terms are smoothed. The results are illustrated with data on UK males from the Office for National Statistics for the age-period model, the age-period-cohort model, the age-period-cohort-improvements model of the Continuous Mortality Investigation and the Lee–Carter model.

Keywords Constraints; Forecasting; Identifiability; Invariance; Mortality

1. Introduction

The modelling and forecasting of mortality play a fundamental part in the working life of an actuary. It is widely known that human mortality depends on an individual's current age, the current year and their year of birth (among many other possible risk factors). These determinants of mortality are generally known as the *age effect*, the *period effect* and the *cohort effect*, respectively. There is a tried and tested method of approaching the problem. We define a model for the force of mortality which depends on the age, the period and usually (but not always) the cohort effect. The parameters are estimated from some suitable data. The period and cohort parameters are then forecast, and from these estimated and forecast values, the forecast values of mortality are obtained. Examples of such models are the Lee–Carter model (Lee & Carter, 1992), the age-period-cohort model (Clayton & Schifflers, 1987b) and various forms of the Cairns–Blake–Dowd (CBD) model (Cairns *et al.*, 2009). More recently, the Continuous Mortality Investigation or CMI introduced the age-period-cohort-improvements model (Continuous Mortality Investigation, 2016a, 2016b, 2016c). For a particular model, the method can be summarised in the following three steps: (1) estimate the age, period and cohort parameters, (2) forecast the period and cohort parameters and (3) obtain the forecast values of mortality.

There is a difficulty. The models mentioned above do not allow the unique estimation of the parameter sets without the introduction of some further assumptions. The standard approach is to

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

place some constraints on the estimates, such as the period effects are constrained to sum to zero. There is no reason why two actuaries working independently should use the same constraints; they will obtain different estimates of the age, period and cohort effects. Indeed, Clayton & Schifflers (1987b) in a carefully argued paper warn against the forecasting of cohort effects in particular. This is often referred to as the "identifiability problem"; see Cairns *et al.* (2009), Continuous Mortality Investigation (2016a, 7.3) and Richards *et al.* (to appear).

In our paper, we take the view that the estimates of the age, period and cohort effects are of interest to the actuary only as intermediate quantities; they are used to obtain the quantities of interest, namely, the forecast values of mortality. Apart from the Lee–Carter model, the models mentioned above are all examples of generalised linear models or GLMs (Nelder & Wedderburn, 1972). It is a basic result that, while the parameters in the GLMs mentioned above are not identifiable, the fitted values (of mortality) are identifiable (Continuous Mortality Investigation, 2016a, 7.3), i.e., invariant with respect to the choice of constraints on the parameters. The main result in our paper is that the forecast values of mortality are also identifiable when the period and cohort effects are forecast with an autoregressive integrated moving average or ARIMA model with fitted mean: two actuaries working independently may not obtain the same estimates of the age, period and cohort effects, but they will obtain the same forecasts of mortality.

The problem of identifiability in the forecasting of mortality has received considerable attention within both the actuarial and the statistical literature. Hunt & Blake (2020a, 2020b, 2020c) in a series of papers discuss the problem for a general class of age-period models (of which the Lee–Carter model is a simple example) and for this class with an added cohort effect. They emphasise the arbitrary nature of a particular set of identifiability constraints and the importance of using a forecasting method which does not depend on this arbitrary choice. They call such a method "well-identified", discuss how to choose such a forecasting method and provide some examples.

The plan of the paper is as follows: in section 2, we describe the data we use to illustrate our results, set out our notation and define the class of model we discuss. In section 3, we state the fundamental result on invariance of fitted values in a GLM in terms of the *null space* of the model matrix; this section also contains the necessary theory for estimation in a GLM in which some parameters are subject to constraints and/or smoothing. Section 4 contains our four examples: the age-period, age-period-cohort, age-period-cohort-improvements and Lee–Carter models. We make some concluding remarks in section 5. There are three appendices. Appendix A gives the matrix theory underlying the use of null spaces as applied to GLMs. Appendix B shows how invariance can be exploited to give a simple way of fitting GLMs with specified constraints. In Appendix C, we give a proof of the time series result used to show the invariance of forecasting with respect to the choice of constraints.

2. Data, Notation and the Basic Model

We illustrate our ideas with data from the Office for National Statistics on UK males. The data comprise the deaths and central exposures for ages 50–104 and years 1971–2015. These data lie naturally in a matrix with rows indexed by age and columns indexed by year. We adopt the convention that matrices are denoted by upper case bold font as in $\boldsymbol{X}$ and column vectors by lower case bold font as in $\boldsymbol{x}$. With this in place, we denote the age indices by $\boldsymbol{x}_a = (1, \ldots, n_a)'$ and the year indices by $\boldsymbol{x}_y = (1, \ldots, n_y)'$, where the $'$ indicates the transpose of a vector (or matrix); this simplifies the notation without jeopardising the presentation. The data thus comprise two matrices: $\boldsymbol{D}_{obs} = (d_{x,y})$, the matrix of the number of deaths at age x in year y, and $\boldsymbol{E}_{obs} = (e_{x,y})$, the total time lived or central exposure at age x in year y. (We will use $\boldsymbol{D}$ later to denote a difference matrix.) Thus, $\boldsymbol{D}_{obs}$ and $\boldsymbol{E}_{obs}$ are both $n_a \times n_y$. Furthermore, we label the cohorts with $x_c = (1, \ldots, n_c)'$, where $n_c = n_a + n_y - 1$ is the number of distinct cohorts. We adopt the convention that the oldest cohort, i.e., for age n_a in year 1, is indexed 1; hence, the cohort index for cell (i, j) is

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

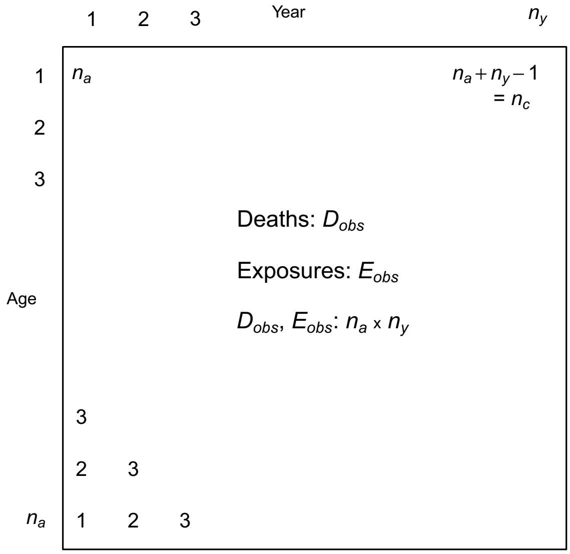

Figure 1. Age of death = rows, year of death = columns, year of birth = diagonals downwards from left to right.

$c(i,j) = n_a - i + j$. In our example, we have $n_a = 55$, $n_y = 45$ and $n_c = 99$. Figure 1 summarises our notation.

In addition, the following notation is useful: $\mathbf{1}_n$ denotes a column vector of 1s of length n and $\boldsymbol{I}_n$ denotes the identity matrix of size n. Furthermore, we let $\mathbf{0}_n$ stand for either a column vector of 0s of length n or an $n \times n$ zero matrix; the context should make clear which applies. We may omit the suffix n if no confusion results. We will make frequent use of the Kronecker product, which we denote $\otimes$; see Macdonald *et al.* (2018, chapter 12) for examples of the Kronecker product as used in expressing models of mortality.

We assume that $D_{x,y}$, the random variable corresponding to the observed deaths $d_{x,y}$, follows the Poisson distribution with mean $e_{x,y}\mu_{x,y}$, where $\mu_{x,y}$ is the force of mortality at age x in year y, i.e., $D_{x,y} \sim \mathcal{P}(e_{x,y}\mu_{x,y})$. We note that there is a slight abuse of notation here; strictly, we should write $\mu_{x+1/2,y+1/2}$ since $e_{x,y}$ is the central exposure; we will use the simpler notation throughout for clarity. Let vec($\boldsymbol{M}$) be the function that stacks the columns of a matrix $\boldsymbol{M}$ on top of each other in column order. Let $\boldsymbol{d} = \text{vec}(\boldsymbol{D}_{obs})$ and $\boldsymbol{e} = \text{vec}(\boldsymbol{E}_{obs})$ be the stacked vectors of deaths and exposures; let $\boldsymbol{\mu}$ be the corresponding vector of forces of mortality. We consider models where $\log \boldsymbol{\mu}$ can be written in the following form:

$$\log \boldsymbol{\mu} = \boldsymbol{X\theta} \tag{1}$$

Together with the Poisson assumption, this defines a generalised linear model or GLM with *model matrix* $\boldsymbol{X}$, log link and Poisson error; the exposure enters into the model definition as an offset, $\log \boldsymbol{e}$. This is a very wide class of model and includes the Gompertz model (in one dimension), the age-period or AP model, the age-period-cohort or APC model, various forms of the CBD models, and the CMI's age-period-cohort-improvements or APCI model. The Lee–Carter model, although not immediately in this class, can also be included, as we will show in section 4.4.

3. Models and Null Spaces

We assume that we have a GLM with model matrix $\boldsymbol{X}$, regression coefficients $\boldsymbol{\theta}$, a log link and a Poisson error distribution. We suppose that $\boldsymbol{X}$ is $n \times p$, $n > p$, with rank $p - q$ where $q \geq 1$; see Appendix A for a brief discussion of rank. We denote the rank of $\boldsymbol{X}$ by $r(\boldsymbol{X})$. The model matrix is

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

not of full rank; this implies that the estimates of $\boldsymbol{\theta}$ are not unique. However, if we specify a set of q linearly independent constraints $\boldsymbol{H\theta} = \mathbf{0}_q$, where

$$\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{H} \end{pmatrix}$$

has full rank p, then we do have a unique estimate of $\boldsymbol{\theta}$. In general, the idea is to choose $\boldsymbol{H}$ so that the components of the resulting estimate have a natural interpretation and can be forecast. All the models considered in this paper have model matrices which are not of full rank.

Suppose $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are two estimates of $\boldsymbol{\theta}$ subject to the constraints $\boldsymbol{H}_1\boldsymbol{\theta} = \boldsymbol{H}_2\boldsymbol{\theta} = \mathbf{0}_q$, respectively. We wish to understand the relationship between $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. This relationship is characterised by the *null space* of $\boldsymbol{X}$, $\mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{v}\colon \boldsymbol{Xv} = \mathbf{0}\}$. To be precise

$$\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \in \mathcal{N}(\boldsymbol{X}) \tag{2}$$

see Appendix A for a proof of this fundamental result. Thus, $\mathcal{N}(\boldsymbol{X})$ will tell us how the estimates of $\boldsymbol{\theta}$ under different constraint systems are related; ideally, a forecast will not depend on this choice. We note in particular that $\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \in \mathcal{N}(\boldsymbol{X})$ implies that

$$\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2 \tag{3}$$

in other words, the fitted values of log mortality are equal, an example of the general result that the fitted values in a GLM are equal under different constraint systems.

It is often advantageous to smooth certain terms in a model of mortality. We will use the P-spline system of smoothing; see Eilers & Marx (1996), Macdonald *et al.* (2018, chapter 11). The idea is to replace a parameter, $\boldsymbol{\alpha}$ say, by a smooth function, $\boldsymbol{Ba}$, where $\boldsymbol{B}$ is a regression matrix evaluated over a basis of B-splines; we use cubic B-splines throughout. In our examples, $\boldsymbol{\alpha}$ will be an age parameter, and for technical reasons, we will always place a knot at the first age. The regression coefficients $\boldsymbol{a}$ are then subject to a penalty which penalises local differences in $\boldsymbol{a}$.

Let $\boldsymbol{\alpha} = \boldsymbol{Ba}$, where $\boldsymbol{B}$ is $n_a \times c_a$ and $\boldsymbol{a} = (a_1, \ldots, a_{c_a})'$. The second-order penalty on $\boldsymbol{a}$ is

$$(a_1 - 2a_2 + a_3)^2 + \ldots + (a_{c_a-2} - 2a_{c_a-1} + a_{c_a})^2 \tag{4}$$

We write this compactly by defining $\boldsymbol{D}$, the difference matrix of order two, as

$$\boldsymbol{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \ (c_a - 2) \times c_a \tag{5}$$

With this notation, the quadratic penalty (4) can be written as

$$\boldsymbol{a}'\boldsymbol{D}'\boldsymbol{Da} \tag{6}$$

We make the important remark that $r(\boldsymbol{D})$ is $c_a - 2$ and so the dimension of the null space of $\boldsymbol{D}$, $\mathcal{N}(\boldsymbol{D})$, is two. A basis for $\mathcal{N}(\boldsymbol{D})$ is $\{\boldsymbol{n}_1, \boldsymbol{n}_2\}$, where $\boldsymbol{n}_1 = \mathbf{1}_{c_a}$ and $\boldsymbol{n}_2 = (1, 2, \ldots, c_a)'$. We will be interested in the relationship between $\mathcal{N}(\boldsymbol{D})$ and $\mathcal{N}(\boldsymbol{X})$ since this can affect the number of constraints required to enable parameters to be estimated uniquely. Proposition 4 in Appendix A describes this relationship, which is illustrated in our discussion of our examples in section 4.

The strength of the penalty is determined by the *smoothing parameter* λ, and, from (6), we define the *penalty matrix*

$$\boldsymbol{P} = \lambda\boldsymbol{D}'\boldsymbol{D} \tag{7}$$

Clearly, if λ is small the coefficients will be less smooth, while if λ is large the coefficients will be more smooth.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

We have described a penalty of order two but penalties of other orders can be used. For example, in his landmark paper on graduation, Whittaker (1923) used a third-order penalty with terms like $(a_1 - 3a_2 + 3a_3 - a_4)^2$. If $\boldsymbol{D}$ is the resulting difference matrix, then $\mathcal{N}(\boldsymbol{D}) = \{\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{n}_3\}$, where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are as in the previous paragraph and $\boldsymbol{n}_3 = (1^2, 2^2, \ldots, c_a^2)'$. The first-order penalty has terms like $(a_1 - a_2)^2$, and the null space of the corresponding difference matrix consists of $\boldsymbol{n}_1$ only. In general, we denote the order of the penalty by d. A basis for the null space of a difference matrix of order d is given by $\{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_d\}$, where

$$\boldsymbol{n}_j = (1^{j-1}, 2^{j-1}, \ldots, c_a^{j-1})',\ j = 1, \ldots, d \tag{8}$$

We describe a general approach to fitting GLMs that are subject to both constraints and smoothing. Nelder and Wedderburn (1972) introduced GLMs and showed that estimation in a GLM is given by

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{9}$$

where the tilde, as in $\tilde{\boldsymbol{\theta}}$, indicates a current estimate, while $\hat{\boldsymbol{\theta}}$ indicates an improved approximation in the iterative scheme. The matrix $\tilde{\boldsymbol{W}}$ is the *diagonal matrix of weights* and the vector $\tilde{\boldsymbol{z}}$ is the so-called *working variable*. For Poisson errors and a log link, $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{z}}$ are

$$\tilde{\boldsymbol{W}} = \text{diag}\{\tilde{\boldsymbol{d}}\},\ \tilde{\boldsymbol{z}} = \boldsymbol{X}\tilde{\boldsymbol{\theta}} + \left(\frac{\boldsymbol{d}}{\tilde{\boldsymbol{d}}} - \boldsymbol{1}\right) \tag{10}$$

where $\tilde{\boldsymbol{d}}$ indicates the current estimate of the fitted deaths and $\boldsymbol{d}/\tilde{\boldsymbol{d}}$ indicates element-by-element division.

