Annals of Actuarial Science, Vol. 5, part 1, pp. 33–52. © Institute and Faculty of Actuaries 2010
doi:10.1017/S1748499510000047

Smoothing dispersed counts with applications to mortality data

By V. A. B. Djeundje and I. D. Currie

Abstract

Mortality data are often classified by age at death and year of death. This classification results in a heterogeneous risk set and this can cause problems for the estimation and forecasting of mortality. In the modelling of such data, we replace the classical assumption that the numbers of claims follow the Poisson distribution with the weaker assumption that the numbers of claims have a variance proportional to the mean. The constant of proportionality is known as the dispersion parameter and it enables us to allow for heterogeneity; in the case of insurance data the dispersion parameter also allows for the presence of duplicates in a portfolio. We use both the quasi-likelihood and the extended quasi-likelihood to estimate models for the smoothing and forecasting of mortality tables jointly with smooth estimates of the dispersion parameters. We present three main applications of our method: first, we show how taking account of dispersion reduces the volatility of a forecast of a mortality table; second, we smooth mortality data by amounts, ie, when deaths are amounts claimed and exposed-to-risk are sums assured; third, we present a joint model for mortality by lives and by amounts with the property that forecasts by lives and by amounts are consistent. Our methods are illustrated with data from the Continuous Mortality Investigation.

Keywords

Amounts; Duplicates; Forecasting; Heterogeneity; Mortality; Over-dispersion; *P*-splines; Quasi-likelihood; Smoothing

Contact address

I. D. Currie, Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK. E-mail: I.D.Currie@hw.ac.uk

1. Introduction

Modelling and forecasting mortality is a problem of fundamental importance to the actuary. Data for this purpose come largely from two sources: (a) population mortality data available from either the Human Mortality Database (2009) or government offices of statistics, (b) insurance data collected and collated by some central agency. The Continuous Mortality Investigation (CMI) fulfils this latter role in the UK. In both cases, data are generally available at the aggregate level, ie, deaths and exposures are classified by age at death and year of death. This gives rise to two problems for model building. For population data, the risk set for each age and year of death is heterogeneous with respect to mortality since it contains smokers and non-smokers, different social classes, etc. For insurance data, the risk set is subject to an additional source of heterogeneity: 'deaths' are claims on policies, and exposed-to-risk is the number of policy-years lived. Very often, some policyholders have more than one policy and so, for these policyholders, a single death gives rise to multiple claims; this is known as the problem of duplicates. Ideally the data would be deduplicated,

ie, policies held by a single life would be consolidated into a single policy; Richards (2008) describes such a process. Unfortunately, such consolidation is not available for historical data such as collected by the CMI. Further, deduplication does not address the problem of the heterogeneity of mortality across the risk set.

Early work tackling the effect of duplicate policies on the estimates of mortality can be found in Seal (1940), Daw (1946, 1951), Beard & Perks (1949), CMI Committee (1957, 1986), and elsewhere. Forfar *et al.* (1988) allowed for the presence of duplicates by scaling the data by a factor known as the 'variance ratio', a measure of the level of duplicate policies in the data. We refer to this approach as the method of scaling and return to it later in the paper. This method is only available if the variance ratios are known and in any case does nothing to address the more general problem of heterogeneity. In an important paper, Renshaw (1992) showed that it is possible to make proper adjustment for both duplicates and heterogeneity within the modelling process itself.

A frequent question is the following: does the existence of duplicates within a portfolio matter for the estimation of mortality? At first sight, the answer might appear to be no, since the multiple deaths in the numerator are balanced by the additional exposed-to-risk in the denominator. In statistical terms, the estimate of mortality remains unbiased. However, the variance of such an estimate is too small, since it is based on more deaths than we have actually observed (Forfar *et al.*, 1988). There are other more subtle consequences. The results of simulating portfolio experience based on policies will be less volatile than they should be, since, for example, the lives with multiple policies should take all their sums assured with them when they die, and not just part of them. More formally, using the result from Shaked & Shanthikumar (1997) and Bäuerle (1997), who compared random vectors with different levels of multiplicity, Denuit (2000) showed how the presence of duplicates leads to a more dangerous portfolio in the supermodular sense.

A standard assumption is that the number of deaths/claims follows a Poisson distribution; see Brouhns *et al.* (2002), Currie *et al.* (2004) and Cairns *et al.* (2009) for example. The Poisson distribution has equal mean and variance but heterogeneity in general and the presence of duplicates in particular will inflate the variance. Such inflation is known as over-dispersion (McCullagh & Nelder, 1989; Renshaw, 1992). Using a Bayesian argument, Li *et al.* (2009) assumed that the force of mortality follows a prior gamma distribution within each age/year cell from which it follows that the number of deaths/claims has a negative binomial distribution with variance larger than the mean, as required. Other approaches to the problem of over-dispersion can be found in Williams (1982), Breslaw & Clayton (1993), Hinde & Demetrio (1998), for example. In this paper, we adopt the two-stage joint-modelling of mean and dispersion through the extended quasi-likelihood, as described in McCullagh & Nelder (1989, chap 10). Our reason for adopting the quasi-likelihood approach (as opposed to the negative binomial approach of Li *et al.*, 2009) is that it allows us to remain within the exponential family, and, as a result, we can essentially adopt a generalized linear model (GLM) approach to model building. Renshaw (1992) also used this approach in his paper on the graduation of mortality data in the presence of duplicates. Our contribution is (a) to extend this work to the smoothing and forecasting of 2-dimensional mortality tables, (b) to apply our method to the modelling and forecasting of mortality data by amounts, and (c) to produce a joint model of mortality by lives and amounts with the property that the forecasts by lives and amounts are constrained to be parallel.

The plan of the paper is as follows: in section 2 we define our notation and describe the joint estimation of regression and dispersion parameters with extended quasi-likelihood; in section 3

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

we show how extended quasi-likelihood applies to *P*-splines, the smoothing method of Eilers & Marx (1996); in section 4 we describe a technique for the adjustment for bias of the estimates of the dispersion parameters. Applications of these methods are described in section 5: (a) we perform a simulation exercise to demonstrate first the effect that over-dispersion has on the smoothing process and second that our methods enable the true underlying smooth curve to be recovered more appropriately; (b) we describe the modelling and forecasting of mortality tables first by lives and then by amounts; (c) we discuss the consistent forecasting of mortality by both lives and amounts. The paper ends with a short conclusion.

2. Model specification and estimation with quasi-likelihoods

Many models for mortality data are based on the Poisson distribution (Brouhns *et al*. 2002; Currie *et al*. 2004; Cairns *et al*. 2009). This strong assumption fails to account for the over-dispersion that is generally found in mortality data. The alternative assumption that the deaths follow the negative binomial distribution, as used by Li *et al*. (2009), takes us outside the exponential family into a less friendly environment for model building. We now describe estimation with extended quasi-likelihood (McCullagh & Nelder, 1989, chap 10) which allows over-dispersion to be modelled within the exponential family framework.

We suppose that we have independent observations $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. In general the Y_i can be counts but we will restrict our description to the case when they are counts of deaths or claims. To simplify the notation, we will not distinguish between random variables and their observations. We supposed that these data can be partitioned into $\mathcal{K}$ classes $\mathcal{C}_1, \ldots, \mathcal{C}_\mathcal{K}, \mathcal{K} \leq n$, in such a way that each class is homogeneous (by homogeneous we mean that the level of dispersion within each class can be assumed constant). For example, at one extreme if $\mathcal{K} = n$ then each count has its own dispersion parameter while at the other extreme if $\mathcal{K} = 1$ then a single dispersion parameter applies to all counts. We will be particularly interested in the intermediate case where the dispersion parameter is age dependent in a mortality table; this is consistent with the approach of Forfar *et al*. (1988), Renshaw (1992) and Li *et al*. (2009) who use age-dependent dispersion parameters. We will denote by $\phi_{\mathcal{C}_k}$ the over-dispersion parameter in the class $\mathcal{C}_k$; we note in passing that $\phi_{\mathcal{C}_k}$ could in theory be less than one, so this approach can also deal with under-dispersion (which the negative binomial cannot). Let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_\mathcal{K}\}$ denote the set of classes and let $\varphi : \{1, \ldots, n\} \to \mathcal{C}$ assign observations to classes.

