J. R. Statist. Soc. A (2001)
164, *Part* 2, *pp.* 275–291

Bayesian analysis of mortality data

Petros Dellaportas,

Athens University of Economics and Business, Greece

Adrian F. M. Smith

Queen Mary and Westfield College, London, UK

and Photis Stavropoulos

University of Glasgow, UK

[Received May 1995. Final revision July 2000]

Summary. Congdon argued that the use of parametric modelling of mortality data is necessary in many practical demographical problems. In this paper, we focus on a form of model introduced by Heligman and Pollard in 1980, and we adopt a Bayesian analysis, using Markov chain Monte Carlo simulation, to produce the posterior summaries required. This opens the way to richer, more flexible inference summaries and avoids the numerical problems that are encountered with classical methods. Particular methodologies to cope with incomplete life-tables and a derivation of joint lifetimes, median times to death and related quantities of interest are also presented.

Keywords: Age; Bayesian inference; Incomplete life-table; Markov chain Monte Carlo methods; Mortality data

1. Introduction

The representation of mortality data via a parametric model has attracted the attention of actuaries, demographers and statisticians for over a century. The most famous such model is that of Gompertz (1825), which is still used by demographers today; see Pollard (1991). This approach of summarizing mortality patterns has many advantages over other ways of describing such data, particularly because it facilitates comparisons over time and space, e.g. between cohorts, periods and regions. In a thorough review of demographic models, Congdon (1993) described in detail the advantages of a parameterized model approach: smoothness, parsimony, interpolation, comparison, trends and forecasting and analytic manipulation. Renshaw (1991) presented a generalized linear and non-linear models approach to mortality graduation and provided arguments in its favour. Haberman and Renshaw (1996) reviewed the application of generalized linear models to several problems arising in actuarial science. Other approaches to further summarizing mortality data, such as numerical tabulations, relational procedures, orthogonal polynomials and splines, can be found in Benjamin and Pollard (1980) and Forfar *et al.* (1988).

A recent attempt to represent mortality across the entire age range has been the eight-parameter non-linear model of Heligman and Pollard (1980). Their suggested model has been

Address for correspondence: Petros Dellaportas, Department of Statistics, Athens University of Economics and Business, 76 Patission Street, GR-10434 Athens, Greece.
E-mail: petros@aueb.gr

0964–1998/01/164275

used in the past for a wide range of mortality data, resulting in satisfactory representations of a variety of patterns; see, for example, Heligman and Pollard (1980), Mode and Busby (1982), Forfar and Smith (1987), Hartmann (1983), Rogers (1986), Kostaki (1991, 1992a, b) and Congdon (1993). However, as noted by Rogers (1986) and Congdon (1993), an estimation of the parameters is problematic owing to the overparameterization of the model, and, also, numerical instabilities in the (commonly used) weighted least squares approach can only be removed by fixing two parameters to be constant. Furthermore, such numerical difficulties create large fluctuations in parameter estimation over time or space, resulting, because of similar inconsistencies in the dispersion matrices, in a lack of substantive interpretability.

We adopt a Bayesian inference approach which has the following advantages over currently used methods. First, because the parameters of the model have a straightforward interpretation, the use of informative prior distributions resolves the problem of overparameterization. Secondly, the non-normality of the likelihood surface in the parameterization that is usually adopted means that the least square estimates are inadequate. Thirdly, an application to an incomplete life-table can be routinely made by using simulation-based Bayesian computation methodology. Fourthly, posterior densities of other quantities of interest such as the joint lifetime of a couple or the median lifetime of a person can be derived. For other Bayesian work related to mortality smoothing and life-table construction, see Kimeldorf and Jones (1967), Hickman and Miller (1977) and Carlin (1992); Carlin (1992) used Markov chain Monte Carlo methods but not in a parametric curve modelling context.

This paper is organized as follows. In Section 2 we present the Heligman and Pollard (1980) formula, suggest two possible modelling approaches for complete life-tables and illustrate them with English and Welsh mortality data. In Section 3 we extend this methodology to incomplete life-tables, and we illustrate it on the same set of data. Finally, in Section 4 we focus on four problems mentioned in Pollard (1991), namely the calculation of survival and first-to-die probabilities and the derivation of joint and median lifetimes.

