Smoothing constrained generalized linear models with an application to the Lee-Carter model

Running headline: Smoothing constrained GLMs

Iain D Currie
Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK.
Email: I.D.Currie@hw.ac.uk, Tel: +44 (0)131 451 3208, Fax: +44 (0)131 451 3249

Abstract: We consider a generalized linear model (GLM) with canonical link function in which parameters can be subject to (a) a set of linear constraints and (b) smoothing. We apply Lagrange methods to give a general Newton-Raphson algorithm for such a GLM in which parameters are estimated, constraints are applied and smoothing is performed simultaneously. We express the Lee-Carter model, an important model for the forecasting of human mortality, in terms of GLMs, and use our method to estimate the parameters in the model. The smoothing option allows us to improve the forecasting properties of the model. We compare the performance of (a) the Poisson model with log link for the force of mortality and (b) the binomial model with logit link for the probability of death in a calendar year. Examples using UK Office for National Statistics data are provided.

Key words: Constraints, generalized linear models, identifiability, Lagrange methods, Lee-Carter, mortality, Newton-Raphson, smoothing.

1 Introduction

The improvement in human mortality over most of the 20th century and into the 21st has been truly remarkable and the increase in the life expectancy of the elderly is particularly striking. For example, the life expectancy of a 65-year-old male in England has risen from 13.1 years in 1980-82 to 18.0 years in 2008-10 (ONS, 2011). These increases are all the more striking since they are *period life expectancies,* ie, they are calculated on the assumption that a life experiences the age-specific mortality of the day for the rest of his life, and thus take no account of any future mortality improvements.

These improvements in mortality have major implications for the funding of public and private pensions, the care of the elderly and the provision of health services. The forecasting of future mortality is thus of crucial importance to the providers of such services. In this paper we will concentrate on the forecasting of mortality in the pensions area, although our methods will have wider application.

Lee and Carter (1992) introduced an important method for the forecasting of mortality in general and life expectancy in particular. Unfortunately, the forecasts produced by their model are not ideal for the fair pricing of pension products since they result in "irregular projected life tables" (Delwarde et al., 2007). This drawback was recognised by Delwarde et al. (2007) who used smoothing to produce more regular forecasts. Their solution addressed most of the irregularity in the forecasts, and the immediate practical purpose of our paper is to show how to extend their solution so that fully regular forecasts are obtained.

Estimation in the Lee-Carter model has to confront two problems: first, the model is not a regression model and second, the model is not identifiable. These are not major issues but do mean that the application of regression methods is not straightforward. The Lee-Carter parameters can be divided into three sets: $\boldsymbol{\alpha}$, a general measure of mortality by age, $\boldsymbol{\kappa}$, a time trend which summarizes the pattern of mortality over time, and $\boldsymbol{\beta}$, an age-dependant factor which modulates the time trend $\boldsymbol{\kappa}$. Brouhns et al. (2002) used maximum likelihood to estimate the parameters in the model. Their approach is simple and direct: condition on some current estimates of two of the sets of parameters and apply the Newton-Raphson method to update the estimate of the third set. Now cycle through the parameter sets. The identifiability constraints are applied after the Newton-Raphson cycle is complete. Delwarde et al. (2007) applied this same method to a penalized log likelihood in which a penalty is used to smooth the elements of the modulation term $\boldsymbol{\beta}$.

The principal contribution of our paper is to show how the theory of generalized linear models (GLMs) can be extended so that both constraints and smoothing are incorporated simultaneously into the estimation algorithm. We express estimation in the Lee-Carter model in terms of GLMs and use our algorithm to give an efficient and unified approach to estimation in the original Lee-Carter model, the smooth model of Delwarde et al. and our own fully smoothed model; we will refer to these models as the LC, the DDE and the LC(S) models respectively.

The plan of the paper is as follows. In section 2 we describe our data set and introduce our approach by showing how some simple models of mortality with both constraints and

smoothing can be expressed in terms of GLMs. In section 3 we first show how the estimating equations in Brouhns et al. (2002) and Delwarde et al. (2007) are special cases of our own approach; we then show how estimation in the LC model can be expressed in terms of GLMs and apply this formulation to all three flavours of the LC model. The maximum likelihood method of Brouhns et al. (2002) was dependent on the assumption that the number of deaths at each age in each year followed the Poisson distribution. We make the same assumption throughout sections 2 and 3 but in section 4 we use an alternative formulation in terms of the binomial distribution. These two distributional assumptions are compared. In section 5 we draw some conclusions and indicate the direction of future work. The proof of our main result is placed in an appendix.

2 Simple GLMs with constraints and smoothing

We suppose that mortality data are available in two matrices, $\boldsymbol{D} = (d_{i,j})$ and $\boldsymbol{E} = (e_{i,j})$, $i = 1, \ldots, n_a$ and $j = 1, \ldots, n_y$; the rows of $\boldsymbol{D}$ and $\boldsymbol{E}$ are labelled by age of death, $\boldsymbol{x}_a' = (x_{a,1}, \ldots, x_{a,n_a})$, and the columns by year of death, $\boldsymbol{x}_y' = (x_{y,1}, \ldots, x_{y,n_y})$. Thus, $d_{i,j}$ is the number of deaths age $x_{a,i}$ in year $x_{y,j}$, and $e_{i,j}$ is the mid-year or central exposed to risk age $x_{a,i}$ in year $x_{y,j}$. Further, let $\boldsymbol{d} = \operatorname{vec} \boldsymbol{D}$ and $\boldsymbol{e} = \operatorname{vec} \boldsymbol{E}$ be the vectors of deaths and exposures corresponding to $\boldsymbol{D}$ and $\boldsymbol{E}$; here and below, the vec operator stacks the columns of a matrix on top of each other in column order. We will use data for ages 40 to 90 and years 1961 to 2009 from the UK Office for National Statistics (ONS) on male mortality in England and Wales to illustrate our methods. Thus our two data matrices have dimension $n_a \times n_y = 51 \times 49$.

We suppose that $d_{i,j}$ is a realization of a Poisson variable $D_{i,j} \sim \mathcal{P}(e_{i,j}\lambda_{i,j})$ where $\boldsymbol{\Lambda} = (\lambda_{i,j})$ and $\boldsymbol{\lambda} = \operatorname{vec} \boldsymbol{\Lambda}$ are the matrix and vector of forces of mortality or hazard rates respectively. We note that strictly we should write $\lambda_{i+0.5,j+0.5}$ since $\lambda_{i,j}$ represents the mid-year force of mortality; here, and below, we interpret $\lambda_{i+0.5,j+0.5}$ as the force of mortality at age $x_{a,i} + 0.5$ at time $x_{y,j} + 0.5$. Further let $\boldsymbol{\mu} = \boldsymbol{e} * \boldsymbol{\lambda}$ denote the vector of expected deaths; here and below, $*$ indicates element-by-element multiplication.

We introduce our method by considering some simple models in which we ignore the change in mortality over time. These models have little practical interest and we use them solely to show how a generalization of the familiar iterative weighted least squares (IWLS) algorithm for a GLM enables us to handle constraints and smoothing within the GLM framework.

2.1 Gompertz models as GLMs

Gompertz (1825) observed that the force of mortality was approximately linear in age (on the log scale) over most of adult life. We express the Gompertz model over the entire mortality table in the language of GLMs as follows

$$\begin{aligned} D_{i,j} &\sim \mathcal{P}(e_{i,j}\lambda_{i,j}),\ i = 1, \ldots, n_a,\ j = 1, \ldots, n_y \\ \log E(D_{i,j}) &= \log e_{i,j} + \alpha_0 + \alpha_1 x_{a,i},\ i = 1, \ldots, n_a,\ j = 1, \ldots, n_y \end{aligned}$$

or more compactly

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\boldsymbol{\theta},\ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes [\mathbf{1}_{n_a} : \boldsymbol{x}_a], \tag{1}$$

with $\mathbf{1}_n$ a vector of 1's of length n and $\boldsymbol{\theta} = (\alpha_0, \alpha_1)'$. Here $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the *Kronecker product* of $\boldsymbol{A}$ and $\boldsymbol{B}$; see Searle (1982) for further information on Kronecker products. We have defined a GLM with *linear predictor* $\boldsymbol{\eta} = \log \boldsymbol{e} + \boldsymbol{X}\boldsymbol{\theta}$, log *link* and Poisson error; the term $\log \boldsymbol{e}$ is an *offset* in the model.

In their landmark paper Nelder and Wedderburn (1972) introduced GLMs and showed that estimation in a GLM is given by

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}. \tag{2}$$

Here 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*. Thus (2) has the *iterative weighted least squares* (IWLS) form. A proof of the IWLS algorithm is given in a more general setting in the Theorem in the Appendix.

In the case of Poisson errors and log link $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{z}}$ are (McCullagh and Nelder, 1989)

$$\tilde{\boldsymbol{W}} = \mathrm{diag}\{\tilde{\boldsymbol{\mu}}\},\ \tilde{\boldsymbol{z}} = \boldsymbol{X}\tilde{\boldsymbol{\theta}} + \left(\frac{\boldsymbol{d}}{\tilde{\boldsymbol{\mu}}} - \mathbf{1}\right); \tag{3}$$

here and below, $\boldsymbol{d}/\tilde{\boldsymbol{\mu}}$ indicates element-by-element division. We fit the model with (2) and (3) where $\boldsymbol{X}$ is given by (1).

For the remainder of the paper we will assume without further comment that we model the force of mortality with a log link and a Poisson error. There is one important exception to this assumption: in section 4 we discuss models for $q_{i,j}$, the probability that a life with exact age $x_{a,i}$ at the start of year $x_{y,j}$ dies in the following year; here a logit link with a binomial error is more appropriate.

2.2 Factor models as GLMs

The factor model in age ignoring time has linear predictor

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\boldsymbol{\alpha},\ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a}, \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_a})'$, $\log \lambda_{i,j} = \alpha_i$, $i = 1, \ldots, n_a$, $j = 1, \ldots, n_y$, and $\boldsymbol{I}_n$ is the identity matrix of size n. We fit the model with (2), (3) and $\boldsymbol{X}$ as in (4). Of course, both the Gompertz model (1) and the factor model (4) would normally be fitted by a standard call in any of a number of statistical packages; for example, the `glm( )` function would be used in R (R Development Core Team, 2011). However, our purpose here is to set up a unified approach for a general class of GLMs with constraints and smoothing.