A possible and easily overlooked complication may arise when $\boldsymbol{\theta}$ (or some portion of $\boldsymbol{\theta}$) is smoothed. With P-splines, equation (9) becomes

$$(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \tag{11}$$

where $\boldsymbol{P}$ is the penalty matrix (Currie *et al.*, 2004). The number of constraints required for (11) to have a unique solution is now determined by $r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})$ and not by $r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}) = r(\boldsymbol{X})$. Fortunately, for most models of mortality, we do have $r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P}) = r(\boldsymbol{X})$, but there are cases when $r(\boldsymbol{X})$ is strictly less than $r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})$, so it is best to be aware of this possible pitfall. In such cases, the number of constraints required to yield a unique estimate of $\boldsymbol{\theta}$ will be reduced; some examples are given in section 4, notably for the APCI model. This discussion suggests the following definition.

Definition: *The* effective rank *of a model with model matrix* $\boldsymbol{X}$ *and penalty* $\boldsymbol{P}$ *is*

$$r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P}) \tag{12}$$

We show in Proposition A.4 in Appendix A that

$$\mathcal{N}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P}) = \mathcal{N}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{P}) = \mathcal{N}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{P}) \tag{13}$$

and so

$$r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P}) \geq r(\boldsymbol{X}) \tag{14}$$

Thus, smoothing can never increase the number of constraints required to obtain unique estimates of $\boldsymbol{\theta}$. The effective rank of a model has connections with the widely used effective dimension of a model but is a distinct idea; see Macdonald et al. (2018, chapter 11) for a discussion of effective dimension.

In our case, $\boldsymbol{X}$ is not of full rank; equation (9) is singular and cannot be solved. The R package (R Core Team, 2018) has its own way of dealing with this problem, but we want solutions that

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

satisfy particular constraints and where some components of $\boldsymbol{\theta}$ may be subject to smoothing. Currie (2013) generalised equation (9) to deal with this case and showed that

$$\begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}+\boldsymbol{P} & : & \boldsymbol{H}' \\ \boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \\ \boldsymbol{k} \end{pmatrix} \tag{15}$$

is a Newton–Raphson scheme whose solution is the maximum likelihood estimate of $\boldsymbol{\theta}$ subject to (a) the linear constraint $\boldsymbol{H\theta}=\boldsymbol{k}$ and (b) smoothing via the penalty matrix $\boldsymbol{P}$. Here, $\hat{\boldsymbol{\omega}}$ is an auxiliary variable; see the appendix in Currie (2013) for a description of its role. We make two remarks: first, in our case, $\boldsymbol{k}$ will usually be a zero vector and second, if there is no smoothing then there is a neat way of recovering estimates of $\boldsymbol{\theta}$ subject to $\boldsymbol{H\theta}=\boldsymbol{k}$ from R's output; see Currie (2016) and Appendix B. All the computations in this paper are done in R and use algorithm (15).

4. Examples

In this section, we present four examples. There is a good argument for smoothing some of the model terms in each of our examples, particularly parameters varying by age. Accordingly, for each example, we begin with the case when no model terms are smoothed; we follow this with a discussion of a smooth model. In each example, we obtain the null space of the model matrix and discuss its influence on forecasting. We start with the age-period model, a simple illustration of our approach to the "problem of identifiability". Our second example, the age-period-cohort model, is non-trivial; we give a fuller discussion of this model. Next, we discuss the CMI's age-period-cohort-improvements model; this model is more complex, and we concentrate on the case when some model terms are smoothed. Lastly, we discuss the Lee–Carter model; this is not immediately in the model class defined by equation (1), so it has its own particular features.

We will forecast model terms with autoregressive integrated moving average models with fitted mean which, in standard notation, we denote by ARIMA(p, δ, q); here p is the order of the autoregressive term, δ is the order of differencing and q is the order of the moving average term; here we use δ instead of the usual d since we use d to denote the order of the penalty. We make three important comments on our forecasting methods. First, with one exception, all our forecasting models are fitted with a mean. Second, we also consider the simple random walk, i.e., a random walk with zero drift, a model often used to forecast cohort effects. Third, plots of estimated period and cohort parameters usually indicate that $d=1$ and $d=2$ are appropriate, and we concentrate on these cases. Shumway & Stoffer (2017) is a standard reference on time series.

4.1 Age-period model

The age-period or AP model is very simple, probably too simple to be useful, but it is a clear demonstration of the method and a straightforward illustration of our results.

4.1.1 AP model

Under the AP model, we have

$$\log \mu_{x,y} = \alpha_x + \kappa_y,\ x=1,\ldots,n_a,\ y=1,\ldots,n_y \tag{16}$$

First, we write the model in the standard form (1) and compute its rank. Let $\boldsymbol{\theta}=(\boldsymbol{\alpha}',\boldsymbol{\kappa}')'$ be the vector of regression coefficients. The model matrix is

$$\boldsymbol{X}=[\boldsymbol{X}_a:\boldsymbol{X}_y]=[\boldsymbol{1}_{n_y}\otimes\boldsymbol{I}_{n_a}:\boldsymbol{I}_{n_y}\otimes\boldsymbol{1}_{n_a}],\ n_a\,n_y\times(n_a+n_y) \tag{17}$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

and $\log \boldsymbol{\mu} = \boldsymbol{X\theta}$. Second, we find a basis for the null space of $\boldsymbol{X}$. Since $\boldsymbol{X}$ is $n_a\, n_y \times (n_a + n_y)$ and $r(\boldsymbol{X}) = n_a + n_y - 1$, the dimension of $\mathcal{N}(\boldsymbol{X})$ is one. In this simple case, we can just write down a basis vector for $\mathcal{N}(\boldsymbol{X})$. We consider

$$\boldsymbol{n} = \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \end{pmatrix} \tag{18}$$

and it is easy to check that $\boldsymbol{Xn} = \mathbf{0}$ and so $\boldsymbol{n}$ spans $\mathcal{N}(\boldsymbol{X})$.

We adopt the following approach to computation. We use two constraint systems: first, a *standard constraint system*, i.e., one that is found in the literature: for the AP model we take $\sum \kappa_y = 0$; second, a *random constraint system:* we set $\sum_1^{n_a+n_y} u_i\theta_i = 0$ where u_i, $i = 1, \ldots, n_a + n_y$, are realisations of independent uniform variables on [0, 1], i.e., $\mathcal{U}(0, 1)$. Let $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ be the estimates of $\boldsymbol{\theta}$ under the two systems. Then, by our fundamental result (2), $\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r \in \mathcal{N}(\boldsymbol{X})$ and so

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \end{pmatrix} \tag{19}$$

for some scalar A. Equating coefficients, we find that

$$\hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r = A\mathbf{1}_{n_a}, \quad \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r = -A\mathbf{1}_{n_y} \tag{20}$$

where $\hat{\boldsymbol{\alpha}}_s$, $\hat{\boldsymbol{\kappa}}_s$, $\hat{\boldsymbol{\alpha}}_r$ and $\hat{\boldsymbol{\kappa}}_r$ are the components of $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$, respectively. The value of A depends on the particular values of u_i; in our computation, we found $A = 20.7$ although any value of A is possible since it corresponds to adding A to α_x and subtracting A from κ_y in (16). We note that (20) confirms what has been widely observed, namely that the $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are only estimable up to an additive constant. A careful discussion of what can and cannot be estimated in the AP model can be found in Clayton & Schifflers (1987a).

Forecasting in the AP model is done by forecasting the $\boldsymbol{\kappa}$ values, keeping the $\boldsymbol{\alpha}$ values fixed at their estimated values and then using equation (16) to forecast the values of $\log \mu$ at each age. Informally, we argue that since $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ are parallel with $\hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r = -A\mathbf{1}_{n_y}$, we require that their forecast values, $\hat{\boldsymbol{\kappa}}_{s,f}$ and $\hat{\boldsymbol{\kappa}}_{r,s}$ say, are also parallel with $\hat{\boldsymbol{\kappa}}_{s,f} - \hat{\boldsymbol{\kappa}}_{r,f} = -A\mathbf{1}_{n_f}$, where n_f is the length of the forecast. Now, the forecast values under both constraint systems will be equal since the change in the $\boldsymbol{\kappa}$ forecast values is exactly compensated for by the change in the $\boldsymbol{\alpha}$ values. This will be achieved with an ARIMA(p, δ, q) model, $\delta = 1$ or 2, since the fitted means of the forecasts for $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ will differ by $-A$.

This informal argument extends to the case when $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ differ by a linear function, as is the case in the age-period-cohort model for example. Here, we want the forecasts to obey the same linear relationship. A formal argument is given in section 4.3.2 when we discuss the age-period-cohort-improvements model.

We make the important remark that not all forecasting models used in the forecasting of mortality lead to forecasts which are invariant with respect to the choice of constraints. For example, if we forecast $\boldsymbol{\kappa}$ with an AR(1) model *without a mean,* then the forecasts will not be invariant. In this paper, we take the position that if two time series differ by a function, then the forecasts should differ by the same function. For example, in the AP model, $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ differ by a constant function, i.e., are parallel; our forecasts should also be parallel. In the APC model, estimates of the period terms and of the cohort terms differ by linear functions. Again, we want our forecasts of the period and cohort terms to obey the same functional relationship. In this way, we will find that not only are fitted values of mortality invariant with respect to the choice of constraints so also are their forecast values.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

4.1.2 Smooth AP model

We turn now to the effect of constraints when some model terms are smoothed. In the AP model, the forecast values of $\log \boldsymbol{\mu}$ are more regular by age if the age parameters $\boldsymbol{\alpha}$ are smoothed (see Delwarde *et al.*, 2007; Currie, 2013). Let $\boldsymbol{\alpha} = \boldsymbol{B}_a \boldsymbol{a}$, where $\boldsymbol{B}_a$ is a B-spline regression matrix along age; here, $\boldsymbol{B}_a$ is $n_a \times c_a$, where c_a is the number of B-splines in the basis. The model matrix (17) becomes

$$\boldsymbol{X} = [\boldsymbol{X}_a : \boldsymbol{X}_y] = [\boldsymbol{1}_{n_y} \otimes \boldsymbol{B}_a : \boldsymbol{I}_{n_y} \otimes \boldsymbol{1}_{n_a}],\ n_a\, n_y \times (c_a + n_y) \tag{21}$$

with rank $c_a + n_y - 1$; the regression coefficients $\boldsymbol{\theta} = (\boldsymbol{a}', \boldsymbol{\kappa}')'$. We distinguish between the regression coefficients denoted $\boldsymbol{\theta}$ and the parameters of interest which we denote $\boldsymbol{\theta}^*$; here $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}', \boldsymbol{\kappa}')'$. Let $\boldsymbol{X}^*$ be the model matrix (17) in the unsmoothed case. We note that

$$\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^* \boldsymbol{\theta}^* \tag{22}$$

since $(\boldsymbol{1}_{n_y} \otimes \boldsymbol{B}_a)\boldsymbol{a} = (\boldsymbol{1}_{n_y} \otimes \boldsymbol{I}_{n_a})\boldsymbol{\alpha}$. We will see that (22) is a very convenient identity. If there is no smoothing then $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and $\boldsymbol{X} = \boldsymbol{X}^*$.

The penalty matrix is

$$\boldsymbol{P} = \text{blockdiag}\{\lambda_a \boldsymbol{D}'\boldsymbol{D}, \boldsymbol{0}\}$$

where $\boldsymbol{D}$ is the difference matrix of order d acting on the smoothed age coefficients $\boldsymbol{a}$, $\boldsymbol{0}$ is the $n_y \times n_y$ matrix of 0s acting on the unsmoothed year coefficients $\boldsymbol{\kappa}$ and λ_a is the smoothing parameter. We compute $\mathcal{N}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})$ for the model in equation (21). Let

$$\boldsymbol{n} = \begin{pmatrix} \boldsymbol{1}_{c_a} \\ -\boldsymbol{1}_{n_y} \end{pmatrix}$$

Using the fact that $\boldsymbol{B}_a \boldsymbol{1}_{c_a} = \boldsymbol{1}_{n_a}$, we see that $\boldsymbol{X}\boldsymbol{n} = \boldsymbol{0}$, i.e., $\boldsymbol{n} \in \mathcal{N}(\boldsymbol{X})$. Furthermore, we have $\boldsymbol{D}\boldsymbol{1}_{c_a} = \boldsymbol{0}$ for any order $d > 0$ of the penalty; hence, $\boldsymbol{P}\boldsymbol{n} = \boldsymbol{0}$, i.e., $\boldsymbol{n} \in \mathcal{N}(\boldsymbol{P})$. Hence, by (13), $\mathcal{N}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P}) = \mathcal{N}(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{P}) = \boldsymbol{n}$. Thus, if $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ are any two estimates of $\boldsymbol{\theta}$, we have

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A \begin{pmatrix} \boldsymbol{1}_{c_a} \\ -\boldsymbol{1}_{n_y} \end{pmatrix}$$

for some scalar A. Pre-multiplying both sides by blockdiag$\{\boldsymbol{B}_a : \boldsymbol{I}_{n_y}\}$, we find

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}_s \\ \hat{\boldsymbol{\kappa}}_s \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\alpha}}_r \\ \hat{\boldsymbol{\kappa}}_r \end{pmatrix} = A \begin{pmatrix} \boldsymbol{1}_{n_a} \\ -\boldsymbol{1}_{n_y} \end{pmatrix}$$

exactly as in the unsmoothed case. We conclude that smoothing has no effect on our invariance results for the AP model.

We have presented a mathematical discussion of the AP model. An important practical point is that from a computational point of view, we do not need the detailed understanding provided by this discussion. We can find the rank and effective rank and check the invariance of the fitted and forecast values easily in R (or other computational tool). We will make further comments on the computational aspects of this work as we go.

4.2 Age-period-cohort model

The age-period-cohort or APC model with almost twice as many parameters as the simple AP model gives a much improved fit.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

4.2.1 APC model

Under the APC model, we have

$$\log \mu_{x,y} = \alpha_x + \kappa_y + \gamma_{c(x,y)}, \; x = 1, \ldots, n_a, \; y = 1, \ldots, n_y \tag{23}$$

where $c(x, y)$ is the cohort index for age x in year y; with our notation, we have $c(x, y) = n_a - x + y$. First, we write model (23) in the standard form (1) and compute its rank. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}', \boldsymbol{\gamma}')'$. The model matrix is

$$\boldsymbol{X} = [\boldsymbol{X}_a : \boldsymbol{X}_y : \boldsymbol{X}_c], \; n_a\, n_y \times (2n_a + 2n_y - 1) \tag{24}$$

where $\boldsymbol{X}_a$ and $\boldsymbol{X}_y$ are defined in (17); the row of $\boldsymbol{X}_c$ corresponding to age x and year y contains a one in column $n_a - x + y$ and zeros elsewhere. The rank of $\boldsymbol{X}$ is $2n_a + 2n_y - 4$ and so the dimension of $\mathcal{N}(\boldsymbol{X})$ is three. The relationship between the estimates of $\boldsymbol{\theta}$ under different constraint systems is determined by $\mathcal{N}(\boldsymbol{X})$.