2. Inference for complete life-tables

Heligman and Pollard (1980) suggested a model which represents the underlying life-table probabilities π_x of dying between exact ages x and $x+1$, for $x = 0, 1, 2, \ldots, n$, via the curve

$$\frac{\pi_x}{1-\pi_x} = A^{(x+B)^C} + D \exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x. \qquad (1)$$

The idea behind this model is to decompose the odds that an individual of age x will die before he or she attains age $x+1$ into three parts: a child mortality curve, an accident hump in early adult life and an adult mortality curve. As an example, in Fig. 1 we present $\log(\pi_x)$ and its three parts evaluated from equation (1) for certain values of A, B, C, D, E, F, G and H.

The interpretation of the parameters in this model is straightforward. The parameter A, taking values in the interval (0, 1), represents the infant mortality rate. B represents the mortality rate for children who are 1 year old, taking values within the interval (0, 1). The parameter C takes values in (0, 1) and is closely associated with the rate of mortality decline, or the rate at which an individual adapts to his environment. The three parameters of the mid-life mortality component, D, E and F, reflect what is often referred to in the demographic literature as the accident hump; see, for example, Congdon (1993). In particular, D indicates the severity of the accident hump and takes values in (0, 1), E is defined in $(0, \infty)$ and is

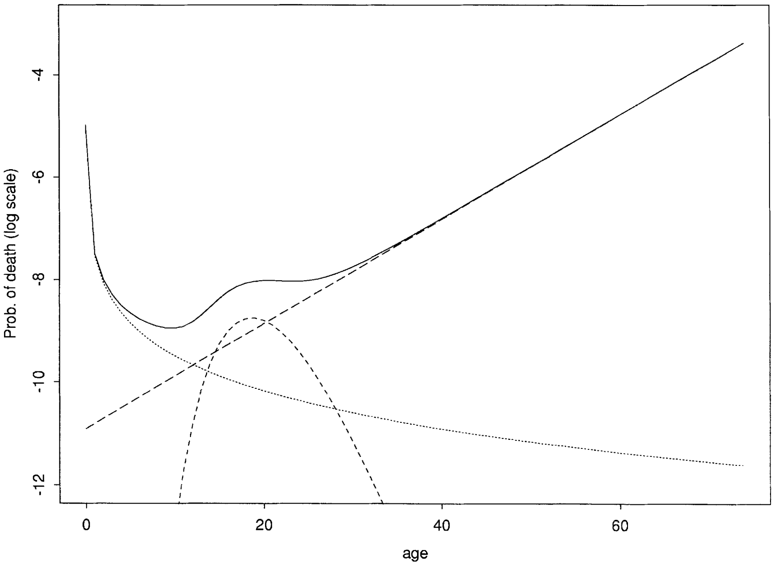

Fig. 1. Progress with age of the logarithm of the probability of dying and of the logarithms of its three parts as given by the Heligman–Pollard formula: the formulae were evaluated at (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*) = (5.44 × 10^{-4}, 1.70 × 10^{-2}, 1.01 × 10^{-1}, 1.58 × 10^{-4}, 10.72, 18.67, 1.83 × 10^{-5}, 1.11); this is an estimate of the posterior mean of (*A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*) for the logistic non-linear model applied to the data of Table 1 (———, full curve; ··········, infant mortality; - - - - -, accident hump; — — —, old age mortality)

related to the spread, with large values indicating a concentrated accident hump, and F is indicative of the location with domain the interval (15, 110). Finally the two parameters in the third term in equation (1), G and H, have domains the intervals (0, 1) and (0, ∞) and indicate the base level of later adult mortality and the rate of increase in mortality at the later adult ages respectively.

Heligman and Pollard (1980) also suggested a slightly alternative model with one more extra parameter in the third part of the model; see also Congdon (1993). Kostaki (1992b) suggested the inclusion of another parameter in the second part of the model. In this paper we shall only focus on the eight-parameter model (1).