2.3 Smooth models as GLMs

The Gompertz model with its straight line form is at one end of the model spectrum; at the other end we have the factor model. The smooth model lies somewhere in between. We will use the P-spline system of Eilers and Marx (1996) for smoothing. Let $\mathcal{B}_a = \{B_{a,1}, \ldots, B_{a,c_a}\}$ be a cubic B-spline basis of dimension c_a defined on equally spaced knots and spanning age; let $\boldsymbol{B}_a = (b_{i,j}) = (B_{a,j}(x_{a,i}))$, $n_a \times c_a$, be the resulting regression matrix. The use of the basis $\mathcal{B}_a$ will result in some smoothing, with larger values of the dimension c_a resulting in rougher fits. We have a GLM with linear predictor

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\boldsymbol{a},\ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes \boldsymbol{B}_a \tag{5}$$

where $\boldsymbol{a}' = (a_1, \ldots, a_{c_a})$. The idea behind the P-spline method is to choose a large value of the dimension c_a and then penalize the resulting undersmoothing. Smoothness of the fitted curve is controlled by a penalty function which penalizes differences between adjacent coefficients. We will use the second order penalty function

$$\tau[(a_1 - 2a_2 + a_3)^2 + \ldots + (a_{c_a-2} - 2a_{c_a-1} + a_{c_a})^2] = \tau \boldsymbol{a}'\boldsymbol{D}_2'\boldsymbol{D}_2\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{P}\boldsymbol{a} \tag{6}$$

where

$$\boldsymbol{P} = \tau \boldsymbol{D}_2'\boldsymbol{D}_2 \tag{7}$$

is the *penalty matrix* and $\boldsymbol{D}_2$, $(c_a - 2) \times c_a$, is the second order difference matrix

$$\boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{8}$$

We emphasize that $\boldsymbol{D}_2$ with the suffix 2 denotes a second order difference matrix while $\boldsymbol{D}$ with no suffix denotes the matrix of observed deaths.

The parameter τ is a tuning constant which controls the amount of smoothing. There are two special cases of interest: (a) if $\tau = 0$ then there is no smoothing and we have a standard GLM with regression matrix $\boldsymbol{X}$ as in (5) and (b) with second order differences if $\tau \to \infty$ then in the limit we recover the Gompertz model. The tuning constant τ is generally known as the *smoothing parameter* (Wood, 2006). Estimation is via penalized likelihood, in which case, conditional on the value of τ, the GLM algorithm (2) extends to

$$(\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P})\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}. \tag{9}$$

Again, a proof is given in a more general setting in the Theorem in the Appendix. Currie et al. (2004) gave equation (9).

Conditional on the smoothing parameter τ we fit the smooth model with the regression matrix $\boldsymbol{X}$ given by (5), the penalty matrix $\boldsymbol{P} = \tau \boldsymbol{D}_2'\boldsymbol{D}_2$ and the coefficients $\boldsymbol{\theta} = \boldsymbol{a}$ in (9). There remains the problem of choosing an appropriate value of the smoothing parameter. For this we must step outside the likelihood framework. One possibility is to choose that

value of τ which optimizes some model selection criterion. In this paper we will minimize the Bayesian Information Criterion or BIC:

$$\text{BIC} = \text{Dev} + \log(n)\,\text{ED} \tag{10}$$

where Dev and ED are the deviance and the effective dimension of the fitted model respectively, and $n = n_a n_y$ is the number of observations; see McCullagh and Nelder (1989) for a discussion of deviance and Hastie and Tibshirani (1990) for one on effective dimension.

2.4 GLMs with constraints

We illustrate how we deal with GLMs with constraints by parameterizing the factor model (4) as follows:

$$\log \lambda_{i,j} = \alpha_0 + \psi_i,\ i = 1, \ldots, n_a,\ j = 1, \ldots, n_y. \tag{11}$$

Clearly, the parameters in (11) are not *identifiable*. We would like a simple test of whether a parameterization is identifiable or not; such a test is given in terms of the *rank* of the regression matrix for the model (11). Let $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\psi}')'$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{n_a})'$, be the vector of regression coefficients. Then the linear predictor for the model (11) is

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\,\boldsymbol{\theta},\ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes [\mathbf{1}_{n_a} : \boldsymbol{I}_{n_a}]. \tag{12}$$

Now $\boldsymbol{X}$ is $n_a n_y \times (n_a + 1)$ but rank n_a. Hence $\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}$, $(n_a + 1) \times (n_a + 1)$, is also rank n_a and so is singular; we cannot solve equation (2).

We can force a unique solution in (11) by placing a condition on the parameters. There is no unique way of doing this but some ways may have a natural appeal. Here we can specify that $\sum \psi_i = 0$ since this allows us to interpret the ψ_i as deviations from average mortality as measured by α_0. Let $\boldsymbol{h}' = (0, \mathbf{1}'_{n_a})$ be the vector corresponding to the constraint $\sum \psi_i = 0$, ie, $\boldsymbol{h}'\boldsymbol{\theta} = 0$ and let $\boldsymbol{H} = \boldsymbol{h}'$ be the *constraints matrix*. Let

$$\boldsymbol{X}_{aug} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{H} \end{pmatrix} \tag{13}$$

be the *augmented* regression matrix. The constraints in $\boldsymbol{H}$ will specify a unique solution in (11) if $\boldsymbol{X}_{aug}$ has full column rank. It is a simple matter to check that $\boldsymbol{X}_{aug}$, $(n_a n_y + 1) \times (n_a + 1)$, has rank $(n_a + 1)$. The unique solution for $\boldsymbol{\theta}$ in (12) subject to $\boldsymbol{H}\boldsymbol{\theta} = 0$ is given by

$$\begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} & : & \boldsymbol{H}' \\ \boldsymbol{H} & : & 0 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\omega} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \\ 0 \end{pmatrix} \tag{14}$$

where $\hat{\omega}$ is a Lagrange multiplier. This is equation (47) in the Theorem in the Appendix with $\boldsymbol{X}$ as in (12), $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{k} = 0$.

2.5 Smooth GLMs with constraints

In general we define three matrices: $\boldsymbol{X}$, the regression matrix, $\boldsymbol{P}$, the penalty matrix, and $\boldsymbol{H}$, the constraints matrix. We then apply the general result in our Theorem which gives the estimate of the regression coefficients as

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

When there is no smoothing, $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ subject to the set of constraints, $\boldsymbol{H}$, if any. When there is smoothing, $\hat{\boldsymbol{\theta}}$ is the maximum penalized likelihood estimate of $\boldsymbol{\theta}$ subject to the set of constraints, $\boldsymbol{H}$, if any.

We complete our introduction by smoothing the deviations $\boldsymbol{\psi}$ in the simple age model (11). We set $\boldsymbol{\psi} = \boldsymbol{B}_a\boldsymbol{a}$. The regression coefficients $\boldsymbol{\theta} = (\alpha_0, a_1, \ldots a_{c_a})'$ and the linear predictor is

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\,\boldsymbol{\theta},\ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes [\mathbf{1}_{n_a} : \boldsymbol{B}_a]. \tag{16}$$

We have exactly the same problem of non-identifiability as we encountered in section 2.4 and we solve it in the same way. The constraint $\sum \psi_i = 0$ is equivalent to $\boldsymbol{h}'\boldsymbol{\theta} = 0$ where $\boldsymbol{h}' = (0, \mathbf{1}'_{n_a}\boldsymbol{B}_a)$. The penalty matrix $\boldsymbol{P}$ comprises two parts: the zero penalty on α_0 and the second order penalty (6) on $\boldsymbol{a}$; this gives

$$\boldsymbol{P} = \text{blockdiag}\{0, \tau\boldsymbol{D}'_2\boldsymbol{D}_2\}. \tag{17}$$

We fit the model by applying (15) with $\boldsymbol{X} = \mathbf{1}_{n_y} \otimes [\mathbf{1}_{n_a} : \boldsymbol{B}_a]$ in (16), $\boldsymbol{H} = \boldsymbol{h}' = (0, \mathbf{1}'_{n_a}\boldsymbol{B}_a)$ and $\boldsymbol{P}$ given by (17). Of course, the optimal value of the smoothing parameter and that of the fitted values are identical whether we use the unconstrained smooth model (5) or the constrained model discussed in this section. We now apply these methods to a discussion of the LC model.

3 Lee-Carter models

Lee and Carter (1992), in an influential paper on mortality forecasting, proposed that

$$\log \lambda_{i,j} = \alpha_i + \beta_i\kappa_j. \tag{18}$$

This simple model has an obvious drawback as a model for mortality: the mortality profile at age $x_{a,i}$ is a scale and location transform of the time index $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{n_y})'$; thus all profiles have the same shape. However, model (18) was designed specifically with forecasting in mind and here the model has one considerable advantage over some of its more complicated and better fitting rivals: the time signal is reduced to a single parameter, $\boldsymbol{\kappa}$. This greatly simplifies the forecasting procedure and generally leads to non-volatile forecasts. Thus, (18) is basically a forecasting model. Our contribution is to show how to improve the forecasting properties of (18).

We will discuss the LC model under four headings: unconstrained fitting methods, constrained fitting methods, the smooth model of Delwarde et al. (2007), the DDE model, and a smooth extension to the Delwarde model, the LC(S) model.

3.1 Unconstrained fitting methods

The difficulty in estimating the parameters in (18) is the product or bi-linear term $\beta_i\kappa_j$. There is no regressor variable and so the usual regression methods cannot be used. The model is also not identifiable, and Lee and Carter imposed the location and scale constraints

$$\sum \kappa_j = 0, \quad \sum \beta_i = 1. \tag{19}$$

Let $\boldsymbol{O} = (o_{i,j})$, $o_{i,j} = \log(d_{i,j}/e_{i,j})$, be the matrix of observed log mortalities. Now under the constraints (19) $\sum_j \log \lambda_{i,j} = n_y\alpha_i$, $i = 1, \ldots, n_a$, and so Lee and Carter estimated $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{n_a})'$ as the row means of $\boldsymbol{O}$. We can now write

$$\boldsymbol{O} - \hat{\boldsymbol{\alpha}}\mathbf{1}'_{n_y} \approx \boldsymbol{\beta}\boldsymbol{\kappa}' \tag{20}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_a})'$. Lee and Carter used the first left and right singular vectors in the singular value decomposition of the left-hand side of (20) as the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$. Thus they viewed estimation in (18) as a matrix approximation problem, rather than as one of statistical estimation.