We now find a basis $\{\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{n}_3\}$ for $\mathcal{N}(\boldsymbol{X})$. We note that this basis is not unique, but we can find a basis that clarifies the relationship between the different estimates of $\boldsymbol{\theta}$. First, we argue that each of $\boldsymbol{X}_a$, $\boldsymbol{X}_y$ and $\boldsymbol{X}_c$ contains a one in each row and zeros elsewhere. Hence, we can take $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ equal to

$$\boldsymbol{n}_1 = \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix}, \quad \boldsymbol{n}_2 = \begin{pmatrix} \mathbf{1}_{n_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix} \tag{25}$$

It remains to find a suitable $\boldsymbol{n}_3$. Some computation is helpful. We fit the APC model under a set of standard constraints, which we take as

$$\sum_1^{n_y} \kappa_y = \sum_1^{n_c} \gamma_c = \sum_1^{n_c} c\gamma_c = 0 \tag{26}$$

where c is the cohort index (Cairns *et al.*, 2009). An alternative is to weight by the number of times cohort c appears in the data, say, w_c. Thus, we could use

$$\sum_1^{n_y} \kappa_y = \sum_1^{n_c} w_c\gamma_c = \sum_1^{n_c} w_c c\gamma_c = 0$$

as in Richards *et al.* (to appear). Our main point is that, although we will get (slightly) different parameter estimates with these constraint systems, our forecasts of mortality will be identical. We also fit the model with a set of random constraints

$$\sum u_{i,j}\theta_j = 0, \; i = 1, 2, 3, \; j = 1, \ldots, n_a + n_y + n_c$$

where the $u_{i,j}$ are independent realisations from the $\mathcal{U}(0, 1)$ distribution. Let $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ be the maximum likelihood estimates of $\boldsymbol{\theta}$ under the two constraint systems. Let $\hat{\boldsymbol{\alpha}}_s$, $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\gamma}}_s$ be the components of $\hat{\boldsymbol{\theta}}_s$ with a corresponding notation for the components of $\hat{\boldsymbol{\theta}}_r$. The null space of $\boldsymbol{X}$ characterises $\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r$, so we define

$$\Delta\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r, \; \Delta\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r, \; \Delta\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r \tag{27}$$

Figure 2 is a plot of $\Delta\hat{\boldsymbol{\alpha}}$, $\Delta\hat{\boldsymbol{\kappa}}$ and $\Delta\hat{\boldsymbol{\gamma}}$ and suggests that these values are linear with slopes C, $-C$ and C, respectively, for some C. Thus,

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r &= a\mathbf{1}_{n_a} + b\boldsymbol{x}_a \\ \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r &= c\mathbf{1}_{n_y} - b\boldsymbol{x}_y \\ \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r &= d\mathbf{1}_{n_c} + b\boldsymbol{x}_c \end{aligned}$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

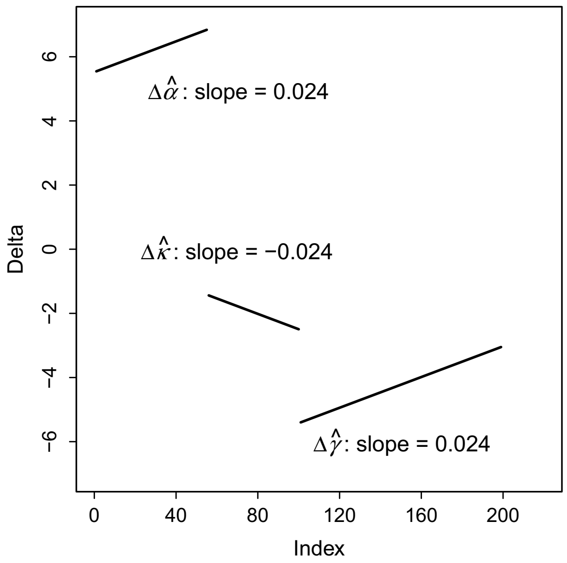

Figure 2. Values of $\Delta\hat{\boldsymbol{\alpha}}$, $\Delta\hat{\boldsymbol{\kappa}}$ and $\Delta\hat{\boldsymbol{\gamma}}$ defined in (27) in the APC model (23).

for some a, b, c and d. Since $\boldsymbol{n}_3$ is defined only up to scale, we may take $b = 1$ and conjecture that $\boldsymbol{n}_3$ has the form

$$\boldsymbol{n}_3 = \begin{pmatrix} a\mathbf{1}_{n_a} + \boldsymbol{x}_a \\ c\mathbf{1}_{n_y} - \boldsymbol{x}_y \\ d\mathbf{1}_{n_c} + \boldsymbol{x}_c \end{pmatrix} \tag{28}$$

Let $\boldsymbol{x}_1'$ be the first row of $\boldsymbol{X}$; we require $\boldsymbol{x}_1'\boldsymbol{n}_3 = 0$. We have

$$\boldsymbol{x}_1'\boldsymbol{n}_3 = (a+1) + (c-1) + (d+n_a) = a + c + d + n_a = 0 \tag{29}$$

Here, and subsequently, brackets used in this way indicate the terms for age, year and cohort, respectively, and help to clarify the argument. Any solution of (29) will do. A convenient choice is $a = c = 0$ and $d = -n_a$; our candidate $\boldsymbol{n}_3$ is

$$\boldsymbol{n}_3 = \begin{pmatrix} \boldsymbol{x}_a \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - n_a\mathbf{1}_{n_c} \end{pmatrix} \tag{30}$$

We check that $\boldsymbol{X}\boldsymbol{n}_3 = \mathbf{0}$ and conclude that $\{\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{n}_3\}$ is a basis for $\mathcal{N}(\boldsymbol{X})$, where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are defined in (25), and $\boldsymbol{n}_3$ in (30). Our estimates $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ then satisfy

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A\begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix} + B\begin{pmatrix} \mathbf{1}_{n_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix} + C\begin{pmatrix} \boldsymbol{x}_a \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - n_a\mathbf{1}_{n_c} \end{pmatrix} \tag{31}$$

for some scalars A, B and C; in our example, we found $A = 1.42$, $B = 4.10$ and $C = 0.024$. We note that (31) confirms what is widely known, namely that the age, period and cohort parameters are only estimable up to linear functions; see for example, Cairns *et al.* (2009).

We use the relationships (31) to investigate forecasting with the APC model. We suppose we forecast with an autoregressive integrated moving average model ARIMA(p, δ, q). We know from

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

(31) that both $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$, and $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$ differ by linear functions. We consider two cases: (a) forecasting the cohort effects with a simple random walk and (b) models with $\delta = 1$ or 2.

Case (a): Forecasting $\hat{\boldsymbol{\gamma}}$ with a simple random walk.

The constraints (26) imply that fitted values of $\boldsymbol{\gamma}$ have mean and slope of zero, and hence the simple random walk is a plausible model for forecasting cohort effects. We suppose that both $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$ are forecast in this way. Then a straightforward computation shows that the resulting forecasts of mortality are not equal, i.e., are not invariant with respect to the choice of constraints. The reason for this is that the forecasts of $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$ are now parallel and the linear relationship between $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$ in (31) has been broken.

We can use a simple device to illustrate further forecasting cohort effects with a simple random walk. If we fit an ARIMA(0,1,0) model, i.e., a random walk with drift, to $\hat{\boldsymbol{\gamma}}_s$, then the estimate of the drift parameter, μ_s say, is

$$\hat{\mu}_s = \frac{\hat{\gamma}_s(n_c) - \hat{\gamma}_s(1)}{n_c - 1} \tag{32}$$

We replace the standard constraints in (26) with

$$\sum_1^{n_y} \kappa_y = \sum_1^{n_c} \gamma_c = \gamma_{n_c} - \gamma_1 = 0 \tag{33}$$

We suppose the resulting estimates are denoted $\hat{\boldsymbol{\alpha}}_z$ $\hat{\boldsymbol{\kappa}}_z$ and $\hat{\boldsymbol{\gamma}}_z$. Now when we fit the ARIMA(0,1,0) model to $\hat{\boldsymbol{\gamma}}_z$, the estimate of the drift parameter is constrained to be zero by (33), i.e., we have fitted a simple random walk. We must forecast $\hat{\boldsymbol{\gamma}}_s$ with the same ARIMA(0,1,0) model; here, the estimate of the drift parameter is not zero, but the forecasts of mortality under both constraint systems are equal.

Case (b): Forecasting with an ARIMA(p, δ, q) model, $\delta = 1, 2$.

We know from (31) that

$$\hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r = -A\mathbf{1}_{n_y} - C\boldsymbol{x}_y \tag{34}$$

Thus $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ differ by a linear function. We should expect that any reasonable forecasts of $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$ obey the same functional relation. We show that this is indeed the case when an ARIMA(p, δ, q) model with a mean, $\delta = 1, 2$, is used to forecast. A formal proof of this is given in Appendix C. In summary, we will only consider forecasting models that obey this functional relationship since, as we shall see, this will preserve the invariance of forecasts with respect to the choice of constraints.

Let $\hat{\boldsymbol{\kappa}}_{s,f}$ and $\hat{\boldsymbol{\kappa}}_{r,f}$ denote the forecast values of $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$, respectively. It follows from (34) and (C.5) in Appendix C with $a = -A$ and $b = -C$ that

$$\hat{\boldsymbol{\kappa}}_{s,f} - \hat{\boldsymbol{\kappa}}_{r,f} = -(A + n_y C)\mathbf{1}_{n_f} - C\boldsymbol{x}_f \tag{35}$$

where $\boldsymbol{x}_f = (1, \ldots, n_f)'$. The analogous argument with $\hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$ shows that

$$\hat{\boldsymbol{\gamma}}_{s,f} - \hat{\boldsymbol{\gamma}}_{r,f} = -(B + (n_y - 1)C)\mathbf{1}_{n_f} + C\boldsymbol{x}_f \tag{36}$$

where $\hat{\boldsymbol{\gamma}}_{s,f}$ and $\hat{\boldsymbol{\gamma}}_{r,f}$ are the forecast values of $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$, respectively.

We can now establish the invariance of the forecast values of $\log \mu$. Define

$$\hat{\boldsymbol{\theta}}_{s,f} = (\hat{\boldsymbol{\alpha}}_s', \hat{\boldsymbol{\kappa}}_s', \hat{\boldsymbol{\kappa}}_{s,f}', \hat{\boldsymbol{\gamma}}_s', \hat{\boldsymbol{\gamma}}_{s,f}')'$$

with a similar definition for $\hat{\boldsymbol{\theta}}_{r,f}$. Let $\boldsymbol{X}_f$ be the model matrix for the APC model for ages $\boldsymbol{x}_a$ and years $(\boldsymbol{x}_y', \boldsymbol{x}_{y,f}')'$, where $\boldsymbol{x}_{y,f} = (n_y + 1, \ldots, n_y + n_f)'$. We can show that $\boldsymbol{X}_f(\hat{\boldsymbol{\theta}}_{s,f} - \hat{\boldsymbol{\theta}}_{r,f}) = \mathbf{0}$. We omit the proof and instead give a detailed proof in the next section for a smooth version of the

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

APCI model; the present case follows in the same fashion. We can therefore conclude that the fitted and forecast values of mortality in the APC model are invariant with respect to the choice of constraints when an ARIMA(p, δ, q) model, $\delta = 1$ or 2, with a mean is used to forecast.

A "computer proof" of the above result is simple and for many models of mortality is all that is reasonably available or indeed required. We compute fitted and forecast values under any two constraint systems and check the invariance of the fitted and forecast values. Once the basic code is written, different constraint systems and different forecasting regimes are easily applied. We have used this approach in parallel with our theoretical discussion.

4.2.2 Smooth APC model

Just as in the AP model, forecasting of mortality in the APC model is improved if the age parameters $\boldsymbol{\alpha}$ are smoothed. We set $\boldsymbol{\alpha} = \boldsymbol{B}_a \boldsymbol{a}$, where $\boldsymbol{B}_a$ is $n_a \times c_a$, and replace $\boldsymbol{X}_a$ in (24) with $\mathbf{1}_{n_y} \otimes \boldsymbol{B}_a$, as in (21). The model matrix becomes

$$\boldsymbol{X} = [\mathbf{1}_{n_y} \otimes \boldsymbol{B}_a : \boldsymbol{I}_{n_y} \otimes \mathbf{1}_{n_a} : \boldsymbol{X}_c],\ n_a n_y \times (c_a + n_a + 2n_y - 1) \tag{37}$$

and the penalty matrix is

$$\boldsymbol{P} = \text{blockdiag}\{\lambda_a \boldsymbol{D}'\boldsymbol{D}, \mathbf{0}\} \tag{38}$$

where $\boldsymbol{D}$ is the difference matrix of order d acting on the smoothed age coefficients $\boldsymbol{a}$, $\mathbf{0}$ is the $(n_y + n_c) \times (n_y + n_c)$ matrix of 0s acting on the unsmoothed year and cohort coefficients $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$, and λ_a is the smoothing parameter. With the P-spline system, it is usual to smooth with a second-order penalty. However, if a first-order penalty is used, we find the effective rank of the model is increased by one; in this case, in order to retain invariance the number of constraints should be reduced to two. We continue with a detailed discussion of the case when a second-order penalty is used.

The regression parameter is $\boldsymbol{\theta} = (\boldsymbol{a}', \boldsymbol{\kappa}', \boldsymbol{\gamma}')'$. It is easy to see that

$$\boldsymbol{n}_1 = \begin{pmatrix} \mathbf{1}_{c_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix}, \quad \boldsymbol{n}_2 = \begin{pmatrix} \mathbf{1}_{c_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix} \tag{39}$$

are in $\mathcal{N}(\boldsymbol{X})$. We use an extension of the argument used in the unsmoothed case to find a suitable third basis vector, $\boldsymbol{n}_3$. We fit with the standard constraints (26) and also with some random constraints but with $\lambda_a = 0$ (since we are interested in determining a basis for $\mathcal{N}(\boldsymbol{X})$). Corresponding to (27), we define

$$\Delta\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}_s - \hat{\boldsymbol{a}}_r,\ \Delta\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r,\ \Delta\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r \tag{40}$$

In our example, we used a knot spacing of $\Delta_a = 5$. From Figure 3, we note that $\Delta\hat{\boldsymbol{\kappa}}$ and $\Delta\hat{\boldsymbol{\gamma}}$ are linear with slopes 0.1754 and -0.1754, respectively, while $\Delta\hat{\boldsymbol{\alpha}}$ is linear with slope $-0.8768 \approx -\Delta_a 0.1754$. Hence, as in (28), we conjecture that $\boldsymbol{n}_3$ has the form

$$\boldsymbol{n}_3 = \begin{pmatrix} a\mathbf{1}_{c_a} + \Delta_a \boldsymbol{x}_{c_a} \\ c\mathbf{1}_{n_y} - \boldsymbol{x}_y \\ d\mathbf{1}_{n_c} + \boldsymbol{x}_c \end{pmatrix}$$

where $\boldsymbol{x}_{c_a} = (1, 2, 3, \ldots, c_a)'$. Let $\boldsymbol{x}_1'$ be the first row of $\boldsymbol{X}$; we set $\boldsymbol{x}_1'\boldsymbol{n}_3 = 0$. We recall that in the computation of $\boldsymbol{B}_a$, we place a knot at the first age. With this assumption, the first row of $\boldsymbol{B}_a$ is $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}, \mathbf{0}'_{c_a - 3})$. We have

$$\boldsymbol{x}_1'\boldsymbol{n}_3 = \big(a + \Delta_a\big(\tfrac{1}{6} + \tfrac{4}{3} + \tfrac{3}{6}\big)\big) + (c - 1) + (d + n_a) = a + c + d + n_a + 2\Delta_a - 1 = 0.$$

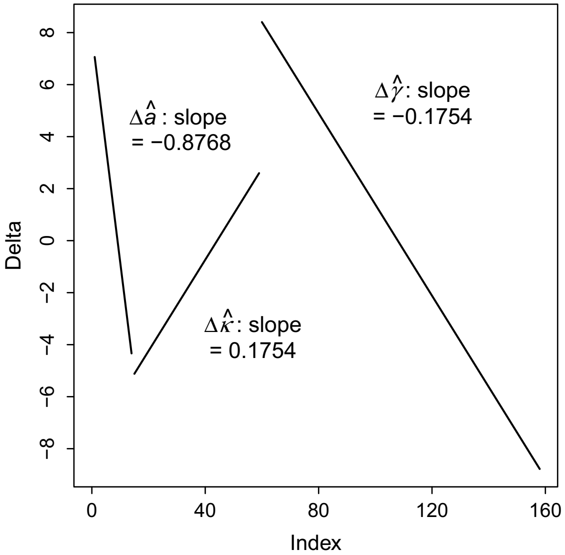

Figure 3. Values of $\Delta\hat{\boldsymbol{a}}$, $\Delta\hat{\boldsymbol{\kappa}}$ and $\Delta\hat{\boldsymbol{\gamma}}$ defined in (40) in the smooth APC model (37) but with $\lambda_a = 0$.