Many researchers have noticed that the estimation of the parameters $\boldsymbol{\theta} = (A, B, C, D, E, F, G, H)$ of model (1) by least squares methods is problematic. In fact, Heligman and Pollard (1980) suggested the use of weighted least squares with weights $w_x = 1/q_x^2$, where q_x is the observed raw frequency estimate of π_x. In an experimental study, Kostaki (1992a) noted that other forms of weights w_x cannot provide any form of convergence, and similar results appear in Hartmann (1983). Rogers (1986) and Congdon (1993) fixed two parameters to avoid numerical instabilities in the non-linear least squares minimization procedure. Note that in general, with the exception of the work of Congdon (1993), none of the researchers who have used model (1) report estimates of the dispersion matrix, a task which poses further computational problems.

2.1. A non-linear logistic model

Following the Bayesian framework, we treat all model parameters as unknowns and we specify prior information via probability density functions. We believe that the demographic interpretability of the parameters will result in informative priors for nearly all the parameters. Let n_x be the population at risk having age in the interval $[x, x+1)$ and d_x the number of people who die at that age. Treating the counts of dead individuals at different ages as independent random variables, each follows a binomial distribution

$$p(d_x) = \binom{n_x}{d_x} \pi_x^{d_x} (1-\pi_x)^{n_x-d_x}$$

and so if $\mathbf{d}$ is the vector with elements d_x, $x = 0, 1, \ldots, n$, we obtain

$$p(\mathbf{d}) = \prod_{x=0}^{n} \binom{n_x}{d_x} \pi_x^{d_x} (1-\pi_x)^{n_x-d_x}.$$

If we denote the right-hand side of equation (1) by $K(x)$ it follows that

$$p(\mathbf{d}) = \prod_{x=0}^{n} \binom{n_x}{d_x} K(x)^{d_x} \{1+K(x)\}^{-n_x}. \tag{2}$$

This is a problem of estimating the parameters of a generalized non-linear model.

Denoting by $p(\boldsymbol{\theta})$ the prior distribution for $\boldsymbol{\theta} = (A, B, C, D, E, F, G, H)$, it follows that for data $\mathbf{d} = d_x$, $x = 0, 1, \ldots, n$, the resulting posterior is of the form

$$p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\boldsymbol{\theta})\, p(\mathbf{d}) = p(\boldsymbol{\theta}) \prod_{x=0}^{n} \binom{n_x}{d_x} K(x)^{d_x} \{1+K(x)\}^{-n_x}. \tag{3}$$

Analytic integration of this eight-dimensional non-normalized posterior joint distribution is not feasible so we adopt a simulation-based approach to obtaining the posterior summaries desired. In particular, we suggest a Markov chain Monte Carlo algorithm, at each iteration of which we update all elements of $\boldsymbol{\theta}$ simultaneously; the background theory is summarized in Appendix A.

For each Metropolis step we first make a transformation of the parameter vector $\boldsymbol{\theta}$ to a new vector $\boldsymbol{\theta}' \in \Re^8$ so that the resulting posterior should be 'close' to normality; see Hills and Smith (1992). For example, we take

$$\begin{aligned} A' &= \log\left(\frac{A}{1-A}\right), \\ E' &= \log(E). \end{aligned} \tag{4}$$

After taking care of the Jacobian, we apply a Metropolis step to $\boldsymbol{\theta}'$ using as an initial proposal distribution a multivariate normal distribution with parameters $\hat{\boldsymbol{\mu}}$ and $c\hat{\boldsymbol{\Sigma}}$. Here $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the maximum likelihood mean vector and covariance (inverse Hessian) matrix derived by using a non-linear weighted least squares algorithm with weights $w_x = 1/q_x^2$ as suggested by Heligman and Pollard (1980); c is a prespecified constant, which is tuned to achieve better convergence behaviour measured with respect to both the sampling efficiency (percentage of accepted proposed moves) and the rate of convergence. After the initial iteration, the mean vector of the proposal density is updated with the current sampled parameter vector.