Brouhns et al. (2002) used the Poisson assumption and maximum likelihood to estimate the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$. They solved the likelihood equations by the uni-dimensional or elementary Newton method, ie, they used the Newton-Raphson method first to update $\boldsymbol{\alpha}$ for given values $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$, second to update $\boldsymbol{\kappa}$ for given values $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and finally to update $\boldsymbol{\beta}$ for given values $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$. The location constraint is imposed on $\hat{\boldsymbol{\kappa}}$ after updating $\boldsymbol{\kappa}$ and the scale constraint is imposed on $\hat{\boldsymbol{\beta}}$ after updating $\boldsymbol{\beta}$. This sequence is then repeated until convergence.

The Brouhns et al. (2002) method can be seen as a special case of our general approach. The LC model is not a GLM but we can use the GLM algorithm for estimation by iterating within a cycle of three GLMs. Motivated by Brouhns et al. (2002) we consider three conditional structures for $\log \lambda_{i,j}$: $\log \lambda_{i,j} = \alpha_i + \tilde{\beta}_i\tilde{\kappa}_j$, $\log \lambda_{i,j} = \tilde{\alpha}_i + \tilde{\beta}_i\kappa_j$ and $\log \lambda_{i,j} = \tilde{\alpha}_i + \beta_i\tilde{\kappa}_j$; here $\tilde{\alpha}_i$, $\tilde{\beta}_i$ and $\tilde{\kappa}_j$ represent current estimates of these parameters. These structures lead to three GLMs defined in turn as follows:

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \tilde{\boldsymbol{\kappa}} \otimes \tilde{\boldsymbol{\beta}} + \boldsymbol{X}\boldsymbol{\alpha}, \ \boldsymbol{X} = \mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a} \tag{21}$$

and we have defined a GLM to estimate $\boldsymbol{\alpha}$ for given $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\kappa} = \tilde{\boldsymbol{\kappa}}$ with model matrix $\boldsymbol{X} = \mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a}$, regression coefficients $\boldsymbol{\theta} = \boldsymbol{\alpha}$ and offset $\log \boldsymbol{e} + \tilde{\boldsymbol{\kappa}} \otimes \tilde{\boldsymbol{\beta}}$. If we insert the model matrix in (21) into the GLM algorithm (2) then the algorithm reduces to

$$\hat{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}} + \boldsymbol{C}_1 \left(\boldsymbol{D} - \tilde{\boldsymbol{D}} \right) \mathbf{1}_{n_y}, \quad \boldsymbol{C}_1 = \mathrm{diag}(\mathbf{1}_{n_a}/\tilde{\boldsymbol{D}}\mathbf{1}_{n_y}), \tag{22}$$

where $\boldsymbol{D}$ is the matrix of observed deaths and $\tilde{\boldsymbol{D}}$ is the matrix of current estimates of deaths. This is precisely the matrix-vector form of the first set of equations on page 379 in Brouhns et al. (2002).

The second GLM is used to estimate $\boldsymbol{\kappa}$ for given $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$:

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\kappa}, \ \boldsymbol{X} = \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}; \tag{23}$$

this GLM has model matrix $\boldsymbol{X} = \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}$, regression coefficients $\boldsymbol{\theta} = \boldsymbol{\kappa}$ and offset $\log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}}$. Insertion of the model matrix into (2) reduces the GLM algorithm to

$$\hat{\boldsymbol{\kappa}} = \tilde{\boldsymbol{\kappa}} + \boldsymbol{C}_2 \left(\boldsymbol{D} - \tilde{\boldsymbol{D}} \right)' \tilde{\boldsymbol{\beta}}, \quad \boldsymbol{C}_2 = \mathrm{diag}(\mathbf{1}_{n_y} / \tilde{\boldsymbol{D}}' \tilde{\boldsymbol{\beta}}^2), \tag{24}$$

where $\tilde{\boldsymbol{\beta}}^2 = \tilde{\boldsymbol{\beta}} * \tilde{\boldsymbol{\beta}}$, the matrix-vector form of the second set of equations in the Brouhns et al. paper.

The third and final GLM is used to estimate $\boldsymbol{\beta}$ for given $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\kappa} = \tilde{\boldsymbol{\kappa}}$:

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta}, \ \boldsymbol{X} = \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}; \tag{25}$$

this GLM has model matrix $\tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}$, regression coefficients $\boldsymbol{\theta} = \boldsymbol{\beta}$ and offset $\log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}}$. Insertion of this model matrix into (2) reduces the GLM algorithm to

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \boldsymbol{C}_3 \left(\boldsymbol{D} - \tilde{\boldsymbol{D}} \right) \tilde{\boldsymbol{\kappa}}, \quad \boldsymbol{C}_3 = \mathrm{diag}(\mathbf{1}_{n_a} / \tilde{\boldsymbol{D}} \tilde{\boldsymbol{\kappa}}^2), \tag{26}$$

where $\tilde{\boldsymbol{\kappa}}^2 = \tilde{\boldsymbol{\kappa}} * \tilde{\boldsymbol{\kappa}}$. Again, this is the matrix-vector form of the third set of equations in the Brouhns et al. paper. We conclude that the maximum likelihood approach of Brouhns et al. (2002) is equivalent to using the above sequence of three GLMs.

3.2 Constrained fitting methods

Brouhns et al. (2002) used equations (22), (24) and (26); the constraints (19) are applied after each iteration. In contrast, equation (14) allows us to include the constraints directly in the estimation process. Our preference is to use the appropriate $\boldsymbol{X}$ and $\boldsymbol{H}$ directly in equation (14) since this allows us to use a general algorithm. Nevertheless, a routine calculation shows how the Brouhns equations are modified when the constraints (19) are included through equation (14). There is no constraint on $\boldsymbol{\alpha}$ so equation (22) stands. Equation (24) to update $\boldsymbol{\kappa}$ becomes

$$\hat{\boldsymbol{\kappa}} = \tilde{\boldsymbol{\kappa}} + \boldsymbol{C}_2 \left(\boldsymbol{D} - \tilde{\boldsymbol{D}} \right)' \tilde{\boldsymbol{\beta}} - (\mathbf{1}'_{n_y} \boldsymbol{C}_2 \mathbf{1}_{n_y})^{-1} \boldsymbol{C}_2 \mathbf{1}_{n_y} \mathbf{1}'_{n_y} \boldsymbol{C}_2 (\boldsymbol{D} - \tilde{\boldsymbol{D}})' \tilde{\boldsymbol{\beta}} \tag{27}$$

where $\boldsymbol{C}_2$ is defined in (24). We note that if $\mathbf{1}'_{n_y} \tilde{\boldsymbol{\kappa}} = 0$ in (27) then $\mathbf{1}'_{n_y} \hat{\boldsymbol{\kappa}} = 0$ so the constraint on $\boldsymbol{\kappa}$ is satisfied. Equation (26) to update $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \boldsymbol{C}_3 \left(\boldsymbol{D} - \tilde{\boldsymbol{D}} \right) \tilde{\boldsymbol{\kappa}} - (\mathbf{1}'_{n_a} \boldsymbol{C}_3 \mathbf{1}_{n_a})^{-1} \boldsymbol{C}_3 \mathbf{1}_{n_a} \mathbf{1}'_{n_a} \boldsymbol{C}_3 (\boldsymbol{D} - \tilde{\boldsymbol{D}}) \tilde{\boldsymbol{\kappa}} \tag{28}$$

where $\boldsymbol{C}_3$ is defined in (26). Similarly, we note that if $\mathbf{1}'_{n_a} \tilde{\boldsymbol{\beta}} = 1$ in (28) then $\mathbf{1}'_{n_a} \hat{\boldsymbol{\beta}} = 1$ so the constraint on $\boldsymbol{\beta}$ is also satisfied. More generally, estimates obtained with (14) do satisfy the constraints; see Corollary 1 in the Appendix.

Our own approach uses the same idea of a sequence of GLMs but there are two important differences. First, we use the constrained version (14) of the GLM algorithm and so the constraints on $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ are built in to the estimation algorithm. Second, our estimation is

based not on three but on two conditional structures for $\log \lambda_{i,j}$ as follows: $\log \lambda_{i,j} = \alpha_i + \tilde{\beta}_i \kappa_j$ and $\log \lambda_{i,j} = \tilde{\alpha}_i + \beta_i \tilde{\kappa}_j$; in other words we combine the GLMs (21) and (23) above into a single GLM. This allows us to use the full Hessian matrix for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\kappa}}$ in the estimation algorithm; see (32) below. These two changes have two consequences: first, the efficiency of the estimation process is improved; second, conditional on the value of $\hat{\boldsymbol{\beta}}$ we can compute the variance matrix of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\kappa}}$ and hence their canonical correlations.

The first GLM is used to estimate $\boldsymbol{\beta}$ for given values of $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ and is defined in (25). Further we have the constraint $\sum \beta_i = 1$ which gives $\boldsymbol{H} = \boldsymbol{h}' = \mathbf{1}'_{n_a}$ and $\boldsymbol{k} = 1$. Estimation uses (14). We refer to this GLM as GLM1 and for clarity its linear predictor is given below together with the linear predictor for GLM2, the GLM used to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ for given value of $\boldsymbol{\beta}$:

$$\text{GLM1}: \qquad \boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{X} = \tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a} \tag{29}$$

$$\text{GLM2}: \qquad \boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \boldsymbol{X}\boldsymbol{\theta},\ \boldsymbol{X} = [\mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a} : \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}]. \tag{30}$$

Thus GLM2 has model matrix $[\mathbf{1}_{n_y} \otimes \boldsymbol{I}_{n_a} : \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}]$, regression coefficients $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}')'$ and offset $\log \boldsymbol{e}$. Further, we have the constraint $\sum \kappa_j = 0$; we define the constraints matrix $\boldsymbol{H} = \boldsymbol{h}' = (\mathbf{0}'_{n_a}, \mathbf{1}'_{n_y})$. Estimation uses the constrained GLM algorithm (14) with $\boldsymbol{k} = 0$.