We include dispersion in the model through the first and second moment assumptions

$$\mathbb{E}[Y_i] = \mu_i, \quad var(Y_i) = \phi_{\varphi(i)} \times \upsilon(\mu_i), \quad i = 1, \ldots, n, \quad \boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{B\alpha}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, $\boldsymbol{B}$ is the regression matrix, $\boldsymbol{\alpha}$ is the unknown vector of coefficients, $\upsilon(\cdot)$ is the variance function, $g(\cdot)$ is the link function, $g(\boldsymbol{\mu}) = (g(\mu_1), \ldots, g(\mu_n))'$, and $\boldsymbol{\eta}$ is the linear predictor. The case of a Poisson GLM is given by $\phi_{\varphi(i)} = 1, \forall i$, and $\upsilon(\cdot)$ equal to the identity function. Further, in the Poisson case we can fit the model with maximum likelihood; this requires the distribution that has generated the data. Unfortunately, such a distribution is not available for (2.1). An alternative is the quasi-likelihood framework of Wedderburn (1974), an extension of the familiar likelihood function that allows estimation to take place in more general settings such as (2.1). Under model (2.1), the *quasi-likelihood* (or more correctly the *quasi-log-likelihood*) of a single observation Y_i is defined as

$$\mathcal{Q}(\mu_i; Y_i) \equiv \mathcal{Q}(\boldsymbol{\alpha}; Y_i) \equiv \mathcal{Q}(\boldsymbol{\eta}; Y_i) = \frac{1}{\phi_{\varphi(i)}} \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\upsilon(t)} dt = -\frac{1}{2\phi_{\varphi(i)}} d_i \tag{2.2}$$

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

where

$$d_i = -2\int_{Y_i}^{\mu_i} \frac{Y_i - t}{\upsilon(t)}\,dt \tag{2.3}$$

is the *deviance component*. The estimates of the dispersion parameters $\phi_{\varphi(i)}$ are based on the d_i. (The normal distribution is a well-known example here since when $\upsilon(t) = 1$ we have $d_i = (Y_i - \mu_i)^2$, the ith component of the residual sum of squares.) The quasi-likelihood of the sample $\boldsymbol{Y}$ is

$$\mathcal{Q}(\boldsymbol{\mu}; \boldsymbol{Y}) \equiv \mathcal{Q}(\boldsymbol{\alpha}; \boldsymbol{Y}) \equiv \mathcal{Q}(\boldsymbol{\eta}; \boldsymbol{Y}) = \sum_{i=1}^{n} \mathcal{Q}(\mu_i; Y_i). \tag{2.4}$$

If the dispersion parameters are known, the fitting of model (2.1) is reduced to the optimization of the quasi-likelihood (2.4). However, since these parameters are generally unknown they also need to be estimated. Thus we also need the derivative of $\mathcal{Q}$ to behave like a log-likelihood with respect to the dispersion parameters, that is $\mathbb{E}\left[\partial\mathcal{Q}/\partial\phi_u\right] = 0$ for all $u \in \mathcal{C}$. For this to be achieved, the quasi-likelihood is usually adjusted (see Nelder & Pregibon, 1987) to the so-called *extended quasi-likelihood* $\mathcal{Q}^+$ as follows:

$$\mathcal{Q}^+(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{Y}) = \mathcal{Q}(\boldsymbol{\alpha}; \boldsymbol{Y}) + f(\boldsymbol{\phi}),\ \boldsymbol{\phi} = (\phi_{\mathcal{C}_1}, \ldots, \phi_{\mathcal{C}_\mathcal{K}})', \tag{2.5}$$

where $f(\cdot)$ is some well chosen function. A simple candidate (that we use here) for $f(\boldsymbol{\phi})$ is $-\frac{1}{2}\sum \log\left(2\pi\phi_i w(Y_i)\right)$ where $w(\cdot)$ is any positive function.

If we set

$$d_u = \frac{\sum_{i\in\varphi^{-1}(u)} d_i}{n_u},\ \text{where } n_u = |\varphi^{-1}(u)|,\ \forall u \in \mathcal{C}, \tag{2.6}$$

then, at the true value of $\boldsymbol{\mu}$ (McCullagh & Nelder, 1989, chap. 10),

$$\mathbb{E}[d_u] \simeq \phi_u,\ \forall u \in \mathcal{C}. \tag{2.7}$$

We make two comments on (2.6). First, there is a possible confusion of notation; we have adopted the convention that the suffix i, $i = 1, \ldots, n$, refers to observations, while the suffix $u, u \in \mathcal{C}$, refers to classes. Second, with the normal distribution, d_u reduces to the familiar maximum likelihood estimate of σ^2 in the class $\mathcal{C}_u$.

Now, corresponding to the model (2.1) for the mean $\boldsymbol{\mu}$, we model the dispersion parameters (for sufficiently large $\mathcal{K}$) with

$$h(\boldsymbol{\phi}) = \check{\boldsymbol{B}}\boldsymbol{\beta} \tag{2.8}$$

for some suitable link function $h(\cdot)$ which we specify below. Within this setting, fitting model (2.1) is reduced to the optimization of the extended quasi-likelihood (2.5) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. This optimization yields the inter-dependent equations

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_{\varphi(i)}\upsilon(\mu_i)} \frac{\partial\mu_i}{\partial\alpha_j} = 0,\ \ j = 1, \ldots, c, \tag{2.9}$$

$$\sum_{u\in\mathcal{C}} \frac{n_u(d_u - \phi_u)}{\phi_u^2} \frac{\partial\phi_u}{\partial\beta_j} = 0,\ \ j = 1, \ldots, \check{c}, \tag{2.10}$$

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

where c and $\check{c}$ are the lengths of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively. Equations (2.10) correspond to the quasi-likelihood estimating equations based on independent responses d_u with $\mathbb{E}[d_u] = \phi_u$ and $var(d_u) = \phi_u^2/n_u$. In the GLM setting, equations (2.10) are identical to the estimating equations based on gamma responses d_u, with shape parameter n_u and scale parameter ϕ_u/n_u. The canonical link for the gamma distribution is the negative inverse function (McCullagh & Nelder, 1989, chap. 2), so we simplify (2.8) by specifying

$$d_u \sim Gamma\left(n_u, \frac{\phi_u}{n_u}\right), \quad u \in \mathcal{C}, \quad h(\boldsymbol{\phi}) = -\frac{1}{\boldsymbol{\phi}} = \check{\boldsymbol{B}}\boldsymbol{\beta}, \tag{2.11}$$

where the quotient sign is interpreted as element-by-element division. In the parametric setting, we have generalized and unified the two-stage joint modelling of mean and dispersion described in McCullagh & Nelder (1989, chap. 10), and used by Renshaw (1992) for graduation in life insurance. However, mortality data often reveal complex patterns which suggest that a smoothing rather than a parametric approach is more appropriate. In the next section, we extend the above results to the P-spline method of Eilers & Marx (1996).

3. Extended quasi-log-likelihood and P-splines

We use the method of P-splines (Eilers & Marx, 1996) and so the regression matrices $\boldsymbol{B}$ and $\check{\boldsymbol{B}}$ are constructed from B-spline bases on the covariates, age at death and year of death. For a description of P-splines from a statistical perspective as applied to modelling mortality see Currie *et al.* (2004); Richards *et al.* (2006) and Richards & Currie (2009) contain descriptions of P-splines in an actuarial context. The key idea is to overfit the data with rich bases of B-splines, and then apply roughness penalties to the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to achieve smoothness; the roughness penalty on $\boldsymbol{\beta}$ here assumes that the number of classes, $|\mathcal{C}|$, is not too small. Combining the penalization with the extended quasi-likelihood (2.5), we derive an optimal criterion, the *penalized extended quasi-likelihood*,

$$\mathcal{Q}_P^+(\boldsymbol{\mu};\boldsymbol{\phi}) = \mathcal{Q}^+(\boldsymbol{\mu};\boldsymbol{\phi}) \; - \; \frac{1}{2}\left(\boldsymbol{\alpha}'\boldsymbol{P}_{\lambda_1}\boldsymbol{\alpha} + \boldsymbol{\beta}'\boldsymbol{P}_{\lambda_2}\boldsymbol{\beta}\right). \tag{3.12}$$

In (3.12), $\boldsymbol{P}_{\lambda_1} = \boldsymbol{P}(\boldsymbol{\lambda}_1, \boldsymbol{\Delta}_1)$ and $\boldsymbol{P}_{\lambda_2} = \boldsymbol{P}(\boldsymbol{\lambda}_2, \boldsymbol{\Delta}_2)$ are penalty matrices acting on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively; here $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ are difference matrices (generally of order 2), and $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are vectors of smoothing parameters. The dimension of $\boldsymbol{\lambda}_1$ depends of the structure of the data $\boldsymbol{Y}$ and the model matrix $\boldsymbol{B}$ while that of $\boldsymbol{\lambda}_2$ is a function of the structure of the C_k's and the model matrix $\check{\boldsymbol{B}}$.