A convenient solution is $a = c = 0$ and $d = 1 - n_a - 2\Delta_a$; our candidate $\boldsymbol{n}_3$ is

$$\boldsymbol{n}_3 = \begin{pmatrix} \Delta_a \boldsymbol{x}_{c_a} \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - \omega_c \mathbf{1}_{n_c} \end{pmatrix} \tag{41}$$

where $\omega_c = -d = n_a + 2\Delta_a - 1$. We now check that $\boldsymbol{X}\boldsymbol{n}_3 = \mathbf{0}$. We conclude that $\{\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{n}_3\}$ is a basis for $\mathcal{N}(\boldsymbol{X})$, where $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ are defined in (39) and $\boldsymbol{n}_3$ in (41). Hence,

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A \begin{pmatrix} \mathbf{1}_{c_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix} + B \begin{pmatrix} \mathbf{1}_{c_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix} + C \begin{pmatrix} \Delta_a \boldsymbol{x}_{c_a} \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - \omega_c \mathbf{1}_{n_c} \end{pmatrix} \tag{42}$$

for some A, B and C; in our example, we found $A = 5.298$, $B = 2.644$ and $C = -0.1754$. Premultiplying (42) through by blockdiag$\{\boldsymbol{B}_a, \boldsymbol{I}_{n_y}, \boldsymbol{I}_{n_c}\}$ and using $\boldsymbol{B}_a \mathbf{1}_{c_a} = \mathbf{1}_{n_a}$ and $\boldsymbol{B}_a \Delta_a \boldsymbol{x}_{c_a} = \boldsymbol{x}_{n_a} + \omega_a \mathbf{1}_{n_a}$, $\omega_a = 2\Delta_a - 1$, we find

$$\hat{\boldsymbol{\theta}}^*_s - \hat{\boldsymbol{\theta}}^*_r = A \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \end{pmatrix} + B \begin{pmatrix} \mathbf{1}_{n_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \end{pmatrix} + C \begin{pmatrix} \boldsymbol{x}_a + \omega_a \mathbf{1}_{n_a} \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - \omega_c \mathbf{1}_{n_c} \end{pmatrix} \tag{43}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}', \boldsymbol{\kappa}', \boldsymbol{\gamma}')'$ denotes the parameters of interest.

We show that forecasting with an ARIMA(p, δ, q), $\delta = 1$ or 2, is invariant with respect to the choice of constraints. We use the same notation as in the unsmoothed case. Comparing (31) and (43), we see that the middle rows are identical. Hence, the forecasts of $\boldsymbol{\kappa}$ under two different constraint systems in the smoothed case obey the same relation (35) as in the unsmoothed case, i.e.,

$$\hat{\boldsymbol{\kappa}}_{s,f} - \hat{\boldsymbol{\kappa}}_{r,f} = -(A + n_y C)\mathbf{1}_{n_f} - C\boldsymbol{x}_f$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

Forecasts of $\boldsymbol{\gamma}$ obey the following

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{s,f} - \hat{\boldsymbol{\gamma}}_{r,f} &= \left(\hat{\boldsymbol{\gamma}}_s(n_y) - \hat{\boldsymbol{\gamma}}_r(n_y)\right)\mathbf{1}_{n_f} + C\boldsymbol{x}_{n_f}\\ &= (-B + (n_c - \omega_c)C)\mathbf{1}_{n_f} + C\boldsymbol{x}_{n_f} \quad \text{by (43)}\\ &= (-B + (n_y - 2\Delta_a)C)\mathbf{1}_{n_f} + C\boldsymbol{x}_{n_f}\end{aligned}$$

It now follows that $\boldsymbol{X}_f^*(\hat{\boldsymbol{\theta}}_{s,f}^* - \hat{\boldsymbol{\theta}}_{r,f}^*) = \mathbf{0}$, where $\boldsymbol{X}_f^*$ is the model matrix (24) in the unsmoothed case; the proof is a special case of that given for the smooth APCI model. We have established the invariance of the fitted and forecast values with respect to the choice of constraint system when an ARIMA(p, δ, q) model, $\delta = 1$ or 2, is used.

4.3 Age-period-cohort-improvements model

The CMI's age-period-cohort-improvements or APCI model was introduced as an industry standard for the forecasting of mortality (Continuous Mortality Investigation, 2016a, 2016b, 2016c; Richards *et al.*, to appear).

4.3.1 APCI model

The APCI model is

$$\log\, \mu_{x,y} = \alpha_x + \kappa_y + \gamma_{c(x,y)} + \beta_x(\bar{y} - y),\ x = 1, \ldots, n_a,\ y = 1, \ldots, n_y \tag{44}$$

where $\bar{y} = (n_y + 1)/2$ (recall the year vector is $\boldsymbol{x}_y = (1, \ldots, n_y)'$). We note that (44) differs in a minor way from the CMI's formulation; we have reversed the sign of the β_x terms, which now correspond to the β_x terms in the Lee–Carter model. The thinking behind the APCI model is as follows: linearise the $\beta_x\kappa_y$ term in the Lee–Carter model and add the cohort term $\gamma_{c(x,y)}$. This gives a model which takes into account any cohort effects and allows the forecast of the year effect to depend on age. The APCI model is a GLM with model matrix

$$\boldsymbol{X} = [\boldsymbol{X}_a : \boldsymbol{X}_y : \boldsymbol{X}_c : \boldsymbol{X}_b],\ n_a n_y \times (3n_a + 2n_y - 1) \tag{45}$$

where $\boldsymbol{X}_a$ and $\boldsymbol{X}_y$ are defined in (17), $\boldsymbol{X}_c$ in (24) and $\boldsymbol{X}_b = (\bar{y}\mathbf{1}_{n_y} - \boldsymbol{x}_y) \otimes \boldsymbol{I}_{n_a}$. We define $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}', \boldsymbol{\gamma}', \boldsymbol{\beta}')'$ and the model has the form (1). The model matrix has $3n_a + 2n_y - 1$ columns and rank $3n_a + 2n_y - 6$, so five constraints are required to enable $\boldsymbol{\theta}$ to be estimated uniquely. The null space of $\boldsymbol{X}$ is spanned by

$$\begin{pmatrix}\mathbf{1}_{n_a}\\ -\mathbf{1}_{n_y}\\ \mathbf{0}_{n_c}\\ \mathbf{0}_{n_a}\end{pmatrix},\ \begin{pmatrix}\mathbf{1}_{n_a}\\ \mathbf{0}_{n_y}\\ -\mathbf{1}_{n_c}\\ \mathbf{0}_{n_a}\end{pmatrix},\ \begin{pmatrix}\boldsymbol{x}_a\\ -\boldsymbol{x}_y\\ \boldsymbol{x}_c - n_a\mathbf{1}_{n_c}\\ \mathbf{0}_{n_a}\end{pmatrix},\ \begin{pmatrix}\mathbf{0}_{n_a}\\ \boldsymbol{x}_y - \bar{y}\mathbf{1}_{n_y}\\ \mathbf{0}_{n_c}\\ \mathbf{1}_{n_a}\end{pmatrix},\ \begin{pmatrix}Q\boldsymbol{x}_a + \boldsymbol{x}_a^2\\ 2n_a\boldsymbol{x}_y + \boldsymbol{x}_y^2\\ n_a^2\mathbf{1}_{n_c} - \boldsymbol{x}_c^2\\ 2\boldsymbol{x}_a\end{pmatrix} \tag{46}$$

where $Q = -2(n_a + \bar{y})$. The first three basis vectors in (46) are the same as the basis vectors for the APC model with $\mathbf{0}_{n_a}$ appended. The fourth basis vector can be deduced by observing that

$$\kappa_y + (\bar{y} - y)\beta_x = \kappa_y - (\bar{y} - y)\omega + (\bar{y} - y)(\beta_x + \omega) = \kappa_y^* + (\bar{y} - y)\beta_x^*$$

for any scalar ω. The fifth basis vector can be found with the same method as used to find the third basis vector for the APC model. We denote the basis vectors in (46) by $\boldsymbol{n}_i,\ i = 1, \ldots, 5$.

We comment briefly on the CMI's approach to identifiability. The CMI (2016a, 7.3) gives the following transformation of the parameters which leaves the values of log $\boldsymbol{\mu}$ unchanged. We give this transformation in our notation and with the CMI's β_x replaced by ours.

$$\begin{pmatrix} \Delta\alpha_x \\ \Delta\kappa_y \\ \Delta\gamma_{y-x} \\ \Delta\beta_x \end{pmatrix} = \theta_1 \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} + \theta_2 \begin{pmatrix} -(x-\bar{x}) \\ y-\bar{y} \\ (x-\bar{x})-(y-\bar{y}) \\ 0 \end{pmatrix} +$$

$$\theta_3 \begin{pmatrix} (x-\bar{x})^2 \\ (y-\bar{y})^2 \\ -\big((x-\bar{x})-(y-\bar{y})\big)^2 \\ 2(x-\bar{x}) \end{pmatrix} + \theta_4 \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} + \theta_5 \begin{pmatrix} 0 \\ -(y-\bar{y}) \\ 0 \\ (y-\bar{y}) \end{pmatrix}$$

Writing the transformation in this way, we see that it is equivalent to a particular null basis for $\boldsymbol{X}$, namely,

$$\begin{pmatrix} \mathbf{1}_{n_a} \\ \mathbf{0}_{n_y} \\ -\mathbf{1}_{n_c} \\ \mathbf{0}_{n_a} \end{pmatrix}, \begin{pmatrix} -(\boldsymbol{x}_a-\bar{a}\mathbf{1}_{n_a}) \\ \boldsymbol{x}_y-\bar{y}\mathbf{1}_{n_y} \\ -(\boldsymbol{x}_c-\bar{c}\mathbf{1}_{n_c}) \\ \mathbf{0}_{n_a} \end{pmatrix}, \begin{pmatrix} (\boldsymbol{x}_a-\bar{a}\mathbf{1}_{n_a})^2 \\ (\boldsymbol{x}_y-\bar{y}\mathbf{1}_{n_y})^2 \\ -(\boldsymbol{x}_c-\bar{c}\mathbf{1}_{n_c})^2 \\ 2(\boldsymbol{x}_a-\bar{a}\mathbf{1}_{n_a}) \end{pmatrix}, \begin{pmatrix} \mathbf{1}_{n_a} \\ -\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \\ \mathbf{0}_{n_a} \end{pmatrix}, \begin{pmatrix} \mathbf{0}_{n_a} \\ \boldsymbol{x}_y-\bar{y}\mathbf{1}_{n_y} \\ \mathbf{0}_{n_c} \\ \mathbf{1}_{n_a} \end{pmatrix} \tag{47}$$

where $\bar{a}$ is the mean age, $\bar{y}$ is the mean year and $\bar{c} = n_a - \bar{a} + \bar{y}$. Three of the basis vectors are the same as ours; the other two have the same form. In particular, the third basis vector in (47) has quadratic and linear functions in the equivalent positions to (46).

We fit the model under the standard constraints (CMI, 2016a)

$$\sum \kappa_y = \sum \gamma_c = \sum c\gamma_c = \sum c^2\gamma_c = \sum y\kappa_y = 0 \tag{48}$$

where c is the cohort index, $c = 1, \ldots, n_c$; again, we could use weighted constraints as in Richards *et al.* (to appear). We also use the random constraints

$$\sum u_{i,j}\theta_j = 0, \; i = 1, \ldots, 5, \; j = 1, \ldots, 3n_a + 2n_y - 1 \tag{49}$$

where the $u_{i,j}$ are independent $\mathcal{U}(0, 1)$. With our usual notation for the estimates of $\boldsymbol{\theta}$ under the standard and random constraints, we have

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A\boldsymbol{n}_1 + B\boldsymbol{n}_2 + C\boldsymbol{n}_3 + D\boldsymbol{n}_4 + E\boldsymbol{n}_5$$

for some A, B, C, D and E. Equating coefficients in (46), we immediately find that $\hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r$, $\hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r$ and $\hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$ are quadratic functions of $\boldsymbol{x}_a$, $\boldsymbol{x}_y$ and $\boldsymbol{x}_c$, respectively, with E, E and $-E$ as the coefficients of the quadratic terms; furthermore, $\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_r$ is linear with slope $2E$. These relationships have implications for invariance when it comes to forecasting.

Figure 4 confirms the relationships among the coefficients. Since $\Delta\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r$ is quadratic with the coefficient of the quadratic term equal to E, first differences of $\Delta\hat{\boldsymbol{\alpha}}$, $D(\Delta\hat{\boldsymbol{\alpha}})$, will be linear with slope $2E$ as shown in Figure 4; the same remark applies to the first differences $D(\Delta\hat{\boldsymbol{\kappa}})$ and $D(\Delta\hat{\boldsymbol{\gamma}})$. The slope of $\Delta\hat{\boldsymbol{\beta}}$ is $2E$. In our example, we found $A = -624.3$, $B = -70.02$, $C = 18.24$, $D = 20.32$ and $E = 0.002385$.

Figure 5 is a plot of the estimated $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under our two constraint systems; the estimates are wildly different both in shape and range. Under the standard constraint, the estimate of $\boldsymbol{\alpha}$ has a

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

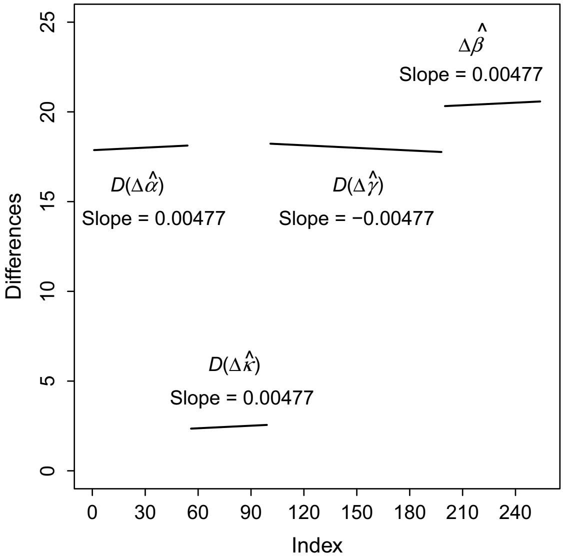

Figure 4. Plot of first differences, denoted D, of $\Delta\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r$, $\Delta\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r$ and $\Delta\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$; plot of $\Delta\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_r$ in the APCI model (44).