We can see the difference between estimation with the unconstrained forms (21), (23) and (25), on the one hand, and the constrained forms GLM1 and GLM2, (29) and (30), on the other. The algorithm (14) for GLM2 reduces to

$$\begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} & : & \boldsymbol{h} \\ \boldsymbol{h}' & : & 0 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \hat{\omega} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{\theta}} + \boldsymbol{X}'(\boldsymbol{d} - \tilde{\boldsymbol{d}}) \\ 0 \end{pmatrix} \tag{31}$$

where

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} = \begin{pmatrix} \operatorname{diag}(\tilde{\boldsymbol{D}}\mathbf{1}_{n_y}) & : & \tilde{\boldsymbol{\beta}} * \tilde{\boldsymbol{D}} \\ \tilde{\boldsymbol{\beta}}' * \tilde{\boldsymbol{D}}' & : & \operatorname{diag}(\tilde{\boldsymbol{D}}'\tilde{\boldsymbol{\beta}}^2) \end{pmatrix}, \tag{32}$$

$$\boldsymbol{X}'(\boldsymbol{d} - \tilde{\boldsymbol{d}}) = \begin{pmatrix} (\boldsymbol{D} - \tilde{\boldsymbol{D}})\mathbf{1}_{n_y} \\ (\boldsymbol{D} - \tilde{\boldsymbol{D}})'\tilde{\boldsymbol{\beta}} \end{pmatrix}, \tag{33}$$

$\boldsymbol{h} = (\mathbf{0}'_{n_a}, \mathbf{1}'_{n_y})'$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\kappa}')'$. If we delete the off-diagonal blocks in $\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}$ and the final row and column in the matrix on the LHS of (31), then equation (31) reduces to the unconstrained algorithms for $\hat{\boldsymbol{\alpha}}$, (22), and $\hat{\boldsymbol{\kappa}}$, (24). The off-diagonal blocks in $\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}$ take account of the covariance between $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\kappa}}$, and increase the efficiency of the estimation algorithm.

The form of equations (32) and (33) deserves comment. The LC model is a "row-and-column model" and the data are in the form of matrices $\boldsymbol{D}$ and $\boldsymbol{E}$. GLMs with this structure are examples of generalized linear array models or GLAMs (Currie et al., 2006). For such models the IWLS algorithm (2) exists in an accelerated form. The estimating equations (22), (24) and (26) are all examples of GLAM-style computations. The Kronecker products in the specifications of the associated models lead to sparse model matrices and inefficient computation in (2); in contrast, the GLAM computations in (22), (24) and (26) are very

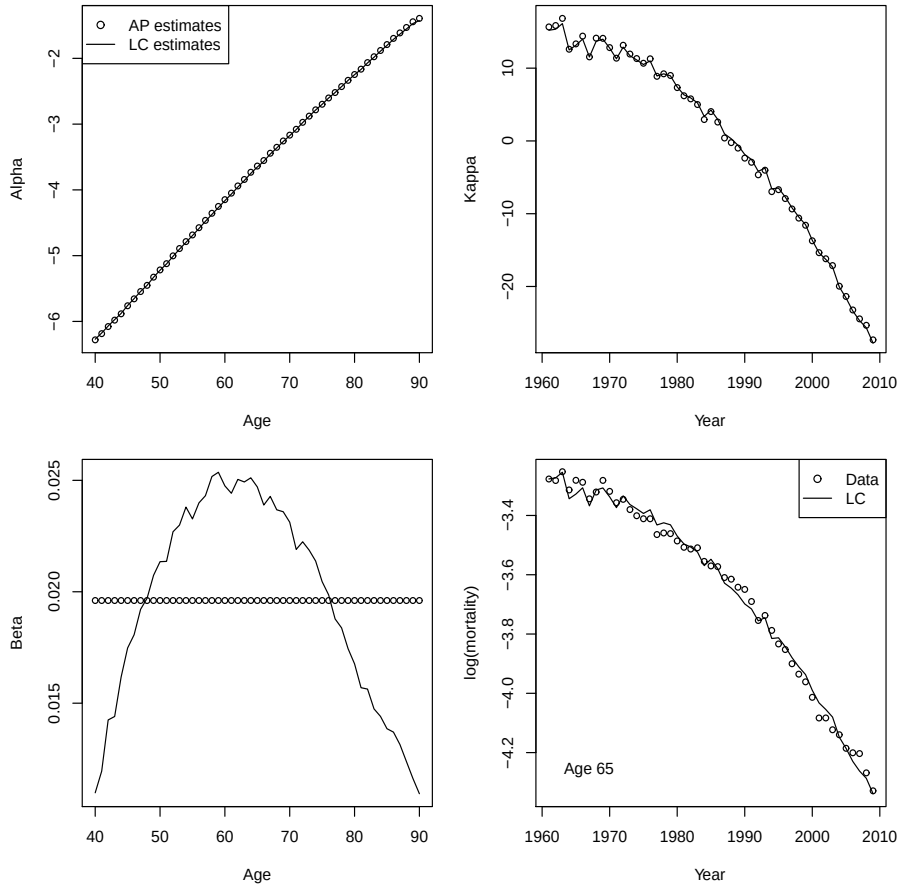

Figure 1: Estimates of $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\kappa}}$ in the Age-Period model (34) (hollow disk, $\circ$) and LC (18) (line, $-$) models; observed and fitted log(mortality) for age 65.

efficient. The evaluation of (31) with (32) and (33) is a clear example of the computational gain possible with GLAM.

The LC model is fitted by iterating between GLM1 and GLM2. Initial values of $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are given by the row and column means of the observed matrix of log mortalities, $\boldsymbol{O}$; initial values of $\boldsymbol{\beta}$ are set to $\mathbf{1}_{n_a}/n_a$, but we note that the initial values of $\boldsymbol{\kappa}$ must be centred (since $\sum \kappa_j = 0$) and then scaled (since $\sum \beta_i = 1$).

We fit the LC model to the ONS data described in section 2. Figure 1 displays the fitted parameter values together with a plot of log(mortality) for age 65. The right hand panels bear out our remark about the shape of the mortality curve for any age being identical to the shape of $\hat{\boldsymbol{\kappa}}$. Figure 1 also highlights an important stability property of the LC model: the introduction of the $\boldsymbol{\beta}$ term into the Age-Period model (Clayton and Schifflers, 1987a)

$$\log \lambda_{i,j} = \alpha_i + \frac{1}{n_a}\kappa_j, \tag{34}$$

ie, the LC model with all $\beta_i = 1/n_a$, has not distorted the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$. The

variance matrix in a constrained GLM is given by (60) and Corollary 2. Conditional on $\hat{\boldsymbol{\beta}}$ we compute the variance matrix of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\kappa}}$ in GLM2 from which we estimate their first canonical correlation; we find $r(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\kappa}}) = 0.21$, a reassuringly small number. The stability of the estimates and the low canonical correlation both suggest that forecasting the mortality table by forecasting $\hat{\boldsymbol{\kappa}}$ is possible.

3.3 The model of Delwarde, Denuit and Eilers

Delwarde et al. (2007) observed that "the estimated β_i's exhibit an irregular pattern in most cases, and this produces irregular projected life tables." The lower left panel in Figure 1 is a typical plot of $\hat{\boldsymbol{\beta}}$. Their solution was to smooth the values of β_i using penalized likelihood. We first show how the DDE model fits into our general approach and then propose a further improvement.

Let $\boldsymbol{B}_a$ be a regression matrix of cubic B-splines on age, as in section 2.3, and then set $\boldsymbol{\beta} = \boldsymbol{B}_a\boldsymbol{b}$. We now consider two GLMs as follows. Corresponding to GLM1 and GLM2 in (29) and (30) respectively, we have

$$\begin{aligned} &\text{GLM1}* : \quad \boldsymbol{\eta} = \log \boldsymbol{\mu} = \log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}} + \boldsymbol{X}\boldsymbol{b},\ \boldsymbol{X} = [\tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}]\boldsymbol{B}_a. \\ &\text{GLM2} \quad : \quad \text{as GLM2 in (30).} \end{aligned} \tag{35}$$

GLM1$*$ is a penalized GLM with regression matrix $[\tilde{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}]\boldsymbol{B}_a$, offset $\log \boldsymbol{e} + \mathbf{1}_{n_y} \otimes \tilde{\boldsymbol{\alpha}}$ and regression coefficients $\boldsymbol{\theta} = \boldsymbol{b}$. The constraint $\sum \beta_i = 1$ is equivalent to $\boldsymbol{H} = \mathbf{1}'_{n_a}\boldsymbol{B}_a$ and $\boldsymbol{k} = 1$, and the penalty matrix is $\boldsymbol{P} = \tau_\beta \boldsymbol{D}'_2\boldsymbol{D}_2$; both $\boldsymbol{H}$ and $\boldsymbol{P}$ act on $\boldsymbol{b}$. For given τ_β, GLM1$*$ is fitted with (15). To estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ for given $\boldsymbol{\beta}$ we use GLM2 without alteration. The smoothing parameter is selected by minimizing BIC.

We note that the model we have described is different in detail from Delwarde et al. (2007) who in effect set $\boldsymbol{B}_a = \boldsymbol{I}_{n_a}$. Their model can be fitted with the above scheme provided the penalty matrix $\boldsymbol{P}$ acts directly on $\boldsymbol{\beta}$ instead of $\boldsymbol{b}$.

The left panel of Figure 2 shows the two estimates of $\boldsymbol{\beta}$; the $n_a = 51$ parameters $\boldsymbol{\beta}$ in the original formulation of the LC model have been replaced with a smooth function with approximately 9 degrees of freedom. Forecasts of mortality are obtained by forecasting $\boldsymbol{\kappa}$ and then applying (18). For illustration we forecast $\boldsymbol{\kappa}$ from 2010 to 2050 with an ARIMA(1,1,1) model. Fitted values from the LC and DDE models can scarcely be distinguished by eye but the forecast with the DDE model improves on the LC model in two respects. First, the right panel of Figure 2 shows a crossover in the forecast mortalities at ages 41 and 42 with the original model which is eliminated (at least up to year 2050) with the DDE model. Second, Figure 3 shows how the general regularity of the forecast over age is improved with the DDE model. Pensions and annuities depend on the future course of mortality so for the pricing and reserving of such contracts actuaries will prefer regular forecasts; the case that they should prefer the DDE model to the original LC model looks strong.