2.2. A model accounting for extra-binomial variation

It can be argued that equation (1) is too restrictive. A deterministic relationship between the age-dependent probabilities of death π_x and the parameters of the Heligman–Pollard model may not explain all the variation exhibited by observed death-rates. We propose to deal with this as follows. We assume that the Heligman–Pollard formula determines age-dependent quantities m_x:

$$\frac{m_x}{1-m_x} = A^{(x+B)^C} + D\,\exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x.$$

The probabilities of death are random quantities following a beta distribution,

$$\pi_x \sim \text{beta}\{\zeta m_x, \zeta(1-m_x)\},$$

where ζ is an unknown positive quantity determining the variance of the beta distribution. More specifically, the expected value and variance of π_x given ζ and m_x are

$$E(\pi_x|\zeta, m_x) = m_x,$$
$$V(\pi_x|\zeta, m_x) = \frac{m_x(1-m_x)}{1+\zeta}.$$

If $\zeta \to \infty$, π_x will again be given by equation (1). The rest of the model remains the same as the non-linear logistic model of the previous section. If we assign to ζ a prior distribution $p(\zeta)$ a full Bayesian analysis leads to the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\pi}, \zeta|\mathbf{d}, \mathbf{n})$, where $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_n)$ We can then express our uncertainties through the marginal posterior distribution for $\boldsymbol{\theta}$.

We again use Markov chain Monte Carlo methods to sample from this analytically intractable posterior distribution. The required full conditional distributions are found to be

$$p(\boldsymbol{\theta}|\mathbf{d}, \mathbf{n}, \boldsymbol{\pi}, \zeta) \propto p(\boldsymbol{\theta}) \prod_{x=0}^{n} \pi_x^{\zeta m_x - 1}(1-\pi_x)^{\zeta(1-m_x)-1},$$

$$p(\boldsymbol{\pi}|\mathbf{d}, \mathbf{n}, \boldsymbol{\theta}, \zeta) \propto \prod_{x=0}^{n} \pi_x^{d_x+\zeta m_x - 1}(1-\pi_x)^{n_x-d_x+\zeta(1-m_x)-1} = \prod_{x=0}^{n} \text{beta}\{d_x+\zeta m_x, n_x-d_x+\zeta(1-m_x)\}$$

and

$$p(\zeta|\mathbf{d}, \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(\zeta) \prod_{x=0}^{n} \pi_x^{\zeta m_x - 1}(1-\pi_x)^{\zeta(1-m_x)-1}.$$

Sampling from the full conditional density of the $\boldsymbol{\pi}$ requires sampling from beta densities. For the other conditional densities we use Metropolis–Hastings steps. For $\boldsymbol{\theta}$ this step utilizes the same proposal distribution as in the previous section, whereas for ζ the proposal distribution is a log-normal density centred on the old value of ζ with variance tuned so that the acceptance rate is around 0.5.

2.3. A log-normal model

The weighted least squares algorithm suggested by Heligman and Pollard (1980) and also advocated by Kostaki (1992a) uses weights $w_x = 1/q_x^2$. This is implicitly based on the assumption of a constant coefficient of variation across age: for each q_x with mean π_x and estimated variance σ_x^2, the use of this weight function, instead of the more common form $w_x = 1/\sigma_x^2$, implies that $\sigma_x^2 \propto q_x^2$ and therefore σ_x/π_x is constant since π_x is estimated by q_x. An equivalent modelling approach can be based on a log-normal model. To clarify this, assume that we have response variables y_i generated by a model of the form $\log(y_i) = \log(f_i) + \epsilon_i$, with f_i being a function of parameters and covariates and $\epsilon_i \sim N(0, \sigma^2)$. Then, $y_i = f_i \exp(\epsilon_i) \simeq f_i(1 + \epsilon_i)$, with $E(y_i) = f_i$ and $V(y_i) = f_i^2 \sigma^2$. The coefficient of variation is then equal to σ.