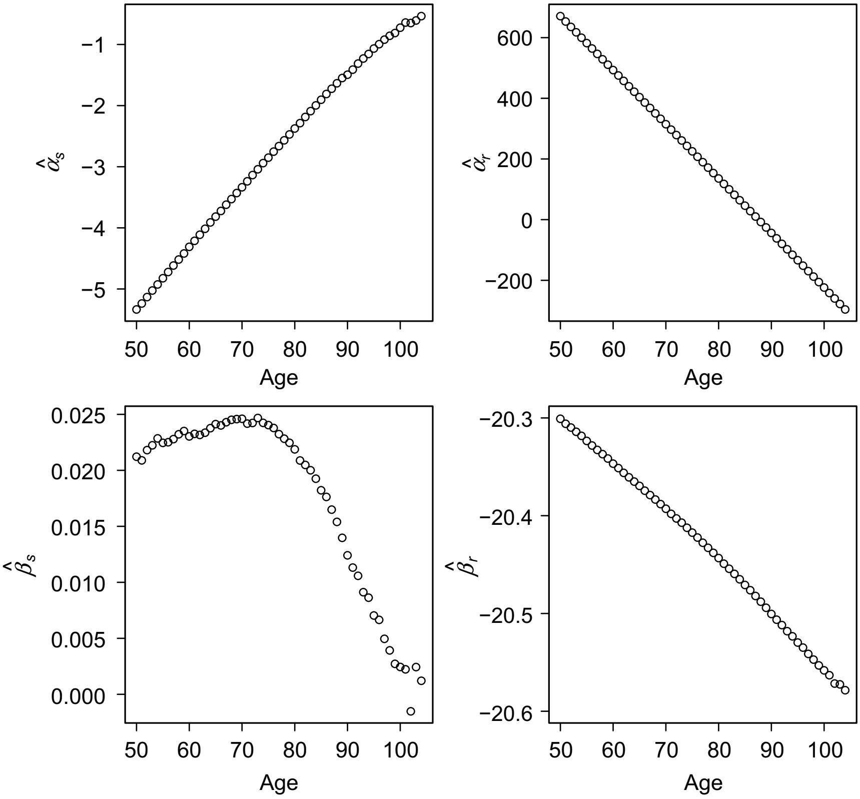

Figure 5. Left: plot of $\hat{\boldsymbol{\alpha}}_s$, $\hat{\boldsymbol{\beta}}_s$; right: plot of $\hat{\boldsymbol{\alpha}}_r$ and $\hat{\boldsymbol{\beta}}_r$ in the APCI model (44).

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

familiar look, while the estimate of $\boldsymbol{\beta}$ is very close to the estimate of $\boldsymbol{\beta}$ in the Lee–Carter model; see Figure 8 for the corresponding plot. However, Figure 5 serves to emphasise our main point: it is how the coefficients join forces that matter, not the individual coefficients themselves.

The CMI smooth both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in their treatment of the APCI model, and we will focus on this case. First, we discuss briefly the unsmoothed case. We forecast with our two constraint systems, namely standard and random. The quadratic nature of the fifth basis vector in the null space of the APCI model (see (46)) implies that a first-order ARIMA model will not give invariant forecasts. In this paper, we concentrate on forecasting models which lead to invariant forecasts so we do not pursue this further.

4.3.2 Smooth APCI model

The case for smoothing both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is strong: forecasts will be much less prone to irregular behaviour by age (see Delwarde *et al.*, 2007; Currie, 2013). We will not smooth either $\boldsymbol{\kappa}$ or $\boldsymbol{\gamma}$, since we are interested in stochastic forecasts of these parameters. This is in contrast to the CMI which in addition to smoothing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ also smooth $\boldsymbol{\kappa}$ or $\boldsymbol{\gamma}$; the CMI's approach is designed to facilitate deterministic targeting of future mortality. We do not comment on this debate here; see Booth & Tickle (2008), Richards *et al.* (to appear).

Let $\boldsymbol{B}_a$, $n_a \times c_a$, be a regression matrix evaluated over a basis of cubic B-splines for age. We use the same regression matrix to smooth both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and set $\boldsymbol{\alpha} = \boldsymbol{B}_a \boldsymbol{a}$ and $\boldsymbol{\beta} = \boldsymbol{B}_a \boldsymbol{b}$. The model matrix is now

$$\boldsymbol{X} = [\boldsymbol{1}_{n_y} \otimes \boldsymbol{B}_a : \boldsymbol{I}_{n_y} \otimes \boldsymbol{1}_{n_a} : \boldsymbol{X}_c : (\bar{y}\boldsymbol{1}_{n_y} - \boldsymbol{x}_y) \otimes \boldsymbol{B}_a] \tag{50}$$

where $\boldsymbol{X}_c$ is defined below (24). The model matrix is $n_a\, n_y \times (2c_a + n_y + n_c)$ with $r(\boldsymbol{X}) = 2c_a + n_y + n_c - 5$, so $\mathcal{N}(\boldsymbol{X})$ has dimension five. We fit model (50) but without smoothing, i.e., with the smoothing parameters set to zero. Let $\boldsymbol{\theta} = (\boldsymbol{a}', \boldsymbol{\kappa}', \boldsymbol{\gamma}', \boldsymbol{b}')'$ be the vector of regression coefficients and $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}', \boldsymbol{\kappa}', \boldsymbol{\gamma}', \boldsymbol{\beta}')'$ denote the parameters of interest. We denote the five vectors

$$\begin{pmatrix} \boldsymbol{1}_{c_a} \\ -\boldsymbol{1}_{n_y} \\ \boldsymbol{0}_{n_c} \\ \boldsymbol{0}_{c_a} \end{pmatrix}, \begin{pmatrix} \boldsymbol{1}_{c_a} \\ \boldsymbol{0}_{n_y} \\ -\boldsymbol{1}_{n_c} \\ \boldsymbol{0}_{c_a} \end{pmatrix}, \begin{pmatrix} \Delta_a \boldsymbol{x}_{c_a} \\ -\boldsymbol{x}_y \\ \boldsymbol{x}_c - \omega_c \boldsymbol{1}_{n_c} \\ \boldsymbol{0}_{c_a} \end{pmatrix}, \begin{pmatrix} \boldsymbol{0}_{c_a} \\ \boldsymbol{x}_y - \bar{y}\boldsymbol{1}_{n_y} \\ \boldsymbol{0}_{n_c} \\ \boldsymbol{1}_{c_a} \end{pmatrix}, \begin{pmatrix} Q_a(\boldsymbol{x}_{c_a}) \\ Q_\kappa(\boldsymbol{x}_{n_y}) \\ Q_\gamma(\boldsymbol{x}_{n_c}) \\ L_b(\boldsymbol{x}_{c_a}) \end{pmatrix} \tag{51}$$

by $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_5$, respectively; here $Q_a(\,\cdot\,)$, $Q_\kappa(\,\cdot\,)$ and $Q_\gamma(\,\cdot\,)$ are quadratic functions and $L_b(\,\cdot\,)$ is a linear function. We can check that $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_4$ are all in $\mathcal{N}(\boldsymbol{X})$. Figure 6 is a plot of differences in estimates under the standard and random constraints (48) and (49) and suggests that the fifth basis vector in $\mathcal{N}(\boldsymbol{X})$ has the form $\boldsymbol{n}_5$. Additionally, the form of the fifth basis in (46) in the unsmoothed case supports this idea. We will not need an exact expression for $\boldsymbol{n}_5$.

Let $\boldsymbol{D}_a$ and $\boldsymbol{D}_b$ be difference matrices of orders d_a and d_b applied to the coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively. The penalty matrix is

$$\boldsymbol{P} = \text{blockdiag}\{\lambda_a \boldsymbol{D}_a' \boldsymbol{D}_a,\ \boldsymbol{0},\ \lambda_b \boldsymbol{D}_b' \boldsymbol{D}_b\}$$

where $\boldsymbol{0}$ is a square matrix of 0s of size $n_y + n_c$. The null space of the penalised model is given by $\mathcal{N}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{P})$; see (13). Now $\boldsymbol{P}\boldsymbol{n}_i = \boldsymbol{0}$ for $i = 1, \ldots, 4$ for $d_a \geq 2$ and $d_b \geq 2$; for $\boldsymbol{P}\boldsymbol{n}_5 = \boldsymbol{0}$, we need $d_a \geq 3$ and $d_b \geq 2$. It is usual to smooth with difference matrices of order two; in this case $\boldsymbol{n}_5 \notin \mathcal{N}(\boldsymbol{P})$. We continue with the case $d_a = d_b = 2$, in which case the null space of the model is given by the first four vectors in (51). Direct computation of $r(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})$ confirms that the effective rank of the smooth APCI model with $d_a = d_b = 2$ is $2c_a + n_y + n_c - 4$. We conclude that $\{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_4\}$ is a basis for the null space of the smooth model.

We make a brief comment on the effect of the choice of order of the penalty. We take $d_a = d_b = 2$. Under the P-spline system of smoothing, the rank of the model is in effect $r(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{P})$,

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

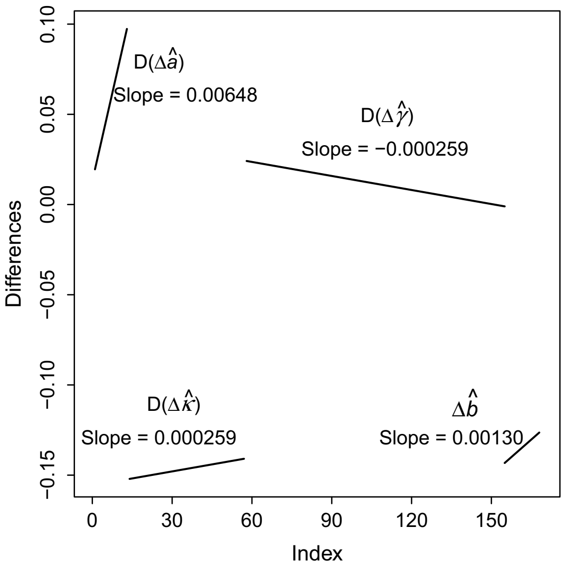

Figure 6. Plot of first differences, denoted D, of $\Delta\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}_s - \hat{\boldsymbol{a}}_r$, $\Delta\hat{\kappa} = \hat{\kappa}_s - \hat{\kappa}_r$ and $\Delta\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$; plot of $\Delta\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}_s - \hat{\boldsymbol{b}}_r$ in APCI model with the smooth model matrix (50) but with $\lambda_a = \lambda_b = 0$.

while the rank of the model without smoothing is $r(\boldsymbol{X}'\boldsymbol{X})$; denote these ranks by r_s and r, respectively. Usually, $r = r_s$; indeed we are not aware that the phenomenon $r_s > r$ has been previously observed. Certainly, the implicit constraint in the smooth model can lead to very different estimates of the parameters. We argue that since $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are not identifiable, we should not be concerned about these differences. The fitted force of mortalities (which are identifiable) under the two models will be very close.

We fit the model with four standard constraints (in (48) we drop the constraint $\sum c^2\gamma_c = 0$) and four random constraints. Let $\hat{\boldsymbol{\theta}}_s$ and $\hat{\boldsymbol{\theta}}_r$ be the estimates of $\boldsymbol{\theta}$ as usual. Then, from (51), we have

$$\hat{\boldsymbol{\theta}}_s - \hat{\boldsymbol{\theta}}_r = A\begin{pmatrix}\boldsymbol{1}_{c_a}\\ -\boldsymbol{1}_{n_y}\\ \boldsymbol{0}_{n_c}\\ \boldsymbol{0}_{c_a}\end{pmatrix} + B\begin{pmatrix}\boldsymbol{1}_{c_a}\\ \boldsymbol{0}_{n_y}\\ -\boldsymbol{1}_{n_c}\\ \boldsymbol{0}_{c_a}\end{pmatrix} + C\begin{pmatrix}\Delta_a\boldsymbol{x}_{c_a}\\ -\boldsymbol{x}_y\\ \boldsymbol{x}_c - \omega_c\boldsymbol{1}_{n_c}\\ \boldsymbol{0}_{c_a}\end{pmatrix} + D\begin{pmatrix}\boldsymbol{0}_{c_a}\\ \boldsymbol{x}_y - \bar{y}\boldsymbol{1}_{n_y}\\ \boldsymbol{0}_{n_c}\\ \boldsymbol{1}_{c_a}\end{pmatrix} \tag{52}$$

for some A, B, C and D; here, as before, $\omega_c = n_a + 2\Delta_a - 1$. In our example, we found $A = 5.742$, $B = -0.701$, $C = -0.0343$ and $D = -0.523$. Pre-multiplying (52) through by blockdiag$\{\boldsymbol{B}_a, \boldsymbol{I}_{n_y}, \boldsymbol{I}_{n_c}, \boldsymbol{B}_a\}$ and using $\boldsymbol{B}_a\boldsymbol{1}_{c_a} = \boldsymbol{1}_{n_a}$ and $\boldsymbol{B}_a\Delta_a\boldsymbol{x}_{c_a} = \boldsymbol{x}_{n_a} + \omega_a\boldsymbol{1}_{n_a}$, we find

$$\hat{\boldsymbol{\theta}}_s^* - \hat{\boldsymbol{\theta}}_r^* = A\begin{pmatrix}\boldsymbol{1}_{n_a}\\ -\boldsymbol{1}_{n_y}\\ \boldsymbol{0}_{n_c}\\ \boldsymbol{0}_{n_a}\end{pmatrix} + B\begin{pmatrix}\boldsymbol{1}_{n_a}\\ \boldsymbol{0}_{n_y}\\ -\boldsymbol{1}_{n_c}\\ \boldsymbol{0}_{n_a}\end{pmatrix} + C\begin{pmatrix}\boldsymbol{x}_a + \omega_a\boldsymbol{1}_{n_a}\\ -\boldsymbol{x}_y\\ \boldsymbol{x}_c - \omega_c\boldsymbol{1}_{n_c}\\ \boldsymbol{0}_{n_a}\end{pmatrix} + D\begin{pmatrix}\boldsymbol{0}_{n_a}\\ \boldsymbol{x}_y - \bar{y}\boldsymbol{1}_{n_y}\\ \boldsymbol{0}_{n_c}\\ \boldsymbol{1}_{n_a}\end{pmatrix} \tag{53}$$

where as before $\omega_a = 2\Delta_a - 1$. It follows immediately that

$$\hat{\boldsymbol{\alpha}}_s - \hat{\boldsymbol{\alpha}}_r = C\boldsymbol{x}_a + (A + B + w_aC)\boldsymbol{1}_{n_a} \tag{54}$$

$$\hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r = (D - C)\boldsymbol{x}_y - (A + D\bar{y})\boldsymbol{1}_{n_y} \tag{55}$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

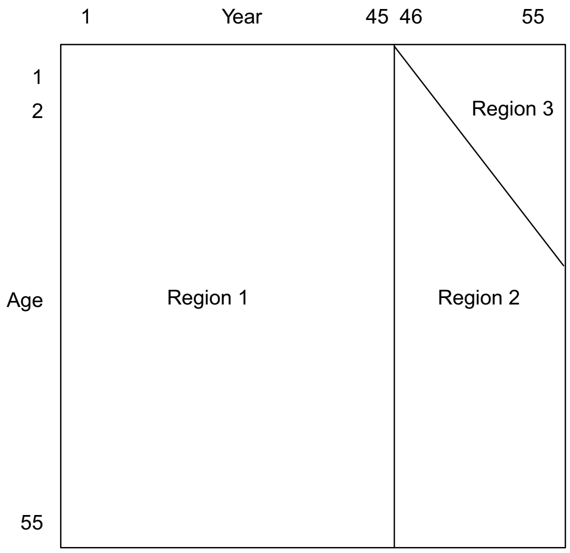

Figure 7. Three regions for fitted and forecast values of $\log \mu$ for ages $1, \ldots, 55$, years $1, \ldots, 45$, and a 10-year-ahead forecast.