There are two advantages in using our general estimating algorithm (15) instead of detailed equations such as (22), (24) and (26). First, all the models in this paper can be fitted by

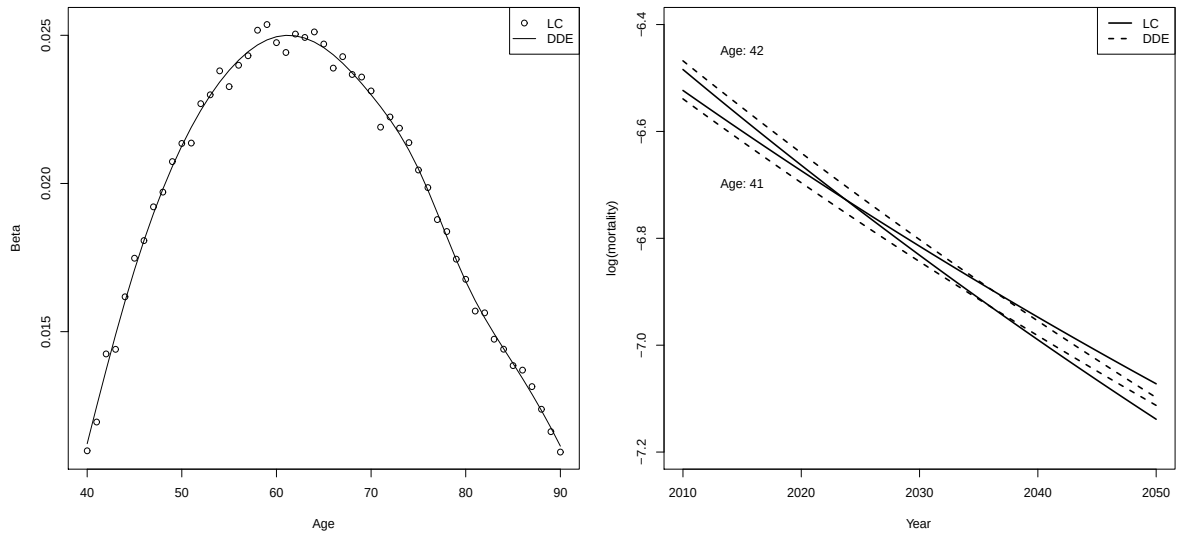

Figure 2: Left: estimated $\boldsymbol{\beta}$ with original LC (circles) and DDE (solid line) models; right: crossover for ages 41 and 42 with LC (solid line) and improved forecast with DDE (dashed line).

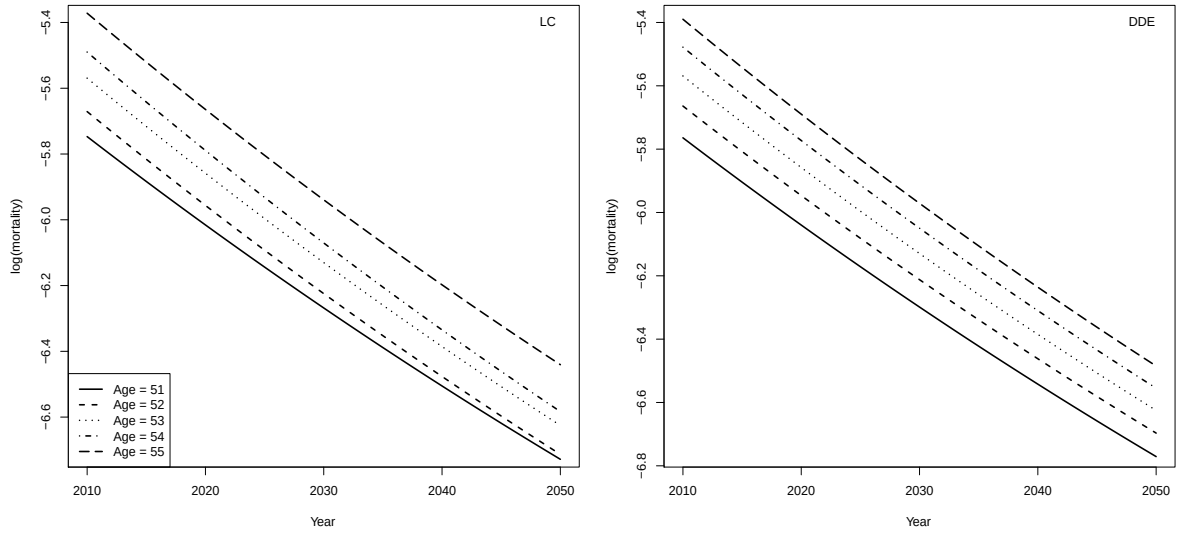

Figure 3: Forecast log mortality for ages 51 to 55 with original LC model (left panel) and DDE model (right panel).

specifying the matrices $\boldsymbol{X}$, $\boldsymbol{H}$ and $\boldsymbol{P}$ and then using the general algorithm (15). Second, we avoid detailed calculation which can be prone to error. Indeed, the estimating equation for $\boldsymbol{\beta}$ on page 39 in Delwarde et al. (2007) should read, in our notation,

$$\left[\text{diag}(\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\kappa}}^2) + \boldsymbol{P}\right]\hat{\boldsymbol{\beta}} = \text{diag}(\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\kappa}}^2)\tilde{\boldsymbol{\beta}} + (\boldsymbol{D} - \tilde{\boldsymbol{D}})\tilde{\boldsymbol{\kappa}} \tag{36}$$

or, in the notation of Delwarde,

$$\left(C_\beta^{(k)} + P_\beta\right)\hat{\boldsymbol{\beta}}^{(k+1)} = C_\beta^{(k)}\hat{\boldsymbol{\beta}}^{(k)} + r_\beta^{(k)}. \tag{37}$$

3.4 Smoothing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

The DDE model is successful in producing more regular life tables but, in the spirit of Delwarde et al. (2007), we suggest one further improvement: the estimate of $\boldsymbol{\alpha}$ be smoothed. The LC(S) model with smooth $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ results in a further loss of fit but the smoothness of the projected lifetable is improved. In GLM2 (30) we set $\boldsymbol{\alpha} = \boldsymbol{B}_a\boldsymbol{a}$. We have the coupled penalized GLMs

$$\begin{aligned} &\text{GLM1}* \;:\; \text{as (35)} \\ &\text{GLM2}* \;:\; \boldsymbol{\eta} = \log\boldsymbol{\mu} = \log\boldsymbol{e} + \boldsymbol{X}\boldsymbol{\theta},\ \boldsymbol{X} = [\mathbf{1}_{n_y} \otimes \boldsymbol{B}_a : \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}] \end{aligned} \tag{38}$$

with model matrix $[\mathbf{1}_{n_y} \otimes \boldsymbol{B}_a : \boldsymbol{I}_{n_y} \otimes \tilde{\boldsymbol{\beta}}]$, offset $\log\boldsymbol{e}$ and regression coefficients $\boldsymbol{\theta}' = (\boldsymbol{a}', \boldsymbol{\kappa}')'$. The constraint $\sum \kappa_j = 0$ is equivalent to $\boldsymbol{H} = (\mathbf{0}'_{c_a}, \mathbf{1}'_{n_y})$ and $\boldsymbol{k} = 0$, and the penalty matrix is $\boldsymbol{P} = \text{blockdiag}\{\tau_\alpha \boldsymbol{D}_2'\boldsymbol{D}_2, 0 * \boldsymbol{I}_{n_y}\}$; both $\boldsymbol{H}$ and $\boldsymbol{P}$ act on $\boldsymbol{\theta} = (\boldsymbol{a}', \boldsymbol{\kappa}')'$. We fit LC(S) by iterating between the two penalized GLMs, GLM1$*$ and GLM2$*$.

Figure 4 shows the differences in log(mortality) at successive ages for the start, 2010, and end, 2050, of the forecast period. A regular lifetable implies that these differences are smooth. We see that the DDE model addresses most of the irregularities that result from the original LC model; the LC(S) model removes what remains.

3.5 Summary of model structure

A constrained smooth GLM is specified by defining its regression matrix, its penalty matrix, its matrix of constraints and its smoothing parameters. Inspection of the model structures GLM1 and GLM2 and their smooth variants, GLM1$*$ and GLM2$*$, reveals that they have a common structure. For example, GLM1 and GLM1$*$ for the estimation of $\boldsymbol{\beta}$ both have regression matrix of the form $[\hat{\boldsymbol{\kappa}} \otimes \boldsymbol{I}_{n_a}]\boldsymbol{X}_\beta$ where $\boldsymbol{X}_\beta = \boldsymbol{I}_{n_a}$ for GLM1 while $\boldsymbol{X}_\beta = \boldsymbol{B}_a$ for GLM1$*$. In a similar way, we set the penalty matrix to $0 * \boldsymbol{I}_{n_a}$ when $\boldsymbol{\alpha}$ is not smoothed and to $\tau_\alpha * \boldsymbol{D}_2'\boldsymbol{D}_2$ when it is; in the latter case τ_α must be estimated. Table 1 provides a summary of this common structure.

Eilers and Marx (1996) discuss the limiting case when a smoothing parameter tends to infinity; the limit is a straight line in the case of a second order penalty. Thus, if we set τ_α to a very large value, eg, 10^{20}, in the LC(S) model this forces a straightline fit on $\boldsymbol{\alpha}$. We think of this model as a Gompertz variant of the LC(S) model.