Optimizing $\mathcal{Q}_P^+$ with respect to $\boldsymbol{\alpha}$ yields the penalized iterative equation

$$\left(\boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{B} + \boldsymbol{P}_{\lambda_1}\right)\hat{\boldsymbol{\alpha}} \; = \; \boldsymbol{B}'\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}\tilde{\boldsymbol{z}}, \tag{3.13}$$

where $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$ represents the diagonal weight matrix in the quasi-likelihood model (2.1) based on the response $\boldsymbol{Y}$, and $\tilde{\boldsymbol{z}}$ is the associated working variable; here, a tilde refers to the current estimates, and a hat refers to the updates. This form is similar to the penalized equation encountered in the penalized GLM setting, the difference being that the dispersion parameters are involved in the smoothing process through the weight matrix $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$, which is a function of the dispersion parameters. Similarly, optimizing $\mathcal{Q}_P^+$ with respect to $\boldsymbol{\beta}$ yields the penalized iterative equation

$$\left(\check{\boldsymbol{B}}'\tilde{\boldsymbol{W}}_d\check{\boldsymbol{B}} + \boldsymbol{P}_{\lambda_2}\right)\hat{\boldsymbol{\beta}} = \check{\boldsymbol{B}}'\tilde{\boldsymbol{W}}_d\tilde{\boldsymbol{z}}_d, \tag{3.14}$$

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

where $\tilde{\boldsymbol{W}}_d$ is the diagonal weight matrix in the GLM based on a gamma response $\boldsymbol{d} = (d_{\mathcal{C}_1}, \ldots, d_{\mathcal{C}_\mathcal{K}})'$, and $\tilde{\boldsymbol{z}}_d$ is the corresponding working variable.

We note that equations (3.13) and (3.14) are the penalized versions of the scoring equations corresponding to (2.9) and (2.10) but written in matrix form. The precise form of the weight functions $\tilde{\boldsymbol{W}}_{\boldsymbol{\phi}}$ and $\tilde{\boldsymbol{W}}_d$ depends on the form of the link functions $g(\cdot)$ and $h(\cdot)$.

For given values of the smoothing parameters $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, the estimation process consists of iterating between (3.13) (the $\boldsymbol{\alpha}$-step) and (3.14) (the $\boldsymbol{\beta}$-step) until convergence is achieved. For the estimation of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, we step outside the likelihood framework and use a model selection criterion. One of the best known criteria is the Bayesian Information Criterion (BIC) (Schwarz, 1978) which attempts to balance (a) fit as measured by the deviance with (b) model complexity as measured by the effective dimension. Under the Poisson assumption, the BIC is given by:

$$\text{BIC} = D \; + \; \log(n) \times v \tag{3.15}$$

where $D = \sum \hat{d}_i$ is the residual deviance, and v is the effective dimension of the model. For count data such as Poisson, (3.15) is appropriate when the value of the dispersion is close to 1; however, if the data are over(under)-dispersed, the deviance will tend to be large(small), with the result that the deviance will also tend to be over(under)-weighted in (3.15). This implies that the effective dimension will also tend to be large(small); we end up by under(over)-smoothing our data. We correct this inappropriate weighting by adjusting the deviance in each class; this gives the scaled BIC:

$$\text{BICs} = \sum_{i=1}^{n} \frac{\hat{d}_i}{\phi_{\varphi(i)}} \; + \; \log(n) \times v, \tag{3.16}$$

a generalization of the scaled criterion given by Heuer (1997). Clearly, if there is no over(under)-dispersion in the data, then BIC and BICs are equivalent. Both BIC and BICs require a value for the dimension of the fitted model. If the smoothing parameters are zero, then penalized regression reduces to ordinary regression and the dimension of the model is the number of linearly independent columns in the regression matrix; in our case, if $\boldsymbol{B}$ has c columns then the dimension of the model is c. With penalization the flexibility of the model is reduced and so the dimension of the model is correspondingly reduced. Following Ye (1998) and Ruppert *et al.* (2003), we approximate the effective dimension v by

$$v = tr\left(\frac{\partial\, g(\hat{\boldsymbol{\mu}})}{\partial \hat{z}}\right) = tr(\boldsymbol{H}) \tag{3.17}$$

where the *hat-matrix* $\boldsymbol{H}$, which maps the working variable $\hat{z}$ in (3.13) to the fitted value of the linear predictor at convergence, is given by $\boldsymbol{B}(\boldsymbol{B}'\hat{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{B} + \boldsymbol{P}_{\boldsymbol{\lambda}_1})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}_{\boldsymbol{\phi}}$. Hence,

$$v = tr[(\boldsymbol{B}'\hat{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{B} + \boldsymbol{P}_{\boldsymbol{\lambda}_1})^{-1}\boldsymbol{B}'\hat{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{B}] \tag{3.18}$$

$$= c \; - \; tr[(\boldsymbol{B}'\hat{\boldsymbol{W}}_{\boldsymbol{\phi}}\boldsymbol{B} + \boldsymbol{P}_{\boldsymbol{\lambda}_1})^{-1}\boldsymbol{P}_{\boldsymbol{\lambda}_1}]. \tag{3.19}$$

The second form (3.19) shows the reduction in the dimension of the model brought about by the penalization. The complete estimation algorithm is as follows:

(i) initialize $\phi_u = 1$, $\forall u \in \mathcal{C}$,

(ii) update $\boldsymbol{\mu}$ by solving (3.13) in $\boldsymbol{\alpha}$ with $\boldsymbol{\lambda}_1$ selected by minimizing (3.16),

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

(iii) - if $|\mathcal{C}|$ is small, update the ϕ_u to their extended quasi-likelihood estimates given by (2.6),
- else, update $\boldsymbol{\phi}$ by solving (3.14) in $\boldsymbol{\beta}$ with $\boldsymbol{\lambda}_2$ selected by minimizing (3.16),

(iv) repeat (ii) and (iii) until convergence is achieved.

In our applications, we will refer to this algorithm as the *full extended quasi-likelihood scheme*.

4. Bias adjustment

We have already remarked after (2.6) that d_u reduces to the maximum likelihood estimate of σ^2 in the normal distribution case. This estimate is biased downward and, in the same way, the maximum extended quasi-likelihood estimate of the dispersion parameters also tends to be biased downward, (see Figure 4). This stems from the fact that (2.7) holds only at the true value of $\boldsymbol{\mu}$ while $\boldsymbol{\mu}$ is generally unknown. An alternative approach is to estimate $\boldsymbol{\mu}$ by maximizing criterion (3.12) as before, ie, by solving the iterative equation (3.13), but to look for a different estimate for $\boldsymbol{\phi}$. A potential candidate (analogous to the unbiased estimate of σ^2 in standard normal regression) is the *bias corrected mean Pearson statistic* in each class:

$$d_u^* = \frac{1}{n_u - v_u} \sum_{i \in \varphi^{-1}(u)} \frac{\left(Y_i - \hat{\mu}_i\right)^2}{v(\hat{\mu}_i)}, \quad u \in \mathcal{C}, \tag{4.20}$$

where v_u is the contribution of the class u to the total dimension v. Intuitively, from (3.17), we estimate the v_u by

$$v_u = \sum_{i \in \varphi^{-1}(u)} \frac{\partial g(\hat{\mu}_i)}{\partial \hat{z}_i} = \sum_{i \in \varphi^{-1}(u)} H_{ii}, \quad u \in \mathcal{C}. \tag{4.21}$$

If the number of classes, $\mathcal{K}$, is small then ϕ_u is estimated by d_u^*; otherwise, we proceed as follows. Instead of relying on the (penalized) extended quasi-likelihood of model (2.1) to estimate $\boldsymbol{\phi}$, we assume a full quasi-likelihood framework for the 'observations' $\boldsymbol{d}^* = (d_{\mathcal{C}_1}^*, \ldots, d_{\mathcal{C}_\mathcal{K}}^*)'$:

$$\mathbb{E}\left[d_u^*\right] = \phi_u, \quad var(d_u^*) = \tau \times v^*(\phi_u), \forall u \in \mathcal{C}, \quad h(\boldsymbol{\phi}) = -\frac{1}{\boldsymbol{\phi}} = \check{\boldsymbol{B}}\boldsymbol{\beta}. \tag{4.22}$$