$$\hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r = C\boldsymbol{x}_c - (B + w_c C)\mathbf{1}_{n_c} \tag{56}$$

$$\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_r = D\mathbf{1}_{n_a} \tag{57}$$

We now turn to forecasting with the smooth APCI model. We use the same notation as in the APC model. Let n_f be the length of the forecast and $\boldsymbol{x}_f = (1, \ldots, n_f)'$. Let $\hat{\boldsymbol{\kappa}}_{s,f}$ and $\hat{\boldsymbol{\kappa}}_{r,f}$ be the forecast values of $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$, and $\hat{\boldsymbol{\gamma}}_{s,f}$ and $\hat{\boldsymbol{\gamma}}_{r,f}$ be the forecast values of $\hat{\boldsymbol{\gamma}}_s$ and $\hat{\boldsymbol{\gamma}}_r$, respectively. We know from (55) and (56) that $\hat{\boldsymbol{\kappa}}_s - \hat{\boldsymbol{\kappa}}_r$ and $\hat{\boldsymbol{\gamma}}_s - \hat{\boldsymbol{\gamma}}_r$ are both linear so, exactly as in the APC model, we can show that

$$\hat{\boldsymbol{\kappa}}_{s,f} - \hat{\boldsymbol{\kappa}}_{r,f} = (n_y(D - C) - (A + D\bar{y}))\mathbf{1}_{n_f} + (D - C)\boldsymbol{x}_f \tag{58}$$

$$\hat{\boldsymbol{\gamma}}_{s,f} - \hat{\boldsymbol{\gamma}}_{r,f} = ((n_c - w_c)C - B)\mathbf{1}_{n_f} + C\boldsymbol{x}_f \tag{59}$$

We now establish the invariance of the forecast values of $\log \boldsymbol{\mu}$. Define

$$\hat{\boldsymbol{\theta}}^*_{s,f} = (\hat{\boldsymbol{\alpha}}'_s, \hat{\boldsymbol{\kappa}}'_s, \hat{\boldsymbol{\kappa}}'_{s,f}, \hat{\boldsymbol{\gamma}}'_s, \hat{\boldsymbol{\gamma}}'_{s,f}, \hat{\boldsymbol{\beta}}'_s)'$$

with a similar definition for $\hat{\boldsymbol{\theta}}^*_{r,f}$. Let $\boldsymbol{X}_f$ be the model matrix for the smooth APCI model for ages $\boldsymbol{x}_a$ and years $(\boldsymbol{x}'_y, \boldsymbol{x}'_{y,f})'$, where $\boldsymbol{x}_{y,f} = (n_y + 1, \ldots, n_y + n_f)'$; let $\boldsymbol{X}^*_f$ be the corresponding model matrix for the unsmoothed model. We wish to show that $\boldsymbol{X}_f(\hat{\boldsymbol{\theta}}_{s,f} - \hat{\boldsymbol{\theta}}_{r,f}) = \mathbf{0}$. We observe that $\boldsymbol{X}_f\hat{\boldsymbol{\theta}}_{s,f} = \boldsymbol{X}^*_f\hat{\boldsymbol{\theta}}^*_{s,f}$, since $(\mathbf{1}_{n_y} \otimes \boldsymbol{B}_a)\hat{\boldsymbol{a}} = (\mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a})\hat{\boldsymbol{\alpha}}$ and $((\bar{y}\mathbf{1}_{n_y} - \boldsymbol{x}_y) \otimes \boldsymbol{B}_a)\hat{\boldsymbol{b}} = ((\bar{y}\mathbf{1}_{n_y} - \boldsymbol{x}_y) \otimes \boldsymbol{I}_{n_a})\hat{\boldsymbol{\beta}}$.

There are three cases to consider, as shown in Figure 7. In region 1, the fitted values depend on the estimated values of $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$; in region 2, the forecast values depend on the estimated values of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and the forecast values of $\boldsymbol{\kappa}$; and in region 3, the forecast values depend on the estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and the forecast values of $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$. We see from (54), (55), (56), (57), (58) and (59) that

$$\boldsymbol{X}^*_f(\hat{\boldsymbol{\theta}}^*_{s,f} - \hat{\boldsymbol{\theta}}^*_{r,f}) = A\boldsymbol{v}_1 + B\boldsymbol{v}_2 + C\boldsymbol{v}_3 + D\boldsymbol{v}_4$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

for some vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ and $\boldsymbol{v}_4$. We show that $\boldsymbol{v}_1 = \boldsymbol{v}_2 = \boldsymbol{v}_3 = \boldsymbol{v}_4 = \boldsymbol{0}$; this will establish invariance of fitted and forecast values.

It is convenient to use a double-suffix notation to identify the entries in $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ and $\boldsymbol{v}_4$. Thus, cell (i, j) is associated with $\boldsymbol{v}_1(i, j)$ and corresponds to the $n_a(j-1)+i$ entry in $\boldsymbol{v}_1$; we have a similar correspondence for $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ and $\boldsymbol{v}_4$. We check invariance region by region.

Consider region 1. Suppose cell (i, j) lies in region 1 with cohort index $n_a - i + j$. We pick out the terms in A, B, C and D from (54), (55), (56) and (57) as appropriate. The bracket notation ( ) indicates which terms in $\boldsymbol{X}_f^*(\hat{\boldsymbol{\theta}}^*_{s,f} - \hat{\boldsymbol{\theta}}^*_{r,f})$ are used, i.e., $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in turn,

$$A : (1) + (-1) + (0) + (0) = 0$$

$$B : (1) + (0) + (-1) + (0) = 0$$

$$C : (i + \omega_a) + (-j) + (n_a - i + j - \omega_c) + (0) = 0$$

$$D : (0) + (j - \bar{y}) + (0) + (\bar{y} - j) = 0$$

and we have verified invariance in region 1. Of course, we already knew this from the general result on fitted values.

In region 2, we consider the j-ahead forecast for $\boldsymbol{\kappa}$. This is cell $(i, n_y + j)$ with cohort index $n_a - i + n_y + j$. Using (54), (56), (57) and (58), we find

$$A : (1) + (-1) + (0) + (0) = 0$$

$$B : (1) + (0) + (-1) + (0) = 0$$

$$C : (i + \omega_a) + (-n_y - j) + (n_a - i + n_y + j - \omega_c) + (0) = 0$$

$$D : (0) + (n_y - \bar{y} + j) + (0) + (\bar{y} - n_y - j) = 0$$

as required. Finally, in region 3, we consider the j-ahead forecast for $\boldsymbol{\gamma}$. For age i, we have the $(i + j - 1)$-ahead forecast for $\boldsymbol{\kappa}$, i.e., cell $(i, n_y + i + j - 1)$. We use (54), (57), (58) and (59) and find

$$A : (1) + (-1) + (0) + (0) = 0$$

$$B : (1) + (0) + (-1) + (0) = 0$$

$$C : (i + \omega_a) + (-n_y - i - j + 1) + (n_c - \omega_c + j) + (0) = 0$$

$$D : (0) + (n_y - \bar{y} + i + j - 1) + (0) + (\bar{y} - n_y - i - j + 1) = 0$$

We conclude that the fitted and forecast values are invariant with respect to the choice of constraints when an ARIMA(p, δ, q) model with fitted mean, $\delta = 1$ or 2, is used to forecast. We note that the proof of invariance in the smooth APC model is a special case of the above. In (52), we omit the final row and column, and we have reduced (52) to (42).

4.4 Lee–Carter model

The Lee–Carter model or LC model (Lee & Carter, 1992) is the benchmark model for modelling and forecasting mortality. The LC model is a non-linear model so the methods of the previous sections do not apply. However, we can still use our extended algorithm (15) to fit two forms of the model: Lee and Carter's original formulation and a smooth version.

4.4.1 LC model

The LC model is

$$\log \mu_{x,y} = \alpha_x + \beta_x \kappa_y, \; x = 1, \ldots, n_a, \; y = 1, \ldots, n_y \tag{60}$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

We require one location and one scale constraint to make the parameters estimable. We use

$$\sum \kappa_y = 0, \ \sum \beta_x = 1 \tag{61}$$

as in Lee & Carter (1992). The following transformation leaves the fitted values of log $\boldsymbol{\mu}$ unchanged for any values of A and B.

$$\alpha_x + \beta_x \kappa_y \mapsto (\alpha_x + A\beta_x) + (\beta_x/B)(B\kappa_y - AB) = \alpha_x^* + \beta_x^* \kappa_y^* \tag{62}$$

Let $\hat{\boldsymbol{\alpha}}_s$, $\hat{\boldsymbol{\beta}}_s$ and $\hat{\boldsymbol{\kappa}}_s$ be the maximum likelihood estimates of $\boldsymbol{\alpha}_s$, $\boldsymbol{\beta}_s$ and $\boldsymbol{\kappa}_s$ under the constraints (61) with the usual corresponding notation for the estimates $\hat{\boldsymbol{\alpha}}_r$, $\hat{\boldsymbol{\beta}}_r$ and $\hat{\boldsymbol{\kappa}}_r$ under some random constraint system. It follows from (62) that

$$\hat{\boldsymbol{\alpha}}_s = \hat{\boldsymbol{\alpha}}_r + A\hat{\boldsymbol{\beta}}_r \tag{63}$$

$$\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_r / B \tag{64}$$

$$\hat{\boldsymbol{\kappa}}_s = B\hat{\boldsymbol{\kappa}}_r - AB\mathbf{1}_{n_y} \tag{65}$$

for some scalars A and B.

Brouhns *et al.* (2002) used maximum likelihood to estimate the parameters. We also use maximum likelihood but use (15); this is a full Newton–Raphson scheme which allows estimation with both smoothing and constraints. Following Currie (2013), we consider two coupled GLMs:

$$\text{GLM1: } \log \boldsymbol{\mu} = \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}_1 \boldsymbol{\beta}, \ \boldsymbol{X}_1 = \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a} \tag{66}$$

$$\text{GLM2: } \log \boldsymbol{\mu} = \boldsymbol{X}_2 \boldsymbol{\theta}, \ \boldsymbol{X}_2 = [\mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a} : \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}] \tag{67}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}')'$; here $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\kappa}}$ represent current estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$, respectively. Both GLM1 and GLM2 are in the class defined in equation (1) but note that in GLM1 the offset is $\log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}}$. In GLM1, we estimate $\boldsymbol{\beta}$ for current values of $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$, while in GLM2 we estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ for given values of $\boldsymbol{\beta}$. We now iterate between GLM1 and GLM2 until convergence.

We fit the model with (a) the standard constraints (61) and (b) the random constraints in GLM1 on $\boldsymbol{\beta}$, i.e., $\sum_1^{n_a} u_i \beta_i = 1$, and the random constraints in GLM2 on $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa})'$, i.e., $\sum_1^{n_a+n_y} u_i \theta_i = 0$; the u_i are independent $\mathcal{U}(0,\ 1)$.

The upper panels and the lower left panel of Figure 8 show the estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$, respectively, under the standard and random constraints. In our example, we found $A = 5.907$, $B = 1.983$ and $AB = 11.71$. The parameter estimates in Figure 8 satisfy (63), (64) and (65) with these values of A and B. The lower right panel shows the single (invariant) estimate of log mortality at age 65.

Forecasting in the LC model is particularly straightforward. As usual, we forecast with an ARIMA(p, δ, q) model, $\delta = 1$ or 2. Let $\hat{\boldsymbol{\kappa}}_{s,f}$ and $\hat{\boldsymbol{\kappa}}_{r,f}$ denote some forecast values of $\hat{\boldsymbol{\kappa}}_s$ and $\hat{\boldsymbol{\kappa}}_r$, respectively. We have from (65) $\hat{\boldsymbol{\kappa}}_s = B\hat{\boldsymbol{\kappa}}_r - AB\mathbf{1}_{n_y}$; it follows immediately from Appendix C that $\hat{\boldsymbol{\kappa}}_{s,f} = B\hat{\boldsymbol{\kappa}}_{r,f} - AB\mathbf{1}_{n_f}$, where n_f is the number of years in the forecast. Thus, the forecasts of $\boldsymbol{\kappa}$ satisfy (65), and hence forecasts of mortality are invariant with respect to the choice of constraint system.

4.4.2 Smooth LC model

One unsatisfactory feature of the Lee–Carter model, particularly for actuaries, is that irregularities in the estimate of $\boldsymbol{\beta}$ lead to irregular forecasts of mortality at individual ages and can even lead to forecasts at adjacent ages crossing over; see Currie (2013) for examples of this behaviour. Delwarde *et al.* (2007) addressed this problem by smoothing the estimate of $\boldsymbol{\beta}$ with the method of P-splines. They set $\boldsymbol{\beta} = \boldsymbol{B}_a \boldsymbol{b}$, where $\boldsymbol{B}_a$ is a regression matrix evaluated on a basis of B-splines along age. The model matrix for GLM1 in (66) becomes

$$\text{GLM1:} \quad \log \boldsymbol{\mu} = \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}_1 \boldsymbol{b}, \ \boldsymbol{X}_1 = \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{B}_a$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

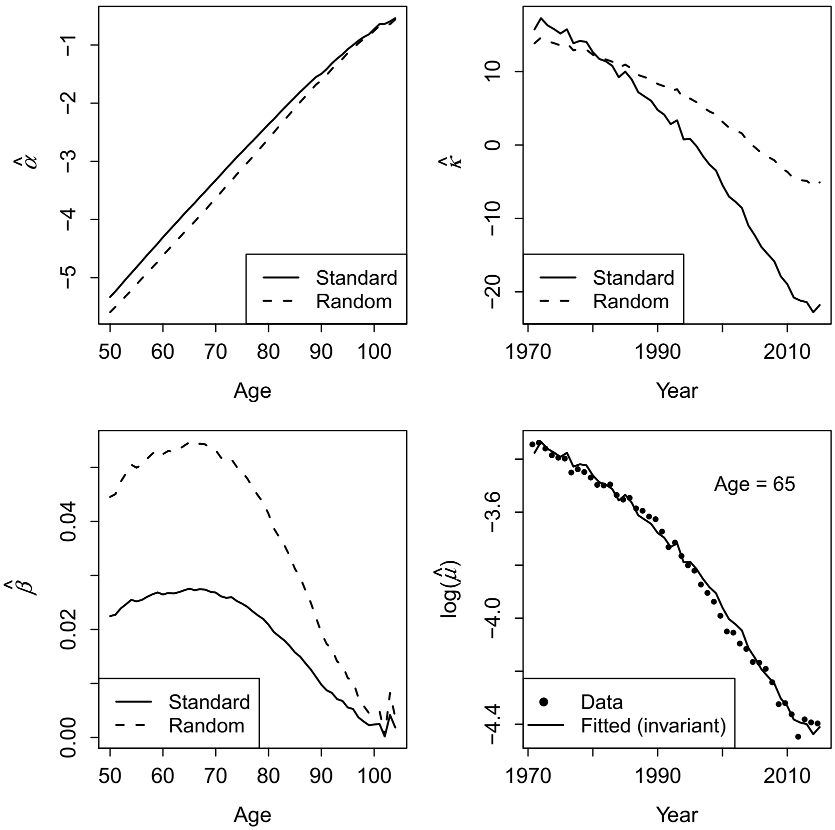

Figure 8. Parameter estimates in the LC model (60) under standard and random constraints: $\hat{\boldsymbol{\alpha}}_s = \hat{\boldsymbol{\alpha}}_r + A\hat{\boldsymbol{\beta}}_r$, $\hat{\boldsymbol{\beta}}_s = \hat{\boldsymbol{\beta}}_r/B$, $\hat{\boldsymbol{\kappa}}_s = B\hat{\boldsymbol{\kappa}}_r - AB\mathbf{1}_{n_y}$; $A = 5.907$, $B = 1.983$ and $AB = 11.71$. Fitted invariant log μ for age 65.