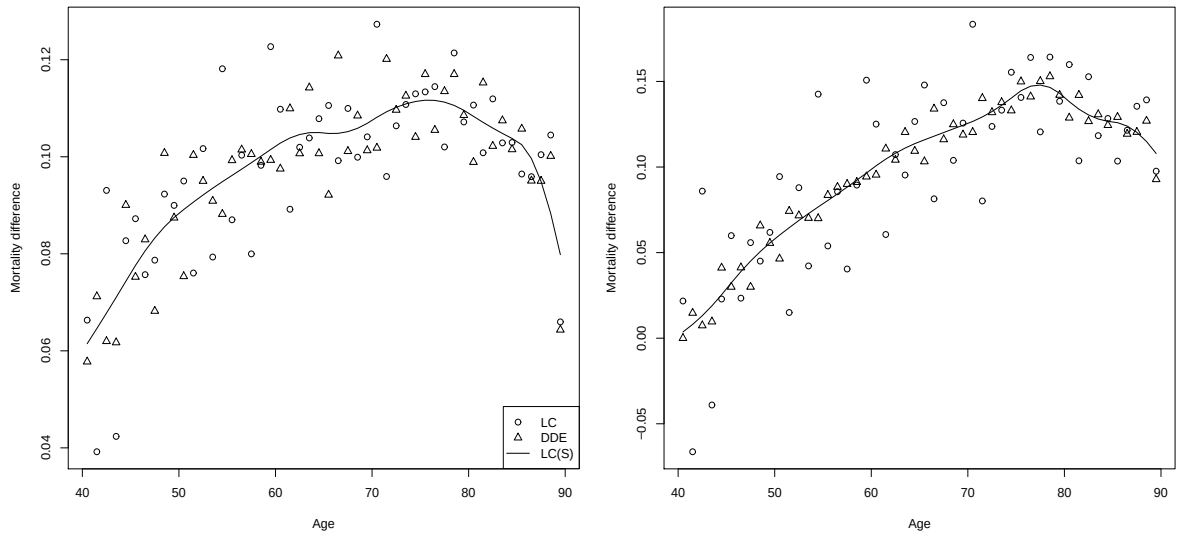

Figure 4: Differences in forecast log(mortality) between successive ages in 2010 (left panel) and 2050 (right panel) for the LC, DDE and LC(S) models.

4 Binomial models for q

The focus of the paper so far has been the Poisson model with log link for the force of mortality. Actuaries are often more interested in q, the probability of death in a single year. In section 2 we let $\lambda_{i,j}$ be the force of mortality at age i in year j; there we noted that strictly, we should write $\lambda_{i+0.5,j+0.5}$. More generally, let $\lambda_{i+s,j+s}$ be the force of mortality at exact age $x_{a,i}+s$ at time $x_{y,j}+s$. Let $q_{i,j}$ be the probability that a life with exact age $x_{a,i}$ at outset in year $x_{y,j}$ dies in the following year, ie, aged $x_{a,i}$. Then

$$q_{i,j} = 1 - \exp\left(-\int_0^1 \lambda_{i+s,j+s}\, ds\right) \approx 1 - \exp(-\lambda_{i+0.5,j+0.5}). \tag{39}$$

The standard procedure is to use this equation to estimate q from the estimates of λ. It is also possible to model q directly. Indeed, Cairns et al. (2009) present a number of models for q; these authors used a logit link on q but retained the Poisson model for the number of deaths. Our approach also uses the logit link but with a binomial error for the number of deaths.

We have the mid-year exposures $e_{i,j}$, so an estimate of the exposure at the start of calendar year $x_{y,j}$ is $e^*_{i,j} = e_{i,j} + 0.5 * d_{i,j}$; actuaries refer to this as the *initial exposure*. We can now model the number of deaths as $D_{i,j} \sim \mathcal{B}(e^*_{i,j},\, q_{i,j})$. If $y_{i,j} = d_{i,j}/e^*_{i,j}$ then we can take y as the dependent variable in a GLM with binomial error and logit link. The estimating equations in the previous sections all hold with three changes: the offset $\log \boldsymbol{e}$ which appears with the Poisson model is set to zero. The diagonal matrix of weights and the working variable in (3)

Table 1: Model specification for LC, DDE and LC(S) models

		$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\kappa}$
LC	Regression	$\boldsymbol{I}_{n_a}$	$\boldsymbol{I}_{n_a}$	$\boldsymbol{I}_{n_y}$
	Penalty	$\boldsymbol{I}_{n_a}$	$\boldsymbol{I}_{n_a}$	-
	Constraint	$\boldsymbol{0}'_{n_a}$	$\boldsymbol{1}'_{n_a}$	$\boldsymbol{1}'_{n_y}$
	$(\tau_\alpha, \tau_\beta)$	0	0	-
DDE	Regression	$\boldsymbol{I}_{n_a}$	$\boldsymbol{B}_a$	$\boldsymbol{I}_{n_y}$
	Penalty	$\boldsymbol{I}_{n_a}$	$\boldsymbol{D}'_2\boldsymbol{D}_2$	-
	Constraint	$\boldsymbol{0}'_{n_a}$	$\boldsymbol{1}'_{n_a}\boldsymbol{B}_a$	$\boldsymbol{1}'_{n_y}$
	$(\tau_\alpha, \tau_\beta)$	0	τ_β	-
LC(S)	Regression	$\boldsymbol{B}_a$	$\boldsymbol{B}_a$	$\boldsymbol{I}_{n_y}$
	Penalty	$\boldsymbol{D}'_2\boldsymbol{D}_2$	$\boldsymbol{D}'_2\boldsymbol{D}_2$	-
	Constraint	$\boldsymbol{0}'_{n_a}\boldsymbol{B}_a$	$\boldsymbol{1}'_{n_a}\boldsymbol{B}_a$	$\boldsymbol{1}'_{n_y}$
	$(\tau_\alpha, \tau_\beta)$	τ_α	τ_β	-

Table 2: Poisson and binomial deviances for (a) Poisson model with log link and (b) binomial model with logit link.

	Poisson model			Binomial model		
Model	Dev_P	Dev_B	ED	Dev_P	Dev_B	ED
LC	17944	19374	149	16624	17854	149
DDE	18070	19509	107	16745	17981	108
LC(S)	18295	19752	69	16951	18204	69

are replaced by

$$\tilde{\boldsymbol{W}} = \text{diag}\{\boldsymbol{e}^*\tilde{\boldsymbol{q}}(\boldsymbol{1}-\tilde{\boldsymbol{q}})\},\ \tilde{\boldsymbol{z}} = \boldsymbol{X}\tilde{\boldsymbol{\theta}} + \frac{\boldsymbol{y}-\tilde{\boldsymbol{q}}}{\tilde{\boldsymbol{q}}(\boldsymbol{1}-\tilde{\boldsymbol{q}})}; \tag{40}$$

here $\boldsymbol{y}$ and $\tilde{\boldsymbol{q}}$ are the vectors of observed and current estimates of the probabilities of death respectively, and $\boldsymbol{e}^*$ is the vector of initial exposures.

It is of interest to compare the two approaches: the Poisson model with log link and the binomial model with logit link. Figures 1, 2, 3 and 4 can scarcely be distinguished under either model. However, there are differences of detail. We define the Poisson and binomial deviances

$$\text{Dev}_P = 2\sum\left[d\log(d/\hat{d}) - (d-\hat{d})\right] \tag{41}$$

and

$$\text{Dev}_B = 2\sum\left\{d\log(d/\hat{d}) + (e^*-d)\log[(e^*-d)/(e^*-\hat{d})]\right\} \tag{42}$$

where the summation is over all observed and fitted deaths (McCullagh and Nelder, 1989).

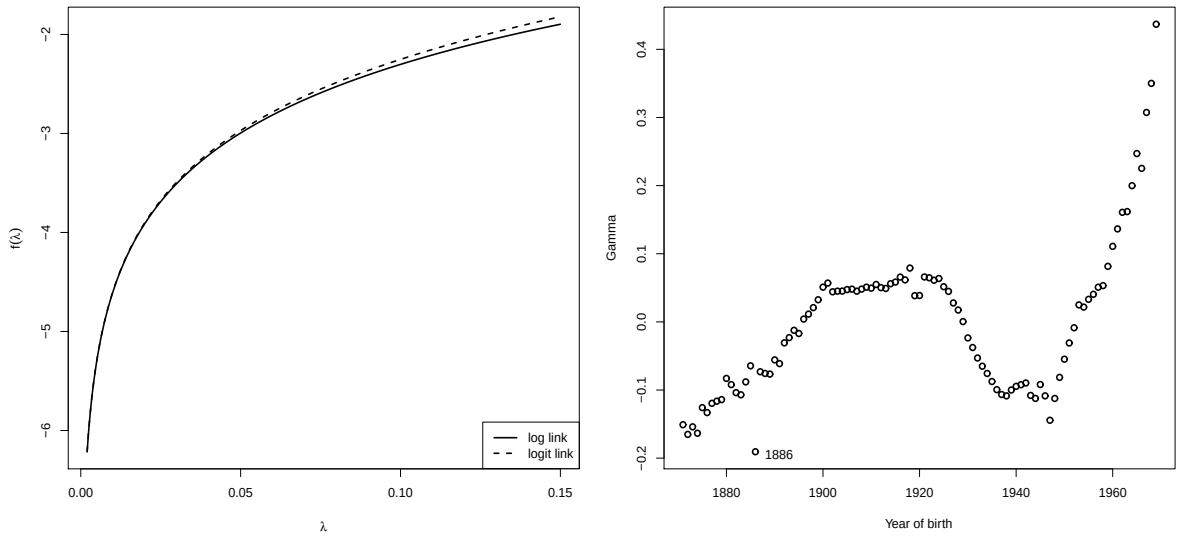

Figure 5: Left: comparison of the log link for the force of mortality, λ, and the logit link for the probability of death in a single year, q; right: cohort parameters in age-period-cohort model.

We interpret Dev_P and Dev_B as measures of fit. Table 2 gives the values of Dev_P, Dev_B and ED, the effective dimension, for all three models. The Poisson deviance, Dev_P, is over 1300 smaller with the fitted values from the binomial model compared to its value with the fitted values from the Poisson model; similarly, the binomial deviance, Dev_B, is over 1500 smaller with the fitted values from the binomial model compared to its value with the fitted values from the Poisson model. We conclude that the binomial model for q with logit link gives a substantially improved fit over the Poisson model for λ with log link. The logit link for q is equivalent to the non-standard link $\eta = \log(\exp(\lambda) - 1)$ for λ. The left panel of Figure 5 compares the two link functions $\log \lambda$ and $\log(\exp(\lambda) - 1)$ over the range $\lambda \in (0.002, 0.15)$ which covers roughly the male force of mortality from age 40 to 90 (ONS, 2009).

Part of the explanation for the better fit of the binomial model may lie with two particular cohorts in the ONS data. We define a residual to be large if its contribution to Dev_P exceeds 100. There are eleven such residuals, seven of them are in the cohort with year of birth 1886 (lives which died age 78 to 84) and the remaining four are in the cohort with year of birth 1919 (lives which died age 85, 87, 88 and 89). These are all lives which died at older ages, the ages at which the log and logit links differ materially. The right panel of Figure 5 displays the cohort parameters $\boldsymbol{\gamma}$ in the age-period-cohort model $\log \lambda_{i,j} = \alpha_i + \kappa_j + \gamma_{j-i}$; the year of birth 1886 stands out from the general pattern in $\boldsymbol{\gamma}$.