We then denote by Q_1 and Q_2 the quasi-likelihood of models (2.1) and (4.22) respectively, and by BICs1 and BICs2 the associated scaled BIC. Conditional on $\boldsymbol{\phi}$, we penalize Q_1 in $\boldsymbol{\alpha}$ to get Q_{1_p}, and conditional on $\boldsymbol{\alpha}$, we penalize Q_2 in $\boldsymbol{\beta}$ to get Q_{2_p}. The estimation algorithm becomes:

(i) initialize $\phi_u = 1$, $\forall u \in \mathcal{C}$,

(ii) update $\boldsymbol{\mu}$ by optimizing Q_{1_p} in $\boldsymbol{\alpha}$ with $\boldsymbol{\lambda}_1$ selected by minimizing BICs1,

(iii) - if $|\mathcal{C}|$ is small, update the components of $\boldsymbol{\phi}$ to their Pearson estimates given by (4.20),
- else, update $\boldsymbol{\phi}$ by optimizing Q_{2_p} in $\boldsymbol{\beta}$ with $\boldsymbol{\lambda}_2$ selected by BICs2,

(iv) repeat (ii) and (iii) until convergence is achieved.

In our applications, we will refer to this algorithm as the *bias corrected scheme*.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

5. Applications

We apply our methods to two CMI data sets, the male assured lives data set where age at death runs from 40 to 90 and year of death from 1950 to 2006, and the male pensioner data where age at death runs from 10 to 108 and year of death from 1983 to 2006; the pensioner data are available both by lives and amounts. To be consistent with the actuarial literature, from now on $\boldsymbol{\mu}$ will represent a force of mortality instead of a mean, as in the previous sections. The applications we consider assume an over-dispersed Poisson model with a systematic structure (as the case may be) for the dispersion parameters. In section 5.1, through a simulation, we illustrate how over-dispersion affects the smoothing process and how the use of the two schemes presented in the previous section leads to improved estimates. In section 5.2, we use both schemes to fit the 2-dimensional mortality surface and we illustrate the effect of over-dispersion on the extrapolated trends and confidence bands. In section 5.3, we show that both schemes can be used directly (without the need of prior scaling, as in Forfar *et al.*, 1988) to handle the high level of over-dispersion such as encountered in mortality data by amounts. Lastly, in section 5.4 we consider the joint modelling and forecasting of mortality by lives and amounts; here, we will be concerned to produce consistent forecasts from the two data sets.

5.1. A simulation exercise

We conduct a simulation exercise with two aims: first, to illustrate how dispersion affects the smoothing process and second, to show how the use of the bias corrected scheme (as well as the full extended quasi-likelihood scheme) gives rise to an improved estimate of the true mortality curve. The simulation exercise will be split into two parts: first, a portfolio without duplicates, and second, one with duplicates.

5.1.1. A simulation exercise without duplicates

Figure 1 shows log mortality for years 1950 to 2005 for a male aged 75 from the CMI assured lives data set. For the purpose of these simulation exercises, we suppose that underlying log mortality follows the fitted quadratic curve shown, ie, $\log \mu_t = Q(t)$ where t is year. We now suppose that we

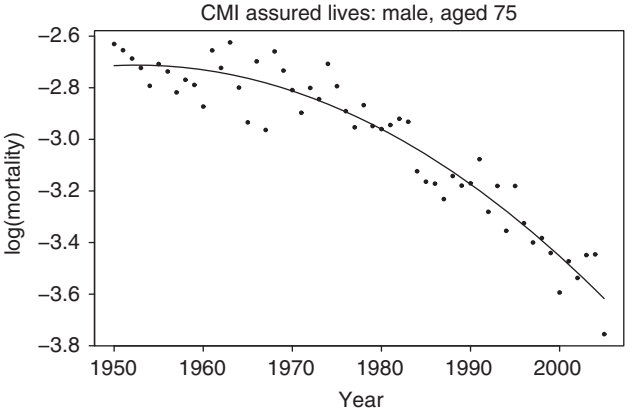

Figure 1. Observed mortality (●●●) for CMI assured lives, males age 75, together with fitted quadratic curve.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

have central exposure $E_t^c = 1000$ at each year t and suppose that the number of deaths (claims) come from the Poisson distribution: $D_t \sim \mathcal{P}\big(E_t^c \exp[Q(t)]\big)$. We simulate from this model and estimate the underlying mortality curve using P-splines with a cubic B-spline basis with $c = 13$ B-splines in the basis, second-order penalty and smoothing parameter chosen by minimizing BIC. This exercise is repeated 2000 times.

With Poisson errors we have $\phi = 1$. For each simulation $i, i = 1, \dots, 2000$, we compute the mean square error

$$\text{MSE} = \frac{1}{n}\sum_1^n (\log \hat{\mu}_t - \log \mu_t)^2 = \frac{1}{n}\sum_1^n (\log \hat{\mu}_t - Q(t))^2, \tag{5.23}$$

an overall measure of the quality of the fit, and we also compute the bias corrected Pearson estimate $\hat{\phi}_i$ of ϕ using (4.20). The mean of the MSEs (over the 2000 simulations) was 0.00168.

We now perform a second round of smoothing for each of our 2000 simulations; for the ith simulation, we set $\phi = \hat{\phi}_i$ and re-estimate the force of mortality with the penalized iterative equation (3.13) and select the smoothing parameter with the scaled BIC defined in (3.16). The mean of the MSEs was very little changed at 0.00170.

In conclusion, since the quasi-likelihood generalizes the usual likelihood approach, the MSEs obtained with the two approaches are essentially equal, even when the Poisson assumption does hold. In the next section we discuss the situation when the presence of duplicates systematically introduces over-dispersion into the problem. Here we will see a much stronger effect of over-dispersion, and both the bias corrected and the full extended quasi-likelihood schemes outperform the likelihood approach.

5.1.2. A simulation exercise with duplicates

In the previous section we considered a portfolio of 1000 distinct policyholders in each year, where each policyholder was exposed to risk for one year. Now we consider males aged 75 again, but we suppose that we have a portfolio of 1000 policies in each year made up as follows: we have 200 policyholders with a single policy, 150 policyholders with two policies, 100 policyholders with three policies and 50 policyholders with four policies. Hence, we have 500 distinct policyholders (classified into four categories) with a total of 1000 policies, an average of two policies per policyholder. If C_t is the number of claims observed in year t, we have $\mathbb{E}[C_t] = 1000\mu_t$ and $var(C_t) = 2500\mu_t$; thus the (theoretical) dispersion parameter is $\phi = 2.5$ in each year, ie, the variance of claim numbers has been inflated by a factor of 2.5. Finally we suppose that a policyholder in year t is subject to the same (quadratic) mortality as in the previous section and we repeat the previous simulation exercise for each category of policyholder.

The black colour in Figure 2 shows the MSEs for each simulation based on the penalized likelihood with Poisson errors. The mean of the MSEs was 0.00843 (compared to 0.00168 without duplicates); the presence of over-dispersion has had a negative impact on the smoothing process. We perform a second round of smoothing with the estimated values of the ϕ's incorporated into the estimation. The red colour in Figure 2 shows the MSEs after this second round of smoothing; the mean of the MSEs was 0.00428, a drop of almost 49%. The mean of the estimated $\hat{\phi}$'s (over the 2000 simulations) was 2.4; this is in agreement with the (theoretical) value $\phi = 2.5$.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

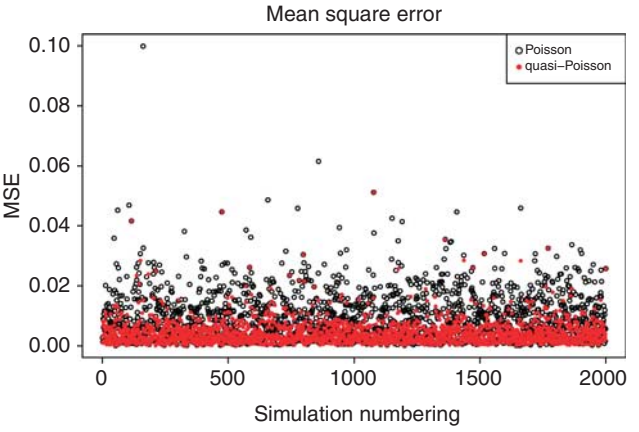

Figure 2. Mean square error, MSE, with duplicates and $\phi = 1$ (○○○). Mean square error, MSE, with duplicates and ϕ set to $\hat{\phi}$ (●●●).