However, $[\tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{B}_a]\boldsymbol{b} = [\tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}]\boldsymbol{\beta}$, i.e., we are back in the model definition (66). The model matrix for GLM2 is not affected by the smoothing of $\boldsymbol{\beta}$. We deduce that (63), (64) and (65) hold in the smooth case too. We conclude that invariance of fitted and forecast values holds for the smooth model. Numerical work supports this conclusion.

4.5 Other models

We have described our method by giving a detailed discussion of a number of examples. The method can be applied to other mortality models and we mention two briefly.

4.5.1 CBD model with cohort effects

Cairns *et al.* (2006) introduced the model $\log \mu_{x,y} = \kappa_y^{(1)} + (x - \bar{x})\kappa_y^{(2)}$ which is often referred to as the CBD model. Cairns *et al.* (2009) modified this model with the addition of cohort effects:

$$\log \mu_{x,y} = \kappa_y^{(1)} + \kappa_y^{(2)}(x - \bar{x}) + \gamma_{c(x,y)},\ x = 1, \ldots, n_a,\ y = 1, \ldots, n_y \tag{68}$$

where $\bar{x}$ is the mean age. The model matrix is

$$\boldsymbol{X} = [\boldsymbol{I}_{n_y} \otimes \mathbf{1}_{n_a} : \boldsymbol{I}_{n_y} \otimes (\boldsymbol{x}_a - \bar{x}\mathbf{1}_{n_a}) : \boldsymbol{X}_c]. \tag{69}$$

The coefficient vector is $\boldsymbol{\theta} = (\boldsymbol{\kappa}^{(1)\prime}, \boldsymbol{\kappa}^{(2)\prime}, \boldsymbol{\gamma}')'$ with length $n_a + 3n_y - 1$. The rank of $\boldsymbol{X}$ is $n_a + 3n_y - 3$, so two constraints are required to give unique parameter estimates. We use $\sum \gamma_c =$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

$\sum c\gamma_c = 0$ as our standard constraints (Cairns *et al.*, 2009) and two sets of random constraints on $\boldsymbol{\theta}$. We compute $\mathcal{N}(\boldsymbol{X})$, the null space of $\boldsymbol{X}$, and conclude that

$$\begin{aligned}\boldsymbol{\kappa}_r^{(1)} &= \boldsymbol{\kappa}_s^{(1)} + [A + (\bar{x} - n_x)B]\mathbf{1}_{n_y} - B\boldsymbol{x}_y \\ \boldsymbol{\kappa}_r^{(2)} &= \boldsymbol{\kappa}_s^{(2)} + B\mathbf{1}_{n_y} \\ \boldsymbol{\gamma}_r &= \boldsymbol{\gamma}_s - A\mathbf{1}_c + B\boldsymbol{x}_c\end{aligned} \tag{70}$$

where we have used our usual notation for the estimates under the standard and random constraints. Once again we see that the parameters are linearly related, so provided we choose a forecasting method which preserves these relationships we can conclude that the forecasts of mortality are invariant with respect to the choice of constraints.

4.5.2 Reduced Plat model

Plat (2009) proposed the model

$$\log \mu_{x,y} = \alpha_x + \kappa_y^{(1)} + (x - \bar{x})\kappa_y^{(2)} + \gamma_{c(x,y)},\ x = 1, \ldots, n_a,\ y = 1, \ldots, n_y \tag{71}$$

Hunt & Blake (2020b) refer to this model as the reduced Plat model. We see that it is formally equivalent to the unsmoothed APCI model (44) so the dimension of its null space, $\mathcal{N}(\boldsymbol{X})$, is five, as is readily verified directly. In particular, $\mathcal{N}(\boldsymbol{X})$ has a quadratic term, as in (46). This means that an ARIMA(p,1,q) model will not lead to invariant forecasts. We can smooth the age term with a second-order penalty; this will reduce the dimension of the null space from five to four, just as in APCI model, removing the quadratic term from $\mathcal{N}(\boldsymbol{X})$. Once more, we are back in the linear case. We will comment further on this increase in the effective rank of the model in our concluding remarks.

5. Conclusions

There are three new ideas in this paper.

(a) The standard approach to modelling and forecasting of mortality specifies a *particular set* of constraints that enables parameters to be estimated uniquely; often there are arguments to support this choice. Our approach is to consider the *difference* in the parameter estimates obtained with two different constraint systems. These differences are characterised by the null space of the model matrix. This approach enabled us to deduce that, while parameters are not identifiable, (i) fitted values of mortality are identifiable (a known result) and (ii) forecast values of mortality too are identifiable for ARIMA(p, δ, q) models with a fitted mean where $\delta = 1$ or 2.

(b) Our second contribution is the idea of the *effective rank* of a model. We could describe a constraint like $\sum \kappa_y = 0$ in the AP model, the APC model or the APCI model as an explicit or *hard constraint*. The P-spline system constrains the regression coefficients by forcing a certain level of smoothness on them; we could call this an implicit or *soft constraint*. We saw in our discussion of the APC model and particularly of the APCI model that in some cases this soft constraint can increase the effective rank of the model, i.e., smoothing can act like a hard constraint.

(c) Our third idea is the use of random constraints. In this paper, we have used random constraints to illustrate our theoretical discussion. Random constraints also have an important practical use since their use avoids the need to calculate a null space for the model. The fitting algorithm (15) makes it straightforward to fit with any constraint system. A practical method of checking whether forecasting is invariant with respect to the choice of constraints is to forecast with both a given and a random constraint system. If the forecasts are equal it is reasonable to assume that forecasting is invariant with respect to the choice of constraints.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

The results in this paper are given for the principal case of interest to actuaries, namely, when we have a GLM for the force of mortality, $\mu_{x,y}$, and deaths are distributed according to the Poisson distribution. However, the results depend only on the assumption of a GLM and the structure of the model matrix $\boldsymbol{X}$. The actuary may choose to model $q_{x,y}$, the probability of death at age x in year y. We assume a binomial distribution, $D_{x,y} \sim \mathcal{B}(E_{x,y}, q_{x,y})$, where $E_{x,y}$ is the initial exposed to risk. The results on the invariance of fitted and forecast values of $q_{x,y}$ all go through without alteration. One could even make the simpler assumption that $\log (D_{x,y}/e_{x,y}) \sim \mathcal{N}(\log \mu_{x,y}, \sigma^2)$; our results apply here too.

The CMI gave a particular set of transformations of the parameters in the APCI model which left the values of $\log \boldsymbol{\mu}$ unchanged (CMI, 2016a, 7.3). We saw in our discussion of this model that this set of transformations corresponded to a particular basis for the null space of the model matrix. This is a good approach if a suitable set of transformations can be found. When smoothing is applied, it does not seem obvious that such a set of transformations can easily be found. In this case, we have provided a systematic method for computing a basis for the null space.

In Appendix B, we gave a simple method for fitting a generalised linear model with specified constraints which exploited the invariance of fitted values to the choice of constraints. The result is given when the parameters are not subject to smoothing. It is possible to extend this approach when there is smoothing of some model terms and we will return to this topic in future work.

This paper has focussed on invariance. This is an attractive mathematical property, but the actuary may have reasons for choosing a forecasting model which does not lead to invariant forecasts; this is particularly true for the forecasting of the cohort parameters. In the APC model, for example, estimates of the cohort parameters under different constraint systems differ by a linear function. We have argued that the choice of forecasting method should preserve this relationship. This places a restriction on the class of forecasting models available, namely to ARIMA models with a fitted mean. The actuary should be aware that a choice outside this class may not lead to invariant forecasts.

Our discussion of effective rank in (b) above raises an interesting and not easily resolved question. Smoothing the age parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with a second-order penalty in the APCI model (and the age parameters $\boldsymbol{\alpha}$ in the reduced Plat model) leads to a reduction in the number of constraints required to produce unique estimates of the parameters from five in the unsmoothed model to four. If a third-order penalty is used, the number of constraints required is again five. Some readers may worry that the number of constraints should depend on the details of the smoothing method adopted. Certainly, smoothing in the APCI and reduced Plat model can have a large effect on parameter estimates, a consequence of the change in the effective rank of the model. In contrast, smoothing has very little effect on the invariant quantities, the fitted and forecast values of mortality.

The R language has its own way of dealing with models which are not of full rank: it deletes columns of the $\boldsymbol{X}$ matrix until the remaining columns give a model matrix which is of full rank. For example, in the APC model, the columns corresponding to the most recent year and the two youngest cohorts are deleted; in the list of the estimated coefficients, these three parameters are reported as `NA`, i.e., not available. Explicit constraints are not used, and R's reporting emphasises the non-identifiability of the parameters without further assumption. A particular set of constraints corresponds to such an assumption; this is a strong assumption and one not easily verified. Of course, there are implicit constraints here and the reported parameters correspond to the explicit constraints $\kappa_{n_y} = \gamma_{n_c-1} = \gamma_{n_c} = 0$.

The practising actuary uses many mortality models in their daily work. Nearly all of these models require identifiability constraints. Our results tell us that we can be relaxed about the choice of constraint system. Forecast values will not depend on this choice provided we stay within the class of models used in this paper. This is a most reassuring conclusion.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

Acknowledgements. I am grateful to Torsten Kleinow for suggesting this problem, to Stephen Richards for many helpful comments on earlier versions of this paper, and to Jim Howie for a most helpful tutorial on matrices and their null spaces. I am also grateful to the anonymous referee for a most helpful and thorough report which led to many clarifications and improvements to the presentation.

References

Booth, H. & **Tickle, L.** (2008). Mortality modelling and forecasting: a review of methods. *Annals of Actuarial Science*, **3**, 3–44.

Brouhns, N., **Denuit, M.** & **Vermunt, J.K.** (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**, 373–393.

Cairns, A.J.G., **Blake, D.** & **Dowd, K.** (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, **73**, 687–718.

Cairns, A.J.G., **Blake, D.**, **Dowd, K.**, **Coughlan, G.D.**, **Epstein, D.**, **Ong, A.** & **Balevich, I.** (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**, 1–35.

Clayton, D. & **Schifflers, E.** (1987a). Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine*, **6**, 449–467.

Clayton, D. & **Schifflers, E.** (1987b). Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine*, **6**, 469–481.

Continuous Mortality Investigation (2016a). CMI Mortality Projections Model consultation. Working Paper 90.

Continuous Mortality Investigation (2016b). CMI Mortality Projections Model consultation - technical paper. Working Paper 91.

Continuous Mortality Investigation (2016c). CMI Mortality Projections Model: Consultation, responses and plans for CMI_2016. Working Paper 93.

Currie, I.D., **Durban, M.** & **Eilers, P.H.C.** (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**, 279–298.

Currie, I.D. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling*, **13**, 69–93.

Currie, I.D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, **2016**, 356–383.

Delwarde, A., **Denuit, M.** & **Eilers, P.** (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized likelihood approach. *Statistical Modelling*, **7**, 29–48.

Eilers, P.H.C. & **Marx, B.D.** (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, **11**, 89–121.

Hunt, A. & **Blake, D.** (2020a). Identifiability in age/period mortality models. *Annals of Actuarial Science*, **14**, 461–499.

Hunt, A. & **Blake, D.** (2020b). Identifiability in age/period/cohort mortality models. *Annals of Actuarial Science*, **14**, 500–536.

Hunt, A. & **Blake, D.** (2020c). A Bayesian approach to modelling and forecasting cohort effects. *North American Actuarial Journal.*

Lee, R.D. & **Carter, L.R.** (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, 659–675.

Macdonald, A.S., **Richards, S.J.** & **Currie, I.D.** (2018). *Modelling Mortality with Actuarial Applications*. Cambridge University Press, Cambridge.

Nelder, J.A. & **Wedderburn, R.W.M.** (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, **135**, 370–384.

Plat, J. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, **45**, 393–404.

R Core Team (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Richards, S.J., **Currie, I.D.**, **Kleinow, T.** & **Ritchie, G.P.** (to appear). A stochastic implementation of the APCI model for mortality projections. *British Actuarial Journal.*

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* Wiley, New York.

Shumway, R.H. & **Stoffer, D.S.** (2017). *Time Series Analysis and Its Applications: With R Examples.* Springer, Berlin.

Whittaker, E.T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

A. Appendix

A.1 Some matrix results on invariance

We consider a model with model matrix $\boldsymbol{X}$ and regression coefficients $\boldsymbol{\theta}$. The number of constraints on $\boldsymbol{\theta}$ required to give a unique estimate of $\boldsymbol{\theta}$ is determined by the rank of $\boldsymbol{X}$ which we

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

denote by $r(\boldsymbol{X})$. The definition of $r(\boldsymbol{X})$ depends on a fundamental result in matrix theory. We need two definitions: the row rank of $\boldsymbol{X}$ is the maximum number of linearly independent rows of $\boldsymbol{X}$; similarly, the column rank of $\boldsymbol{X}$ is the maximum number of linearly independent columns of $\boldsymbol{X}$. Then, it can be proved that the row rank of $\boldsymbol{X}$ equals its column rank; this common value is known as the rank of $\boldsymbol{X}$. For a proof of this result see, for example, Searle (1982, chapter 6).

The relationship between different estimates of $\boldsymbol{\theta}$ under different constraint systems is determined by the null space of $\boldsymbol{X}$. We denote the null space of $\boldsymbol{X}$ by $\mathcal{N}(\boldsymbol{X})$ and define

$$\mathcal{N}(\boldsymbol{X}) = \{\boldsymbol{v} : \boldsymbol{X}\boldsymbol{v} = \boldsymbol{0}\}$$

For our purposes, the following result is fundamental.

Proposition A.1. *For any matrix* $\boldsymbol{X}$

$$\mathcal{N}(\boldsymbol{X}'\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}) \tag{A.1}$$

Proof. First, let $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{X}'\boldsymbol{X})$. Then

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{v}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} = 0 \Rightarrow (\boldsymbol{X}\boldsymbol{v})'(\boldsymbol{X}\boldsymbol{v}) = 0 \Rightarrow \boldsymbol{X}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{v} \in \mathcal{N}(\boldsymbol{X})$$

Conversely, let $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{X})$, then $\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{X}'\boldsymbol{X}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{v} \in \mathcal{N}(\boldsymbol{X}'\boldsymbol{X})$.

Hence, $\mathcal{N}(\boldsymbol{X}'\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X})$. □

We wish to show that the fitted values in a GLM are invariant with respect to the choice of constraints. We consider first the standard linear regression model with normal errors and common variance; the result for a GLM will follow.

Proposition A.2. *Define the regression model*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon},\ \boldsymbol{X},\ n \times p, n > p, r(\boldsymbol{X}) = p - q, q \geq 1 \tag{A.2}$$

Let $\hat{\boldsymbol{\theta}}_1$ *and* $\hat{\boldsymbol{\theta}}_2$ *be any two solutions of the normal equations*

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{y} \tag{A.3}$$

Then,

$$\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2 \tag{A.4}$$

Proof: Since $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ satisfy the normal equations (A.3)

$$\begin{aligned} &\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0} \\ \Rightarrow\ &\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \in \mathcal{N}(\boldsymbol{X}'\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}) \quad \text{by (A.1)} \\ \Rightarrow\ &\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2 \end{aligned}$$

as required. □

A square matrix $\boldsymbol{A}$ is *positive semi-definite* if $\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v}$. We first prove:

Proposition A.3. *Let* $\boldsymbol{A}$ *be a symmetric, positive semi-definite matrix. If* $\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} = 0$, *then* $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{A})$.