5 Concluding remarks

We have given a general solution to the problem of fitting a GLM with smooth components and subject to linear constraints. One of the main attractions of the method is that the solution is specified at a high level. We define three matrices: the regression matrix $\boldsymbol{X}$, the constraints matrix $\boldsymbol{H}$ and the smoothing or penalty matrix $\boldsymbol{P}$; additionally we define the constraints vector $\boldsymbol{k}$. The matrices $\boldsymbol{X}$, $\boldsymbol{H}$ and $\boldsymbol{P}$, and the vector $\boldsymbol{k}$ are then supplied as arguments to a general estimating function.

We have supplied software in R together with the ONS data; this allows the reader to reproduce the results in the paper for the Poisson based models. The software can also be used to fit any of our models to the reader's own data. One of the attractions of the original LC model is its ease of fitting. The smooth variants of the LC model are conceptually as simple as Lee and Carter's original proposal but they are less accessible to the general user. Our software should help to ease this difficulty.

There are many variants of the LC model; Booth et al. (2006) provide a good summary. Generally, these variants aim at improving the fit of the model. The focus of our paper is different: we aim to produce smooth forecasts, and our method of smoothing $\boldsymbol{\beta}$ and perhaps $\boldsymbol{\alpha}$ could be applied to any Lee-Carter style model.

We have illustrated our method with a detailed discussion of the LC model but the method can be applied to many other models of mortality. In a comprehensive paper Cairns et al. (2009) discussed eight models of mortality. These models range from GLMs with no identifiability constraints through GLMs subject to linear constraints up to models, like the LC model, where we have coupled GLMs subject to linear constraints. The age-period-cohort or APC model (Clayton and Schifflers, 1987b)

$$\eta_{i,j} = \log \mu_{i,j} = \log e_{i,j} + \alpha_i + \kappa_j + \gamma_{j-i} \tag{43}$$

is a GLM and so can be fitted immediately with the `glm( )` function in R, for example. However, the model is not identifiable and our approach allows additional possibilities: the specification of particular constraints and the smoothing of any of the three components; see Currie (2012) for a discussion of using constrained methods to fit the APC model.

Renshaw and Haberman (2006) generalized both the LC and the APC models:

$$\eta_{i,j} = \log \mu_{i,j} = \log e_{i,j} + \alpha_i + \beta_i\kappa_j + \delta_i\gamma_{j-i}. \tag{44}$$

This model requires two location constraints and two scale constraints to ensure identifiability. It can be fitted by considering two conditional GLMs: (a) for given β_i and δ_i we estimate α_i, κ_j and γ_{j-i} and (b) for given α_i, κ_j and γ_{j-i} we estimate β_i and δ_i. Cairns et al. (2009) reported very slow convergence in the Renshaw-Haberman model. Our algorithm can be used to fit the Renshaw-Haberman and other models with cohort effects, and we will report on an investigation of this approach in a subsequent paper.

GLMs subject to linear constraints occur widely in many applied areas of statistics. For example, Goodman (1979) described several row-column association models for the analysis

of contingency tables. These models invariably involve constraints and so our methods apply. For example, the RC(1) model is defined

$$\log \mu_{i,j} = \alpha_i + \beta_j + \gamma_i \delta_j \tag{45}$$

where $\mu_{i,j}$ is a cell mean. The model can be fitted with our coupled GLMs with constraints method, and smoothing is an option if appropriate.

In conclusion, the LC model is an important model for the forecasting of mortality. Actuaries require regular forecasts across age for the fair pricing of pension products. Irregularities in the forecast are caused by irregularities (a) in the coefficients $\boldsymbol{\beta}$ modulating the time index $\boldsymbol{\kappa}$ and, to a lesser extent, (b) in the age coefficients $\boldsymbol{\alpha}$. We have developed a general approach to estimation in GLMs subject to linear constraints and smoothing. Application of this method allows us to smooth the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, and regular forecasts result.

Acknowledgements: I am grateful to Andrew Cairns, Viani Djeundje and Stephen Richards for useful discussions, to Maria Durban and the Spanish Ministry of Science and Innovation (project MTM2011-28285-C02-02) for financial support and to the Office for National Statistics for access to their data.

References

Booth, H., Hyndman, R.J., Tickle, L. and de Jong, P. (2006) Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, **15**, 289-310.

Brouhns, N., Denuit, M. and 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., Coughlan, G.D. *et al.* (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. and 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. and Schifflers, E. (1987b) Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine,* **6**, 469-481.

Currie, I. D. (2012) Forecasting with the age-period-cohort model? *Proceedings of 27th International Workshop on Statistical Modelling,* Prague, 87-92.

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

Currie, I. D., Durban, M. and Eilers, P. H. C. (2006) Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society*, Series B, 68, 259-280.

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

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

Gompertz, B. (1825) On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, **115**, 513-583.

Goodman, L.A. (1979) Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537-552.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models.* London: Chapman and Hall, 39-81.

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* London: Chapman and Hall, 15-34.

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

ONS (2009) Office for National Statistics, English Life Tables No 16 (2000-2002).

ONS (2011) Office for National Statistics, England Interim Life Tables, 1980-82 to 2008-10.

R Development Core Team (2011) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Renshaw, A.R. and Haberman, S. (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**, 556-570.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* New York: Wiley, 257-271.

Strang, G. (1986). *Introduction to Applied Mathematics.* Cambridge, USA: Wellesley Cambridge Press, 96-107.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* London: Chapman & Hall/CRC, 121-143.

6 Appendix

We prove (15) in the general setting of a GLM with canonical link function. Let Y be a random variable with $E(Y) = \mu$ and log likelihood

$$\ell = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{46}$$

for some functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. Thus Y is a member of the *exponential family*; see McCullagh and Nelder (1989), for example. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ be a vector of regressor variables and suppose $\theta = \boldsymbol{x}'\boldsymbol{\beta}$ for some regression coefficients $\boldsymbol{\beta}$. When such a connection between θ and the regressor variables exists, θ is known as the *canonical parameter*. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be independent observations on Y and let $\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the regression matrix. Suppose $\boldsymbol{X}$, $n \times p$, has rank $p - q$ and $\boldsymbol{H}$, $q \times p$, is such that $(\boldsymbol{X}' : \boldsymbol{H}')$ has rank p. (Such a matrix exists by Lemma 1 below). Suppose that $\boldsymbol{\beta}$ is subject to a penalty $\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}$ where $\boldsymbol{P}$ is positive semi-definite. (For example, we may wish to smooth $\boldsymbol{\beta}$ in some way.) Then

Theorem: The maximum likelihood estimate of $\boldsymbol{\beta}$ subject to

(i) the linear constraints $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{k}$, where $\boldsymbol{k}$ is a vector of known constants, and

(ii) the penalty $\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}$

is the unique solution of the iterative least squares equation

$$\begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P} & : & \boldsymbol{H}' \\ \boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} \\ \boldsymbol{k} \end{pmatrix}; \tag{47}$$

here, the tilde represents an approximate solution, as in $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ is an improved estimate,

$$\tilde{\boldsymbol{W}} = \operatorname{diag}\left\{\frac{b''(\theta_i)}{a_i(\phi)}\right\} \tag{48}$$

is the *diagonal matrix of weights* and

$$\tilde{\boldsymbol{z}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \operatorname{diag}\left\{\frac{1}{b''(\theta_i)}\right\}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) \tag{49}$$

is the *working variable*.

Comment. If $\boldsymbol{H} = \boldsymbol{0}$ and $\boldsymbol{P} = \boldsymbol{0}$ the equations (47) reduce to the standard GLM algorithm. The algorithm has the form of iterative weighted least squares with dependent variable $\tilde{\boldsymbol{z}}$ and weight $\tilde{\boldsymbol{W}}$; the algorithm is often known as IWLS. We will refer to (47) as the extended IWLS algorithm.

Proof. The log likelihood for a single observation y from (46) is

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi),\ \theta = \boldsymbol{x}'\boldsymbol{\beta} \\
\Rightarrow \frac{\partial \ell}{\partial \beta_j} &= \frac{y - b'(\theta)}{a(\phi)} x_j,\ j = 1,\ldots,p \qquad (50)\\
&= \frac{y-\mu}{a(\phi)} x_j,\ \text{since } b'(\theta) = \mu \text{ (McCullagh and Nelder, 1989)} \\
\Rightarrow \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} x_{ij} \text{ on summing over the sample, } j = 1,\ldots,p \\
\Rightarrow \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \text{ where } \boldsymbol{\Lambda}_a = \text{diag}\,\{a_i(\phi)\} \text{ and } \boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)'. \qquad (51)
\end{aligned}$$

If we introduce restrictions on $\boldsymbol{\beta}$ with the penalty $\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}$ then we maximize the penalized log likelihood

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta}.$$

Furthermore we must maximize $\ell_p(\boldsymbol{\beta})$ subject to $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{k}$. Hence we consider the Lagrangian

$$G(\boldsymbol{\beta}, \boldsymbol{\omega}) = \ell(\boldsymbol{\beta}) - \tfrac{1}{2}\boldsymbol{\beta}'\boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{\omega}'(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{k})$$

where $\boldsymbol{\omega}$ is the vector of Lagrange multipliers; see Strang (1986) for a discussion of Lagrange methods. Using (51), we must solve

$$\begin{aligned}
\frac{\partial G}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{H}'\boldsymbol{\omega} = \boldsymbol{0} \qquad (52)\\
\frac{\partial G}{\partial \boldsymbol{\omega}} &= -\boldsymbol{H}\boldsymbol{\beta} + \boldsymbol{k} = \boldsymbol{0}. \qquad (53)
\end{aligned}$$

We solve this system by Newton-Raphson. Let $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\omega}')'$. Suppose $\boldsymbol{\xi}$ solves (52) and (53) and let $\tilde{\boldsymbol{\xi}}$ be an approximate solution. Then expanding $\partial G/\partial \boldsymbol{\xi}$ about $\tilde{\boldsymbol{\xi}}$ we find

$$\begin{aligned}
\boldsymbol{0} = \frac{\partial G}{\partial \boldsymbol{\xi}} &\approx \frac{\partial G}{\partial \boldsymbol{\xi}}_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}} + \frac{\partial^2 G}{\partial \boldsymbol{\xi}^2}_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}} (\boldsymbol{\xi} - \tilde{\boldsymbol{\xi}}) \\
\Rightarrow \frac{\partial^2 G}{\partial \boldsymbol{\xi}^2}_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}} \boldsymbol{\xi} &= \frac{\partial^2 G}{\partial \boldsymbol{\xi}^2}_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}} \tilde{\boldsymbol{\xi}} - \frac{\partial G}{\partial \boldsymbol{\xi}}_{\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}} \qquad (54)
\end{aligned}$$