The effective dimension of the fitted model gives another perspective on the effect of over-dispersion. Ignoring over-dispersion gave a mean effective dimension of 7, well in excess of 3, the true dimension of the model, while including over-dispersion reduced the mean effective dimension to 3.44. In general (by examining (3.13) and (3.16)), the flexibility of the fitted curve is reduced by the inclusion of over-dispersion parameters into the estimation process. This has important consequences for forecasting where less volatile curves lead to more stable forecasts.

5.2. Modelling and forecasting over-dispersed mortality tables

Here we consider the problem of modelling and forecasting a mortality table. For illustration, we use the CMI assured lives data with ages 40 to 90 and years 1950 to 2006.

5.2.1. Model specification

We use 2-dimensional *P*-splines. Details of this method from an actuarial perspective can be found in Richards *et al.* (2006) but, in brief, it consists of supposing at age x in year t that

$$\text{Model 1}: \quad D_{x,t} \sim \mathcal{P}(E_{x,t} \times \mu_{x,t}), \quad \log(\boldsymbol{\mu}) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\alpha} \tag{5.24}$$

where $\boldsymbol{B}_x$ and $\boldsymbol{B}_t$ are 1-dimensional cubic *B*-spline bases along age and year respectively, and $\otimes$ is the Kronecker product; here $D_{x,t}$ is the number of deaths/claims and $E_{x,t}$ is the central exposed-to-risk at age x in year t. Smoothness is then achieved by penalizing the regression coefficients separately in the age and year directions, that is, the vector of smoothing parameters $\boldsymbol{\lambda}_1$ (see equation (3.12)) is a 2-dimensional vector: $\boldsymbol{\lambda}_1 = (\lambda_{1,x}, \lambda_{1,t})'$, where $\lambda_{1,x}$ and $\lambda_{1,t}$ quantify the amount of smoothing in the age and year directions respectively. The penalization not only gives a smooth mortality surface but also allows forecasting to take place as described by Currie *et al.* (2004).

Here we are interested in incorporating over-dispersion into the 2-dimensional smoothing process in the same fashion as the 1-dimensional case just considered; in other words, we are interested in the effects of replacing the Poisson assumption with a first and second moment assumption, as in (2.1). A starting model is to assume an over-dispersed Poisson model with a common over-dispersion

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

parameter for all observations:

$$\text{Model 2}: \quad \mathbb{E}\left[D_{x,t}\right] = E_{x,t} \times \mu_{x,t}, \quad var(D_{x,t}) = \phi \times \mathbb{E}\left[D_{x,t}\right], \quad \log(\boldsymbol{\mu}) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\alpha}. \tag{5.25}$$

Note that (5.25) is a special case of model (2.1), in which all the observations are assumed to belong to the same class; the variance function is the identity. The structure of the over-dispersion here is very simple, but it is useful for understanding the effect of the dispersion parameters on the smoothing process. A refinement of Model 2 is to allow the dispersion to be age dependent, that is

$$\text{Model 3}: \quad \mathbb{E}\left[D_{x,t}\right] = E_{x,t} \times \mu_{x,t}, \quad var(D_{x,t}) = \phi_x \times \mathbb{E}\left[D_{x,t}\right], \quad \log(\boldsymbol{\mu}) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\alpha}; \tag{5.26}$$

once again, (5.26) is a special case of model (2.1), where the dispersion classes $\mathcal{C}$ comprise the observations of the same age, and the variance function is the identity. Renshaw (1992) presented an age dependant dispersion parameter model for graduation in life insurance. We extend Renshaw's work to two dimensions with the possibility of a general smooth structure for the dispersion parameters (as suggested by Renshaw); furthermore we consider forecasting and discuss the effect of over-dispersion on the associated confidence bands.

5.2.2. Estimation

For models 1, 2 and 3, the space between knots was five years for both age and year, and the smoothing parameters were chosen by minimizing the scaled BIC. In each case, the generalized linear array representation (Currie *et al.* 2006) was used to speed up the computation. Forecasting to 2050 was performed via the penalization by extending the basis in the year direction to 2050, as described in Currie *et al.* (2004).

Model 1 was fitted as described in Currie *et al.* (2004), that is by the penalized GLM for $\boldsymbol{D} = (D_{x,t})$ with regression matrix $\boldsymbol{B}$, offset $\log(\boldsymbol{E})$, $\boldsymbol{E} = (E_{x,t})$, log link, Poisson error and penalty matrix $\boldsymbol{P}_{\lambda_1}$, where

$$\boldsymbol{B} = \boldsymbol{B}_t \otimes \boldsymbol{B}_x \quad \text{and} \quad \boldsymbol{P}_{\boldsymbol{\lambda}_1} = \boldsymbol{\lambda}_{1,x}\boldsymbol{I}_{c_t} \otimes \boldsymbol{\Delta}'_x\boldsymbol{\Delta}_x + \lambda_{1,t}\boldsymbol{\Delta}'_t\boldsymbol{\Delta}_t \otimes \boldsymbol{I}_{c_x}. \tag{5.27}$$

In (5.27), $\boldsymbol{\Delta}_x$ and $\boldsymbol{\Delta}_t$ are difference matrices of order 2 (Eilers & Marx, 1996) in the age and year direction, c_x and c_t are the number of B-splines in the age and year direction (that is the number of columns in $\boldsymbol{B}_x$ and $\boldsymbol{B}_t$), and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Models 2 and 3 were each fitted with both the full extended quasi-likelihood and the bias corrected schemes. In both schemes we applied the penalty matrix (5.27) to the coefficient $\boldsymbol{\alpha}$ to achieve smoothness. Model 2 (with a single dispersion parameter) does not require second stage smoothing; we simply update the value of ϕ either to its extended quasi-likelihood estimate or to its Pearson estimate, as the case may be. In contrast, for Model 3, a second stage smoothing process was implemented to get smooth estimates of the dispersion parameters. This second stage modelling process is easier in the full extended quasi-likelihood scheme in comparison with the bias corrected scheme, since the full extended quasi-likelihood scheme uses the gamma distribution (2.11) with a known shape parameter, whereas the bias corrected scheme assumes a full extended quasi-likelihood scheme (with an unknown dispersion parameter that needs to be estimated) based on the bias corrected mean Pearson statistics (4.20); in the bias corrected scheme the variance function, $v^*(\phi_u)$, in (4.22) is set to ϕ_u^2/n_u. In both cases, we have a 1-dimensional smoothing process in the

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

age direction, and so $\boldsymbol{\lambda}_2$ is reduced to a scalar, λ_2. Hence, we set the regression matrix $\check{\boldsymbol{B}}$ and the penalty matrix $\boldsymbol{P}_{\boldsymbol{\lambda}_2}$ to

$$\check{\boldsymbol{B}} = \boldsymbol{B}_x \quad \text{and} \quad \boldsymbol{P}_{\boldsymbol{\lambda}_2} = \lambda_2 \boldsymbol{\Delta}'_x \boldsymbol{\Delta}_x. \tag{5.28}$$

For both schemes, we have used the gamma canonical link function (that is the negative inverse) in the smoothing of the dispersion parameters. Wherever needed, the convergence criterion was the relative error between the current estimate and its update with a tolerance of 10^{-5}.

5.2.3. Results and comments

The estimated forces of mortality obtained with the full extended quasi-likelihood scheme can scarcely be distinguished by eye from those obtained with the bias corrected scheme, since the difference between the estimated over-dispersion parameters from both schemes is not substantial; see Figure 4 and Table 1. For precision however, all the graphics and most of the estimates discussed in this section are those implemented with the bias corrected scheme.

Figure 3 shows the profile views for ages 45 and 70 which result from fitting Models 1 and 2; some statistics are given in Table 1. Figure 4 shows the raw values and the smooth estimates of the ϕ_u's under Model 3. This graphic shows how the full extended quasi-likelihood scheme under-estimates the dispersion parameters compared to the bias corrected scheme.

We make some comments on the results in Table 1. First, as measured by BICs, Model 2 gives a much superior fit to the data compared to Model 1, with Model 3 a further improvement. Second, the less flexible the fitted model, the larger the deviance; however, the deviance in Model 1 is computed under the assumption that $\phi = 1$, and the relative increase in deviance from 4968 to 5076 as we go from Model 1 to Model 2 is more than compensated for by the additional variance of Model 2 (as measured by its assumed over-dispersion parameter $\hat{\phi} = 1.76$). Similar remarks apply to Model 3 compared to Model 2. Third, Figure 3 does not include the output from Model 3; the plotted lines for Model 3 are very close to those shown for Model 2, and we have chosen to omit them; however, the difference between the two fitted models is clearly seen in Table 1.