Proof. Since $\boldsymbol{A}$ is a symmetric, positive semi-definite matrix, there exists a matrix $\boldsymbol{K}$ such that $\boldsymbol{A} = \boldsymbol{K}'\boldsymbol{K}$; see Searle (1982, chapter 7) for example. Hence,

$$\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} = 0 \Rightarrow \boldsymbol{v}'\boldsymbol{K}'\boldsymbol{K}\boldsymbol{v} = 0 \Rightarrow (\boldsymbol{K}\boldsymbol{v})'(\boldsymbol{K}\boldsymbol{v}) = 0 \Rightarrow \boldsymbol{K}\boldsymbol{v} = \boldsymbol{0} \Rightarrow \boldsymbol{K}'\boldsymbol{K}\boldsymbol{v} = \boldsymbol{0}$$

and $\boldsymbol{v} \in \mathcal{N}(\boldsymbol{A})$. □

We need a definition: the *dimension* of the null space of a matrix $\boldsymbol{X}$ is the number of vectors in its basis; we denote this by $\dim[\mathcal{N}(\boldsymbol{X})]$. In general, we have

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

Proposition A.4. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be symmetric, positive semi-definite matrices of the same size. Then*
(a) $\mathcal{N}(\boldsymbol{A}+\boldsymbol{B})=\mathcal{N}(\boldsymbol{A})\cap\mathcal{N}(\boldsymbol{B})$, and
(b) $r(\boldsymbol{A}+\boldsymbol{B})\geq max\{r(\boldsymbol{A}), r(\boldsymbol{B})\}$.

Proof. Let $\boldsymbol{v}\in\mathcal{N}(\boldsymbol{A}+\boldsymbol{B})$. Then $(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v}=\boldsymbol{0}\Rightarrow \boldsymbol{v}'(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v}=0\Rightarrow \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}+\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v}=0\Rightarrow \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}=0$ and $\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v}=0$ since both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semi-definite. Hence, by Proposition A.3, $\boldsymbol{v}\in\mathcal{N}(\boldsymbol{A})$ and $\boldsymbol{v}\in\mathcal{N}(\boldsymbol{B})\Rightarrow \boldsymbol{v}\in\mathcal{N}(\boldsymbol{A})\cap\ \mathcal{N}(\boldsymbol{B})$.

Clearly, if $\boldsymbol{v}\in\mathcal{N}(\boldsymbol{A})\cap\mathcal{N}(\boldsymbol{B})$, then $\boldsymbol{v}\in\mathcal{N}(\boldsymbol{A}+\boldsymbol{B})$ and (a) is proved.

Furthermore, it follows from (a) that

$$\dim[\mathcal{N}(\boldsymbol{A}+\boldsymbol{B})]\leq\min\{\dim[\mathcal{N}(\boldsymbol{A})],\dim[\mathcal{N}(\boldsymbol{B})]\}$$

and (b) follows immediately from the rank-nullity theorem; see Searle, (1982, chapter 6) for example. □

In our case, we take $\boldsymbol{A}=\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{B}=\boldsymbol{P}$, where $\boldsymbol{P}$ is a penalty matrix. The result tells us that smoothing may reduce the number of constraints required to obtain a unique estimate of $\boldsymbol{\theta}$.

An immediate corollary of Proposition A.4 is the invariance result for fitted values in a penalised regression.

Corollary: *Let $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ be any two solutions of the penalised normal equations*

$$(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{P})\hat{\boldsymbol{\theta}}=\boldsymbol{X}'\boldsymbol{y}$$

Then

$$\boldsymbol{X}\hat{\boldsymbol{\theta}}_1=\boldsymbol{X}\hat{\boldsymbol{\theta}}_2$$

Proof. By assumption, we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_2 &\in \mathcal{N}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{P})\\ &=\mathcal{N}(\boldsymbol{X}'\boldsymbol{X})\cap\mathcal{N}(\boldsymbol{P}) \text{ by Proposition A.4}\\ &\subseteq\mathcal{N}(\boldsymbol{X}'\boldsymbol{X})\\ &=\mathcal{N}(\boldsymbol{X}) \text{ by Proposition A.1}\end{aligned}$$

and $\boldsymbol{X}\hat{\boldsymbol{\theta}}_1=\boldsymbol{X}\hat{\boldsymbol{\theta}}_2$ as required. □

The extension of these results to a GLM is straightforward. The normal equations are replaced by the estimating equations (Nelder & Wedderburn, 1972)

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\theta}}=\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z} \tag{A.5}$$

where $\boldsymbol{W}$ is the diagonal matrix of weights and $\boldsymbol{z}$ is the working variable; $\hat{\boldsymbol{\theta}}$ is the updated estimate of $\boldsymbol{\theta}$. Both $\boldsymbol{W}$ and $\boldsymbol{z}$ depend on the current estimated value of the mean. Suppose that $\boldsymbol{W}_0$ and $\boldsymbol{z}_0$ are the initial estimates of $\boldsymbol{W}$ and $\boldsymbol{z}$ which are computed from the initial estimate of the mean (usually the observed values y). Let $\hat{\boldsymbol{\theta}}_{0,1}$ and $\hat{\boldsymbol{\theta}}_{0,2}$ be any two initial solutions of (A.5). We write (A.5) as

$$\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\hat{\boldsymbol{\theta}}=\boldsymbol{X}^{*\prime}\boldsymbol{z}^*$$

where $\boldsymbol{X}^*=\boldsymbol{W}^{1/2}\boldsymbol{X}$, $\boldsymbol{z}^*=\boldsymbol{W}^{1/2}\boldsymbol{z}$ and $\boldsymbol{W}^{1/2}$ is the diagonal matrix whose elements are the square roots of those of $\boldsymbol{W}$. (We note that $\boldsymbol{W}^{1/2}$ exists since $\boldsymbol{W}$ is diagonal with positive diagonal entries.) Then by the invariance result (A.4), we have

$$\boldsymbol{X}^*\hat{\boldsymbol{\theta}}_{0,1}=\boldsymbol{X}^*\hat{\boldsymbol{\theta}}_{0,2}\Rightarrow\boldsymbol{X}\hat{\boldsymbol{\theta}}_{0,1}=\boldsymbol{X}\hat{\boldsymbol{\theta}}_{0,2}$$

Hence, the updated values of the mean are equal and so the updated values $\boldsymbol{W}_1$ and $\boldsymbol{z}_1$ are also equal. Hence, as we iterate (A.5) until convergence, we have $\boldsymbol{X}\hat{\boldsymbol{\theta}}_{i,1}=\boldsymbol{X}\hat{\boldsymbol{\theta}}_{i,2}$ at the i th iteration.

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

Hence, at convergence $\boldsymbol{X}\hat{\boldsymbol{\theta}}_1 = \boldsymbol{X}\hat{\boldsymbol{\theta}}_2$ where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are the converged values. Thus, $\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \in \mathcal{N}(\boldsymbol{X})$ just as in the normal regression case. The extension to smoothing in a GLM follows as in the normal case.

B. Appendix

B.1 Some R code on invariance

The fundamental invariance result (A.4) implies that we can compute a model fit subject to a particular set of constraints from any other fit of the same model under different constraints. Currie (2016, equation (31)) gives the following formula:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{X}' \log \hat{\boldsymbol{\mu}} \tag{B.1}$$

Here, $\boldsymbol{X}$ is the model matrix for, say, the APC model, $\log \hat{\boldsymbol{\mu}}$ is the (invariant) vector of fitted log mortalities and $\boldsymbol{\theta}$ is the parameter vector which is subject to the desired constraint $\boldsymbol{H}\boldsymbol{\theta} = \boldsymbol{0}$. In the case of the APC model, $\boldsymbol{H}$ is $3 \times (n_a + n_y + n_c)$.

For example, let Dth, Exp, X and H be the R objects that contain the vectors of deaths and exposures, the model matrix and the constraints matrix, respectively. Then, the following computes $\log \hat{\boldsymbol{\mu}}$ with R's glm function and then computes $\hat{\boldsymbol{\theta}}$ subject to $\boldsymbol{H}\boldsymbol{\theta} = \boldsymbol{0}$ from (B.1).

```
Fit.glm <- glm(Dth ~ X + offset(log(Exp)), family = poisson)
Log.Mu.hat <- log(Fit.glm$fit/Exp)
Theta.hat <- solve(t(X) %*% X + t(H) %*% H, t(X) %*% Log.Mu.hat)
```

We note in particular that we do not need to know what constraints the glm function has used to fit the model.

C. Appendix

C.1 A time series result

We show that if two time series are linearly related and they are forecast with an ARIMA(p, δ, q) model with a fitted mean, $\delta = 1$ or 2, then the forecasts obey the same linear relationship. First, we prove a preliminary result. The proof uses discrete integration where the usual constant of integration corresponds to how the forecast is joined to the series to be forecast. The results correspond to how R computes a forecast.

Proposition A.5. *Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$. There are two cases.*

Case 1: $\delta = 1$. *Consider forecasting $\boldsymbol{y}$ with an ARIMA$(p, 1, q)$ model. Let $\boldsymbol{u} = (u_1, \ldots, u_{n_f})'$ be the forecast of length n_f of first differences of $\boldsymbol{y}$ with an ARMA(p, q) model. Let $\boldsymbol{y}_f = (y_{n+1}, \ldots, y_{n+n_f})'$ be the forecast of length n_f* of *$\boldsymbol{y}$ with the ARIMA$(p, 1, q)$ model. Then*

$$y_{n+j} = y_n + \sum_{k=1}^{j} u_k,\ j = 1, \ldots, n_f \tag{C.1}$$

Case 2: $\delta = 2$. *In a similar fashion, let $\boldsymbol{v} = (v_1, \ldots, v_{n_f})'$ be the forecast of length n_f of second differences of $\boldsymbol{y}$ with an ARMA(p, q) model. Now let $\boldsymbol{y}_f = (y_{n+1}, \ldots, y_{n+n_f})'$ be the forecast of length n_f* of *$\boldsymbol{y}$ with the ARIMA$(p, 2, q)$ model. Then*

$$y_{n+j} = y_n + \sum_{k=1}^{j} w_k,\ j = 1, \ldots, n_f \tag{C.2}$$

where

$$w_k = y_n - y_{n-1} + \sum_{\ell=1}^{k} v_\ell, \; k = 1, \ldots, n_f \tag{C.3}$$

Proof. Case 1. Forecasting $\boldsymbol{y}$ with an ARIMA(p, 1, q) is performed by fitting an ARMA(p, q) model to first differences of $\boldsymbol{y}$ and then reversing the differencing. With the above notation, we have

$$\begin{aligned} y_{n+1} &= y_n + u_1 \\ y_{n+2} &= y_{n+1} + u_2 = y_n + u_1 + u_2, \text{ and in general} \\ y_{n+j} &= y_{n+j-1} + u_j = y_n + \sum_{k=1}^{j} u_k \end{aligned}$$

which is (C.1).

Case 2. Forecasting $\boldsymbol{y}$ with an ARIMA(p, 2, q) is performed by fitting an ARMA(p, q) model to second differences of $\boldsymbol{y}$ and then reversing the differencing. Applying (C.1) once, we find

$$y_{n+j} = y_n + \sum_{k=1}^{j} w_k$$

where $\boldsymbol{w} = (w_1, \ldots, w_{n_f})'$ is the forecast of first differences of $\boldsymbol{y}$. Applying (C.1) a second time to the series $\boldsymbol{w}$ yields

$$w_k = y_n - y_{n-1} + \sum_{\ell=1}^{k} v_\ell$$

as required. □

We can now prove

Proposition A.6. *Let* $\boldsymbol{y} = (y_1, \ldots, y_n)'$ *and* $\boldsymbol{z} = (z_1, \ldots, z_n)'$ *be two times series which are linearly related, i.e.,*

$$\boldsymbol{y} = \boldsymbol{z} + a\mathbf{1} + b\boldsymbol{x} \tag{C.4}$$

for some a and b; here $\mathbf{1}$ *is the vector of 1s of length n and* $\boldsymbol{x} = (1, \ldots, n)'$. *Let* $\boldsymbol{y}_f = (y_{n+1}, \ldots, y_{n+n_f})'$ *and* $\boldsymbol{z}_f = (z_{n+1}, \ldots, z_{n+n_f})'$ *be the forecasts of* $\boldsymbol{y}$ *and* $\boldsymbol{z}$ *of length* n_f *with an* ARIMA(p, δ, q), $\delta = 1, 2$ *model. Then,* $\boldsymbol{y}_f$ *and* $\boldsymbol{z}_f$ *obey the same linear relationship as* $\boldsymbol{y}$ *and* $\boldsymbol{z}$, *i.e.,*

$$\boldsymbol{y}_f = \boldsymbol{z}_f + a\mathbf{1} + b(n\mathbf{1} + \boldsymbol{x}_f) \tag{C.5}$$

where $\boldsymbol{x}_f = (1, \ldots, n_f)'$.

Proof: Case 1: $\delta = 1$. We have from (C.4)

$$\Delta(\boldsymbol{y}) - b\mathbf{1}_{n-1} = \Delta(\boldsymbol{z}) \tag{C.6}$$

where Δ indicates first-order differencing and $\mathbf{1}_{n-1}$ is the vector of 1s of length $n-1$. Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be the forecasts, respectively, of $\Delta(\boldsymbol{y})$ and $\Delta(\boldsymbol{z})$ with an ARIMA(p, 0, q) model with a mean. Denote the fitted means by μ_u and μ_v. It follows immediately from (C.6) and the definition of an ARIMA model with a mean that $\mu_u = \mu_v + b$. Hence

$$\boldsymbol{u} = \boldsymbol{v} + b\mathbf{1}_{n-1} \tag{C.7}$$

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press

We have

$$\begin{aligned} y_{n+j} - z_{n+j} &= y_n + \sum_{k=1}^{j} u_k - z_n - \sum_{k=1}^{j} v_k \text{ by (C.1)} \\ &= y_n + \sum_{k=1}^{j} u_k - y_n + a + nb - \sum_{k=1}^{j} (u_k - b) \text{ by (C.4) and (C.7)} \\ &= a + (n+j)b \end{aligned}$$

as required.

Case 2: $\delta = 2$. We have

$$\Delta^2(\boldsymbol{y}) - \Delta^2(\boldsymbol{z}) = \boldsymbol{0}_{n-2} \tag{C.8}$$

and so $\boldsymbol{u} = \boldsymbol{v}$ where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the forecasts of $\Delta^2(\boldsymbol{y})$ and $\Delta^2(\boldsymbol{z})$ with an ARIMA($(p, 0, q)$ model, respectively. We apply (C.2) and (C.3). We have

$$\begin{aligned} y_{n+j} - z_{n+j} &= y_n + \sum_{k=1}^{j} \left(y_n - y_{n-1} + \sum_{\ell=1}^{k} u_\ell \right) - z_n - \sum_{k=1}^{j} \left(z_n - z_{n-1} + \sum_{\ell=1}^{k} v_\ell \right) \\ &= y_n + \sum_{k=1}^{j} (y_n - y_{n-1}) - y_n + a + nb - \sum_{k=1}^{j} (y_n - a - nb - y_{n-1} + a + (n-1)b) \\ &= \sum_{k=1}^{j} (y_n - y_{n-1}) + a + nb - \sum_{k=1}^{j} (y_n - y_{n-1} - b) \\ &= a + (n+j)b \end{aligned}$$

as required. □

Cite this article: Currie ID (2020). Constraints, the identifiability problem and the forecasting of mortality. *Annals of Actuarial Science* **14**, 537–566. https://doi.org/10.1017/S1748499520000020

https://doi.org/10.1017/S1748499520000020 Published online by Cambridge University Press