Differentiating (50) wrt β_k we find

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\frac{b''(\theta)}{a(\phi)} x_j x_k,\ j,k = 1,\ldots,p$$

and so

$$\frac{\partial^2 G}{\partial \boldsymbol{\beta}^2} = -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} - \boldsymbol{P} \qquad (55)$$

where $\boldsymbol{W}$ is given by (48). Furthermore, from (52) and (53), we have

$$\frac{\partial^2 G}{\partial \boldsymbol{\beta}\partial \boldsymbol{\omega}} = -\boldsymbol{H}', \quad \frac{\partial^2 G}{\partial \boldsymbol{\omega}\partial \boldsymbol{\beta}} = -\boldsymbol{H}, \quad \frac{\partial^2 G}{\partial \boldsymbol{\omega}^2} = \boldsymbol{0}. \qquad (56)$$

Substituting (55) and (56) in (54), we find

$$\text{LHS} = \begin{pmatrix} -\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} - \boldsymbol{P} & : & -\boldsymbol{H}' \\ -\boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\omega}} \end{pmatrix}$$

and

$$\begin{aligned} \text{RHS} &= \begin{pmatrix} -\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} - \boldsymbol{P} & : & -\boldsymbol{H}' \\ -\boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\omega}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{X}'\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) - \boldsymbol{P}\tilde{\boldsymbol{\beta}} - \boldsymbol{H}'\tilde{\boldsymbol{\omega}} \\ -\boldsymbol{H}\tilde{\boldsymbol{\beta}} + \boldsymbol{k} \end{pmatrix} \\ &= \begin{pmatrix} -\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) \\ -\boldsymbol{k} \end{pmatrix} \end{aligned}$$

on multiplying out. Equation (47) follows since

$$\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}}) = \boldsymbol{X}'\tilde{\boldsymbol{W}}\left(\boldsymbol{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{W}}^{-1}\boldsymbol{\Lambda}_a^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{\mu}})\right) = \boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}$$

from the definitions of $\boldsymbol{W}$, $\boldsymbol{\Lambda}_a$ and $\boldsymbol{z}$. □

There now follow various formulae for computing $\hat{\boldsymbol{\beta}}$, $\operatorname{Var}\hat{\boldsymbol{\beta}}$, the effective dimension of the model and the standard errors of the fitted values, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Lemma 1: Let $\boldsymbol{X}$, $n \times p$, have rank $p - q$. Let $\boldsymbol{W}$, $n \times n$, be a diagonal matrix with strictly positive diagonal entries, ie, $w_{i,i} > 0$. Let $\boldsymbol{P}$, $p \times p$, be positive semi-definite. Then there exists $\boldsymbol{H}$, $q \times p$, of rank q such that

$$\boldsymbol{X}_{aug} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{H} \end{pmatrix}$$

has rank p. Moreover, if

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{P} & : & \boldsymbol{H}' \\ \boldsymbol{H} & : & \boldsymbol{0} \end{pmatrix} \tag{57}$$

then $\boldsymbol{A}$ is non-singular and

$$\boldsymbol{A}^{-1} = \begin{pmatrix} \boldsymbol{\Psi} & : & \boldsymbol{\Delta}^{-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1} \\ (\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1}\boldsymbol{H}\boldsymbol{\Delta}^{-1} & : & \boldsymbol{I} - (\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1} \end{pmatrix} \tag{58}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X} + \boldsymbol{P} + \boldsymbol{H}'\boldsymbol{H}, \text{ and} \tag{59}$$

$$\boldsymbol{\Psi} = \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1}\boldsymbol{H}\boldsymbol{\Delta}^{-1}. \tag{60}$$

Proof. First, $\boldsymbol{H}$ exists since we can add q linearly independent rows to $\boldsymbol{X}$ and so $\boldsymbol{X}_{aug}$ has rank p.

Second, we show $\boldsymbol{\Delta}$ is positive definite. Let

$$\boldsymbol{X}^*_{aug} = \begin{pmatrix} \boldsymbol{W}^{1/2}\boldsymbol{X} \\ \boldsymbol{H} \end{pmatrix}$$

where $\boldsymbol{W}^{1/2}$ is the diagonal matrix with diagonal entries $w_{i,i}^{1/2}$. We know that $\boldsymbol{W}^{1/2}$ is non-singular and so $\boldsymbol{W}^{1/2}\boldsymbol{X}$, $n \times p$, also has rank $p - q$. Suppose that there exists $\boldsymbol{a} \neq \boldsymbol{0}$ such that $(\boldsymbol{W}^{1/2}\boldsymbol{X})'\boldsymbol{a} = \boldsymbol{h}$ where $\boldsymbol{h}'$ is any row of $\boldsymbol{H}$. This implies that $\boldsymbol{X}'\boldsymbol{b} = \boldsymbol{h}$ for some $\boldsymbol{b} \neq \boldsymbol{0}$. But this contradicts the fact that the rows of $\boldsymbol{H}$ are linearly independent of the rows of $\boldsymbol{X}$. Hence the rows of $\boldsymbol{H}$ are linearly independent of the rows of $\boldsymbol{W}^{1/2}\boldsymbol{X}$ and so $\boldsymbol{X}^*_{aug}$ also has rank p. In particular, $(\boldsymbol{X}^*_{aug})'\boldsymbol{X}^*_{aug}$ is positive definite. Now $\boldsymbol{P}$ is positive semi-definite so

$$(\boldsymbol{X}^*_{aug})'\boldsymbol{X}^*_{aug} + \boldsymbol{P} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{P} + \boldsymbol{H}'\boldsymbol{H} = \boldsymbol{\Delta}$$

is also positive definite.

Next we show $(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1}$ exists. Since $\boldsymbol{H}$ is of full row rank $\boldsymbol{H}'\boldsymbol{a} = \boldsymbol{b}$, $\boldsymbol{b} \neq \boldsymbol{0}$, for any $\boldsymbol{a} \neq \boldsymbol{0}$. Hence $\boldsymbol{a}'\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}'\boldsymbol{a} = \boldsymbol{b}'\boldsymbol{\Delta}^{-1}\boldsymbol{b} > 0$ since $\boldsymbol{\Delta}$ and hence $\boldsymbol{\Delta}^{-1}$ is positive definite and $\boldsymbol{b} \neq \boldsymbol{0}$. Hence $\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}'$ is positive definite and so $(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1}$ exists.

The result (58) follows by checking that $\boldsymbol{A}\boldsymbol{A}^{-1}$ reduces to $\boldsymbol{I}_{p+q}$. □

Corollary 1: Let $\boldsymbol{\Psi}$ be the upper left block in (58). Then the solution of $\hat{\boldsymbol{\beta}}$ in (14) is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Psi}\boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}} + \boldsymbol{\Delta}^{-1}\boldsymbol{H}'(\boldsymbol{H}\boldsymbol{\Delta}^{-1}\boldsymbol{H}')^{-1}\boldsymbol{k} \tag{61}$$

and satisfies $\boldsymbol{H}\hat{\boldsymbol{\beta}} = \boldsymbol{k}$. In the important special case that $\boldsymbol{k} = \boldsymbol{0}$ we have

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Psi}\boldsymbol{X}'\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}. \tag{62}$$

Proof. Multiply the inverse (58) in Lemma 1 onto the RHS in (47) and (61) follows. Further, $\boldsymbol{H}\boldsymbol{\Psi} = \boldsymbol{0}$ from (60) so $\boldsymbol{H}\hat{\boldsymbol{\beta}} = \boldsymbol{k}$ as required. □

Corollary 2: The variance of $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\Psi}$, as defined in (60). If $\boldsymbol{H} = \boldsymbol{0}$ then $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Delta}^{-1}$.

Proof. From (55) and (56) the information matrix of $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\omega}')'$ is $\boldsymbol{A}$ and is given by (57). The variance of $\hat{\boldsymbol{\beta}}$ is given by the upper left block of $\boldsymbol{A}^{-1}$, ie, by $\boldsymbol{\Psi}$. □

Corollary 3: Suppose $\boldsymbol{k} = \boldsymbol{0}$ and let $\mathcal{H} = \boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}'\tilde{\boldsymbol{W}}$ be the hat-matrix. Then the trace of $\mathcal{H}$, denoted $\mathrm{tr}\,(\mathcal{H})$, is $p - q - \mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{P})$. If $\boldsymbol{P} = \boldsymbol{0}$ then $\mathrm{tr}\,(\mathcal{H}) = p - q$, the rank of $\boldsymbol{X}$.

Proof. The trace of the hat-matrix is

$$\begin{aligned}
\mathrm{tr}\,(\mathcal{H}) &= \mathrm{tr}\left\{\boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}'\tilde{\boldsymbol{W}}\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Psi}\boldsymbol{X}'\tilde{\boldsymbol{W}}\boldsymbol{X}\right\} \\
&= \mathrm{tr}\left\{\boldsymbol{\Psi}(\boldsymbol{\Delta} - \boldsymbol{H}'\boldsymbol{H} - \boldsymbol{P})\right\}. \\
\text{Now } \mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Delta}) &= \mathrm{tr}(\boldsymbol{I}_p) - \mathrm{tr}(\boldsymbol{I}_q) = p - q \\
\text{and } \mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{H}'\boldsymbol{H}) &= 0
\end{aligned}$$

and the result follows. □

Corollary 4: The variances of the fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ are given by $\text{diag}\{\boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}'\}$ where $\boldsymbol{\Psi}$ is defined in (60). We can avoid computing the $n \times n$ matrix $\boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}'$, a potentially large matrix, by using

$$\text{diag}\{\boldsymbol{X}\boldsymbol{\Psi}\boldsymbol{X}'\} = [(\boldsymbol{X}\boldsymbol{\Psi}) * \boldsymbol{X}]\,\mathbf{1}_p, \tag{63}$$

where $*$ denotes element-by-element multiplication. □