There are two important conclusions to be drawn from this example: the first concerns the central forecasts and the second the width of the confidence intervals. First, we consider the central forecasts. The effective dimension of the model under the Poisson assumption is 46. If we include

Table 1. Comparative statistics for Models 1, 2 and 3. FEQS and BCS stand for the full extended quasi-likelihood and the bias corrected schemes respectively.

	Model 1	Model 2		Model 3	
		FEQS	BCS	FEQS	BCS
$\boldsymbol{\lambda}_1$	(228, 676)	(277, 926)	(278, 935)	(295, 1077)	(296, 1091)
$\boldsymbol{\phi}$	1	1.74	1.76	see Figure 4	see Figure 4
$tr(\boldsymbol{H})$	46	36	36	33	33
Deviance	4968	5073	5076	5118	5121
BICs	5333	3192	3158	2828	2796
Iterations	1	7	7	6	6

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

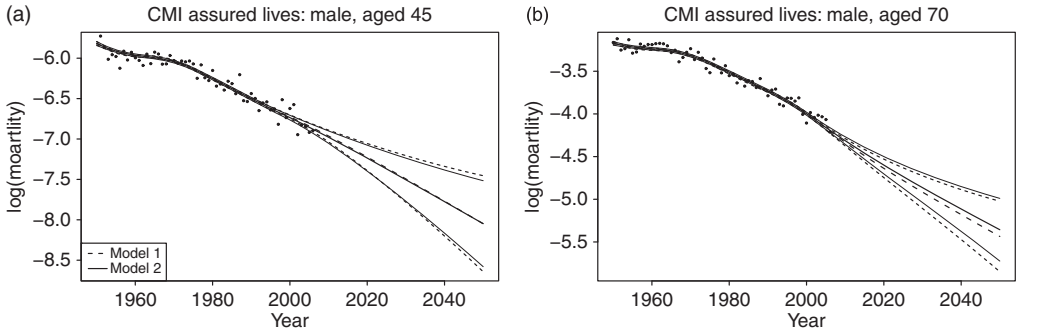

Figure 3. Fitted and forecast mortality with confidence intervals.

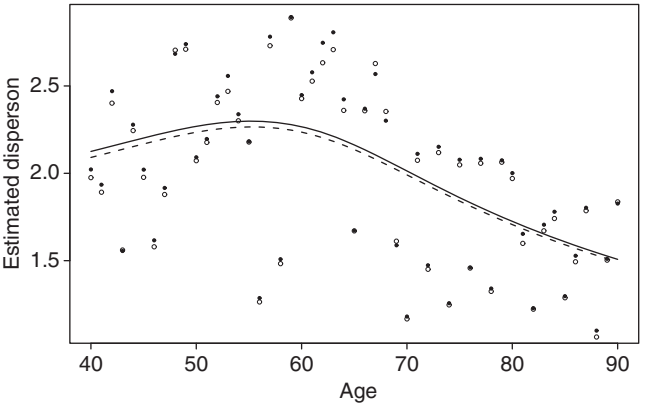

Figure 4. Raw (○○○) and smoothed estimates (- - -) of the ϕ_x's using the full extended quasi-likelihood scheme. Raw (●●●) and smoothed estimates (——) of the ϕ_x's using the bias corrected scheme.

the dispersion parameter in the estimation process, the effective dimension is reduced to 36 with Model 2, and to 33 with Model 3; this corresponds to a more robust, ie, less volatile fit. In general, this seems to us to be a desirable property for a forecast. Second, we consider the effect on the confidence intervals. Taking account of the over-dispersion has led to narrower confidence intervals. Why should this be? We argue as follows: smoothing is a compromise between (a) increasing roughness, ie, improved fit to data and (b) increasing smoothness, ie, poorer fit to data. When we include ϕ we down-weight the fit to data (the deviance is increased from 4968 for Model 1 first to 5076 for Model 2, and then again to 5121 for Model 3) and so decrease the volatility of the fitted model (the effective dimension is decreased first from 46 to 36 and then again to 33). The width of the forecast confidence intervals reflects our faith in the selected model and we will have more faith in the future direction of a forecast in a less volatile model; we conclude that the width of the confidence interval will be decreased. Both of these effects can be seen in Figure 3.

5.3. Smoothing mortality data by amounts

We consider CMI male pensioner data. These data are available both by lives and by amounts from 1983 to 2006 and ages 10 to 108. For this exercise we consider separately the data for those pensioners who are (a) age 65 and (b) age 73 (we will see below why we have chosen these ages).

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

Let $\boldsymbol{D} = (D_1, \ldots, D_n)'$ and $\boldsymbol{E} = (E_1, \ldots, E_n)'$ denote the numbers of claims and central exposed-to-risk, and let $\boldsymbol{D}^{[a]} = (D_1^{[a]}, \ldots, D_n^{[a]})'$, $\boldsymbol{E}^{[a]} = (E_1^{[a]}, \ldots, E_n^{[a]})'$ and $\boldsymbol{\mu}^{[a]} = (\mu_1^{[a]}, \ldots, \mu_n^{[a]})'$ denote the amount of pension of those that died, the amount at risk and the force of mortality (by amounts) for those pensioners age 65 (or 73 as the case may be). For amounts data, the problem of duplicates arises in two ways:

A: the original problem of duplicate policies in the portfolio, and

B: the much larger problem that amounts data by its very nature contains duplication on a grand scale, since a single death, even of a pensioner with a single pension, generates not one but multiple claims, namely the amount of pension at risk; see Forfar *et al.* (1988).

Figure 5 shows a plot of the observed log mortality by amounts, ie, $\log(\boldsymbol{D}^{[a]}/\boldsymbol{E}^{[a]})$ (where the quotient sign is interpreted as element-by-element division). We consider four approaches to smoothing the observed mortality rates; these approaches reflect different attitudes to problems A and B.

(a) Assume $D_i^{[a]} \sim \mathcal{P}\left(E_i^{[a]} \times \mu_i^{[a]}\right)$, (ie, ignore both A and B).

(b) Define $\boldsymbol{A} = \boldsymbol{E}^{[a]}/\boldsymbol{E}$, the mean amount at risk per life. Then the vector of raw mortalities by amounts is

$$\frac{\boldsymbol{D}^{[a]}}{\boldsymbol{E}^{[a]}} = \frac{\boldsymbol{D}^{[a]}/\boldsymbol{A}}{\boldsymbol{E}^{[a]}/\boldsymbol{A}} = \frac{\boldsymbol{D}^*}{\boldsymbol{E}} \tag{5.29}$$

where $\boldsymbol{D}^* = \boldsymbol{D}^{[a]}/\boldsymbol{A}$. If all policies are for the same amount then $\boldsymbol{D}^* = \boldsymbol{D}$ so we assume

$$D_i^* \sim \mathcal{P}\left(E_i \times \mu_i^{[a]}\right), \quad \text{(ie, adjust for B by scaling but ignore A).}$$

This is the 'method of scaling' referred to in our introduction.

(c) Assume a quasi-likelihood framework for $D_i^{[a]}$, (ie adjust for A and B simultaneously by quasi-likelihood):

$$\mathbb{E}\left[D_i^{[a]}\right] = E_i^{[a]} \times \mu_i^{[a]}, \quad var\left(D_i^{[a]}\right) = \phi \times \mathbb{E}\left[D_i^{[a]}\right].$$

This model is similar to that discussed by Renshaw & Hatzopoulos (1996).

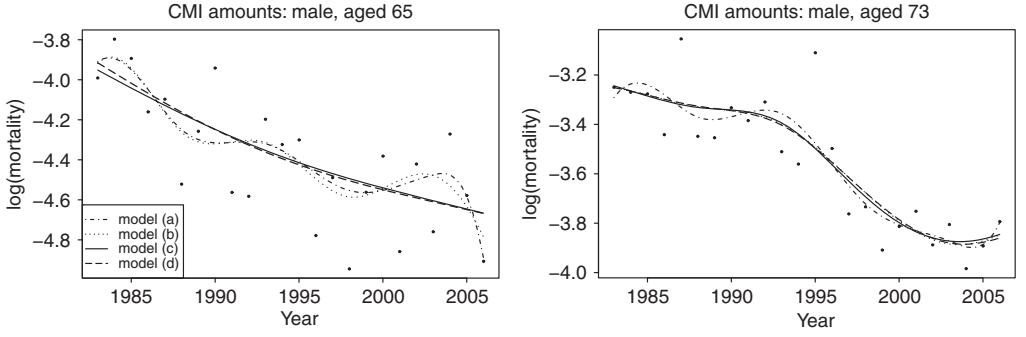

Figure 5. Various smooths of mortality by amounts.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

Table 2. Estimates of the over-dispersion parameter ϕ, the smoothing parameter λ and the effective dimension $tr(\boldsymbol{H})$.

	Age 65			Age 73		
	$\hat{\phi}$	λ	$tr(\boldsymbol{H})$	$\hat{\phi}$	λ	$tr(\boldsymbol{H})$
Model (a)	2020	0.01	8	2471	0.037	8
Model (b)	1.4	0.1	6.8	1.43	6.1	5.55
Model (c)	3940	148	2.7	3078	5.8	5
Model (d)	2.2	110	2.8	1.37	7.1	5.25

(d) Assume a quasi-likelihood framework for D_i^*, (ie adjust for B by scaling and for A by quasi-likelihood):

$$\mathbb{E}[D_i^*] = E_i \times \mu_i^{[a]}, \quad var(D_i^*) = \phi \times \mathbb{E}[D_i^*].$$

The smooth estimates which result from these four approaches are presented in Figure 5; various summary statistics obtained with the bias corrected scheme are provided in Table 2. We comment on each approach in turn. There are two serious objections to approach (a). Approach (a) is equivalent to assuming that $E_i^{[a]}$, the amount at risk, corresponds to $E_i^{[a]}$ independent lives, each with an amount at risk of £1. Thus, over-dispersion is an essential part of model (a). Indeed, on smoothing the age 65 data under model (a) we find (after the fitting) $\hat{\phi} = 2020$ while $\hat{\phi} = 2471$ at age 73. An even more serious objection to model (a) is that it depends on the unit of currency used to measure amounts. Simply by changing the units to hundreds of pounds, say, we alter the amounts claimed and the amounts at risk. The consequence of model (a) is that the fitted curve is substantially under-smoothed, since the exaggerated exposures force the fitted curve to follow the data, as can be seen in both panels of Figure 5.

Approach (b) is an attempt to solve the problems with (a). We scale the amounts claimed and the amounts at risk in such a way that the exposed-to-risk is once again measured in terms of policyholders' lives. Note that the problem with the units of currency is also solved by this scaling. The effect of scaling varies with our two illustrative ages: there is little effect on the fitted smooth at age 65 while at age 73 scaling results in a satisfactory smooth. There seems to be no clue in the values of $\hat{\phi}$ and λ in Table 1 for these different behaviours, but we can remark from (3.13) that a smoother curve results when the product of $\hat{\phi}$ and λ is large.

Approach (c) is very simple. We let the quasi-likelihood approach look after everything. We note in particular that the problem with the units of currency disappears since if $\boldsymbol{D}^\bullet = \boldsymbol{D}^{[a]}/c$ and $\boldsymbol{E}^\bullet = \boldsymbol{E}^{[a]}/c$ for some constant c then $\mathbb{E}[D_i^\bullet] = E_i^\bullet \times \mu_i^{[a]}$ and $var(D_i^\bullet) = (\phi/c) \times \mathbb{E}[D_i^\bullet]$ and the over-dispersion parameter adjusts to take account of the change of currency. The estimated smooth mortalities at both ages seem satisfactory. Approach (d) is a hybrid which combines the scaling argument of approach (b) with the over-dispersion modelling of approach (c). Approaches (c) and (d) give very similar results; see, for example, the values of $tr(\boldsymbol{H})$ in Table 2.

We remark that the over-dispersion only approach in (c) is capable on its own of adjusting for both the dispersion caused by amounts data per se and the 'intrinsic' dispersion caused by heterogeneity and duplicate policies. Renshaw & Hatzopoulos (1996) use this partition of

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

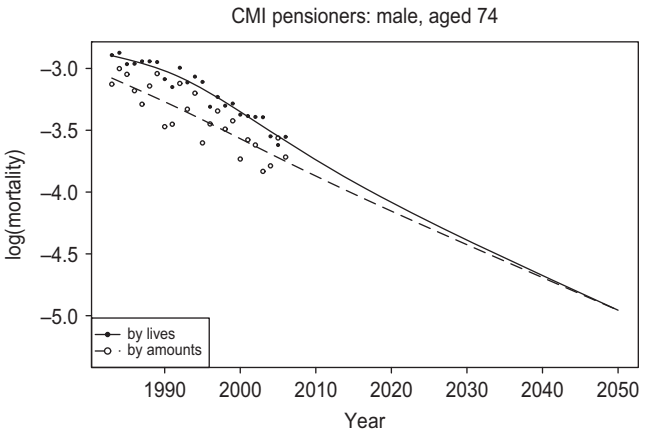

Figure 6. Profile view illustrating the crossing effect in the independent extrapolations of mortality by lives and by amounts for CMI pensioners.

dispersion to calculate dispersion in their model for amounts data. Model (b) uses scaling to adjust for the overdispersion caused by amounts data, but this is insufficient on its own and further over-dispersion modelling, as in model (d), is required. Approach (d) also suffers from the (possible) defect that it depends on a knowledge of the exposed-to-risk by lives; while this would generally be available, approach (c) does not use this information.

5.4. Joint modelling and forecasting of mortality by lives and by amounts

The CMI pensioner data consists of data on both lives and amounts. It seems natural to require that forecasts of mortality based on the lives data should be consistent with forecasts based on the amounts data. Yet, independent forecasts of mortality by lives and by amounts will inevitably lead to inconsistencies, as illustrated in Figure 6. It is well known that mortality by amounts is lighter than that by lives so we propose a joint model of mortality by lives and amounts that has consistency built into the model. Our joint model is an additive model with two components: the first component is a smooth 2-dimensional surface and the second component is a smooth age dependent curve which is constant in time. Thus, the mortality surface by lives sits on top of the mortality surface by amounts in such a way that the cross-sections in time by age are parallel. This model was discussed in Currie *et al.* (2004) but the treatment there used the Poisson assumption to model the lives data and the 'method of scaling' to model the amounts data (described in (b) in the previous subsection). Here we use the full quasi-Poisson assumptions as follows

$$\begin{cases} \mathbb{E}[D_{x,t}^{[a]}] = E_{x,t}^{[a]} \times \mu_{x,t}^{[a]}, \;\; var(D_{x,t}^{[a]}) = \phi_{x,t}^{[a]} \times \mathbb{E}[D_{x,t}^{[a]}] \\ \mathbb{E}[D_{x,t}^{[l]}] = E_{x,t}^{[l]} \times \mu_{x,t}^{[l]}, \;\; var(D_{x,t}^{[l]}) = \phi_{x,t}^{[l]} \times \mathbb{E}[D_{x,t}^{[l]}] \end{cases} \tag{5.30}$$

where the upper indexes "$[l]$" and "$[a]$" refer to the lives data and the amounts data respectively; for instance, for the amounts data at age x in calendar year t, $E_{x,t}^{[a]}$ is the exposure and $\phi_{x,t}^{[a]}$ is the dispersion parameter. Formulation (5.30) is quite general but here we will set $\phi_{x,t}^{[a]} = \phi_x^{[a]}$ and $\phi_{x,t}^{[l]} = \phi_x^{[l]}$, and then smooth these parameters across age as described in sections 3 and 4. The joint aspect of the model is then constructed through the linear

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

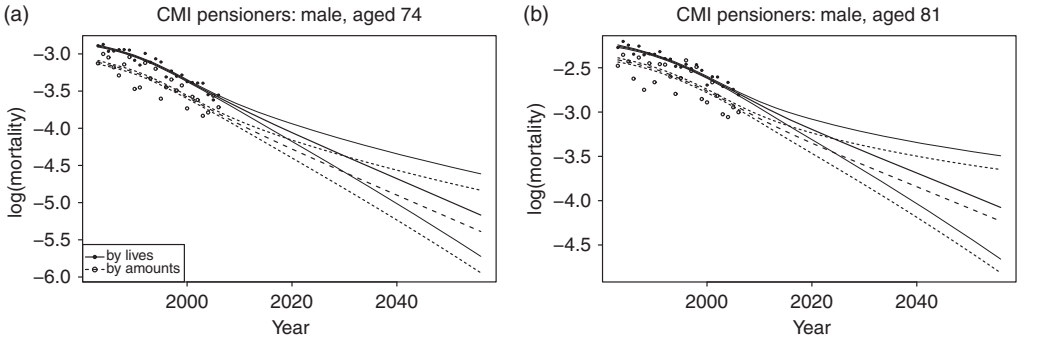

Figure 7. Profile views of the joint modelling of mortality by lives and by amounts for CMI pensioners.

predictor as follows:

$$\begin{cases} \log(\boldsymbol{\mu}^{[a]}) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta} \\ \log(\boldsymbol{\mu}^{[l]}) = (\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta} \ + \ (\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\delta} \end{cases} \tag{5.31}$$

where n_t is the number of years; that is, there is an underlying smooth surface, $(\boldsymbol{B}_t \otimes \boldsymbol{B}_x)\boldsymbol{\theta}$ (viewed as the reference surface and corresponding to the mortality surface by amounts). This surface drives an important part of the common dynamism in the mortality by lives and by amounts; and then, the relative variation between these two types of mortality is captured by the gap, $(\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\delta}$, which is smooth in age and constant in time. Thus, our modelling of the gap with $(\mathbf{1}_{n_t} \otimes \boldsymbol{B}_x)\boldsymbol{\delta}$ is designed to achieve flexibility in the age direction and parallelism in the time direction. One may argue that this is a strong assumption but it is a convenient and simple approach if we wish to avoid any crossing effects in the two extrapolated surfaces. Moreover, for the CMI data sets considered here, it produces satisfactory results (as we report below) because the underlying dynamism of mortality by lives and amounts supports such a model. Smoothness is obtained by applying the penalty in (5.27) on $\boldsymbol{\theta}$ and a penalty similar to (5.28) on $\boldsymbol{\delta}$.

We now define the joint vectors of the forces of mortality, $\boldsymbol{\mu} = vec(\boldsymbol{\mu}^{[a]}, \boldsymbol{\mu}^{[l]})$, and the coefficients, $\boldsymbol{\alpha} = vec(\boldsymbol{\theta}, \boldsymbol{\delta})$; the linear predictor in (5.31) can then be expressed compactly as

$$\log(\boldsymbol{\mu}) = \boldsymbol{B}\boldsymbol{\alpha}, \quad \text{with } \boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}_t \otimes \boldsymbol{B}_x & 0 \\ \boldsymbol{B}_t \otimes \boldsymbol{B}_x & \mathbf{1}_{n_t} \otimes \boldsymbol{B}_x \end{bmatrix}. \tag{5.32}$$

This compact formulation allows the joint model specified through (5.30) and (5.31) to be fitted with both schemes presented earlier. An illustration of the result (for selected profile views) fitted with the bias corrected scheme is displayed in Figure 7. An attractive point in this joint modelling approach is that the order of the two types of mortality in the joint predictor (5.31) does not matter, ie, flipping the two types of mortality will lead to the same fit. Indeed, the fitting here is not sequential; the coefficients $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ in (5.31) are estimated simultaneously.

6. Concluding remarks

We have described a general class of models for count data which allows the joint modelling of mean and dispersion effects through the extended quasi-likelihood. Renshaw (1992) was the first to

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

apply this method to actuarial data and we have extended his work to models for 2-dimensional data in age and time with general smooth functions for both the mean and dispersion surfaces. Smoothing is accomplished with the penalized *B*-spline method of Eilers & Marx (1996) which fits naturally into the generalized linear model framework. This enables more complex models such as the joint model for lives and amounts to be formulated. This last model is computationally very demanding since five smoothing parameters (two for the reference surface, one for the gap, and one for each of the age-dependent dispersion parameters for lives and amounts respectively) must be chosen in the context of a large regression model. The efficient array algorithms described in Currie *et al*. (2006) enable these calculations to be performed; these algorithms impact on the computations but not on model formulation.

Forfar *et al*. (1988, Sect. 17.2) used variance ratios to adjust for the presence of duplicate policies in a portfolio. Renshaw (1992) realised that there is sufficient information in the claims data alone to adjust for the presence of duplicates. This was a considerable advance since special investigations needed to be conducted to determine such variance ratios. We have discussed models for the smoothing and forecasting of mortality tables (by lives, by amounts, and by lives and amounts jointly). These models make proper adjustment for the presence of duplicates in particular and heterogeneity in general. An attractive feature of our models is that the treatments of mortality data by lives and by amounts are formally the same; this emphasizes that amounts data is fundamentally the problem of duplicates, a point originally made by Forfar *et al*. (1988, Sect. 16.4).

We close by mentioning a number of extensions to model (5.32). One possibility is a model to produce consistent forecasts for more than two populations; one application of actuarial interest is where claims are classified by age, year and duration of policy. A second possibility of interest in demography is to generalize model (5.32) to enable comparisons of mortality between different countries and/or different genders. In such cases, it may be of interest to allow the gap to vary not only by age (as in our lives/amounts example) but also by year; some examples of these models are given in Biatat & Currie (2010).

Acknowledgements

We are most grateful to Prof Angus Macdonald and Stephen Richards for useful discussions, and to the referees for their constructive comments. The work of Viani Djeundje was supported by a Dorothy Hodgkin Studentship and by a grant from the CMI; the work of Iain Currie was supported by a grant from the Actuarial Profession. We are also grateful to the CMI for permission to use their data.

References

Bäuerle, N. (1997). Inequalities for stochastic models via supermodular orderings. *Communications in Statistics – Stochastic Models*, **13**, 181–201.

Beard, R.E. & Perks, W. (1949). The relation between the distribution of sickness and the effect of duplicates on the distribution of deaths. *Journal of the Institute of Actuaries*, **75**, 75–86.

Biatat, V.D. & Currie, I.D. (2010). Joint models for classification and comparison of mortality in different countries. In *Proceedings of the 25th International Workshop on Statistical Modelling* (Ed: A.W. Bowman), 89–94. Glasgow: University of Glasgow.

Breslaw, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

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., 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.

CMI Committee (1957). Continuous investigation into the mortality of assured lives. *Journal of the Institute of Actuaries*, **83**, 34–36.

CMI Committee (1986). An investigation into the distribution of policies per life assured in the cause of death investigation data. *CMIR*, **8**, 49–58.

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

Currie, I.D., Durban, M. & 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.

Daw, R.H. (1946). On the validity of statistical tests of the graduation of a mortality table. *Journal of the Institute of Actuaries*, **72**, 174–190.

Daw, R.H. (1951). Duplicate policies in mortality data. *Journal of the Institute of Actuaries*, **77**, 261–267.

Denuit, M. (2000). Stochastic analysis of duplicates in life insurance portfolios. *German Actuarial Bulletin*, **24**, 507–514.

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

Forfar, D.O., McCutcheon, J.J. & Wilkie, A.D. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, **115**, 1–149.

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, **53**, 161–177.

Hinde, J. & Demetrio, C.G.B. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Human Mortality Database (2009). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (January, 2009).

Li, J.S.H., Hardy, M.R. & Tan, K.S. (2009). Uncertainty in mortality forecasting: an extension to the classic Lee-Carter approach. *Astin Bulletin*, **39**, 137–164.

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

Nelder, J.A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Renshaw, A.E. (1992). Joint modelling for actuarial graduation and duplicate policies. *Journal of the Institute of Actuaries*, **119**, 69–85.

Renshaw, A.E. & Hatzopoulos, P. (1996). On the graduation of 'amounts'. *British Actuarial Journal*, **2**, 185–205.

Richards, S.J. (2008). Applying survival models to pensioner mortality data. *British Actuarial Journal*, **14**, 257–326.

Richards, S.J. & Currie, I.D. (2009). Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal*, **15**, Part II (to be published).

Richards, S.J., Kirkby, J.G. & Currie, I.D. (2006). The importance of year of birth in two-dimensional mortality data. *British Actuarial Journal*, **12**, 5–61.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
Seal, H.L. (1940). Tests of a mortality table graduation. *Journal of the Institute of Actuaries*, **71**, 5–67.
Shaked, M. & Shanthikumar, J.G. (1997). Supermodular stochastic orders and positive dependence of random vectors. *Journal of Multivariate Analysis*, **61**, 86–101.
Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144–148.
Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.

https://doi.org/10.1017/S1748499510000047 Published online by Cambridge University Press