For our mortality data problem, assuming that the probability odds have a constant coefficient of variation, the log-normal model can be stated as

$$\log\left(\frac{q_x}{1-q_x}\right) = \log\left(A^{(x+B)^C} + D \exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x\right) + \epsilon_x \tag{5}$$

where ϵ_x are independent $N(0, \sigma^2)$ variables. In this case, for data $\mathbf{X} = q_x = d_x/n_x$, $x = 0, 1, \ldots, n$, and an independent prior $p(\sigma^2)$ for σ^2, the posterior is of the form

$$p(\boldsymbol{\theta}, \sigma^2|\mathbf{X}) \propto p(\boldsymbol{\theta})\, p(\sigma^2) \sigma^{-(n+1)} \exp\left[-\frac{1}{2\sigma^2} \sum_{x=0}^{n} \left\{\log\left(\frac{q}{1-q_x}\right) - \log\left(A^{(x+B)^C} + D \exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x\right)\right\}^2\right]. \tag{6}$$

Here the Markov chain Monte Carlo algorithm requires the updating of both the parameter vector $\boldsymbol{\theta}$ and σ^2. For the former we use the same Metropolis sampler as in the previous section, whereas for the latter we use a Gibbs step, noting that with an uninformative prior $p(\sigma^2) = \sigma^{-2}$ the conditional density $p(\sigma^2|\boldsymbol{\theta})$ is

$$p(\sigma^2|\boldsymbol{\theta}) \equiv \mathrm{IG}\left(\frac{n+1}{2}, 2\left[\sum_{x=0}^{n} \left\{\log\left(\frac{q_x}{1-q_x}\right) - \log\left(A^{(x+B)^C} + D \exp\left[-E\left\{\log\left(\frac{x}{F}\right)\right\}^2\right] + GH^x\right)\right\}^2\right]^{-1}\right). \tag{7}$$

This model is clearly less appropriate than the model of Section 2.1 because of the assumption of constant error variance across age. For the non-linear logistic model note that the variance of $\log\{q_x/(1-q_x)\}$ is, using a Taylor expansion, equal to $1/\{n_x \pi_x(1-\pi_x)\}$ which is not constant. If we assume that $1/n_x$ does not change across age, a suitable variance for ϵ_x in equation (5) would be $\sigma_x^2 = \sigma^2/\{\pi_x(1-\pi_x)\}$ estimated by $\sigma^2/\{q_x(1-q_x)\}$. However, n_x is rarely constant across age.

Carlin (1992) also adopted a Bayesian perspective and used Monte Carlo techniques for mortality graduation. Nevertheless, he did not use a parametric model to describe the behaviour of the true probabilities of death. Instead, he considered the latter as the unknown individual parameters.

2.4. An illustrative example

We illustrate the methodology of the previous subsections by using the 1988–1992 mortality data of English and Welsh females illustrated in Table 1. The estimated resident populations on June 30th of each year and the yearly counts of deaths were summed across the five years to avoid having the observed mortality rates influenced by extreme random fluctuations.

All the models are applied to these data for comparison. The prior for σ^2 is the prior mentioned in the previous section, whereas in all the models the same prior distribution for $\boldsymbol{\theta}$ is chosen as follows. To each component parameter, we assign, independently, a log-normal distribution $\log(\theta_i) \sim N(\mu_i, \sigma_i^2)$, where μ_i and σ_i are implicitly defined by specifying z_{i1} and z_{i2} to be suitable lower and upper percentiles of the normal distribution (in this application, we

Table 1. English and Welsh mortality data, females, 1988–1992

x	n_x	d_x	x	n_x	d_x	x	n_x	d_x
0	1682000	11543	25	2078400	689	50	1375500	4130
1	1666400	940	26	2084300	698	51	1365400	4564
2	1644700	538	27	2067000	712	52	1373500	5017
3	1634400	420	28	2021400	799	53	1361300	5417
4	1610000	332	29	1963000	795	54	1335900	5786
5	1581800	250	30	1903800	787	55	1313900	6567
6	1564500	254	31	1844600	935	56	1306200	7173
7	1554700	228	32	1788000	978	57	1306200	8068
8	1549800	208	33	1745500	977	58	1314600	8809
9	1544600	215	34	1714800	1131	59	1325400	10148
10	1514300	182	35	1690300	1219	60	1330600	11390
11	1482500	200	36	1671400	1270	61	1332100	12789
12	1453900	215	37	1668000	1435	62	1328200	13999
13	1436700	204	38	1684600	1516	63	1322300	15528
14	1443000	294	39	1707600	1693	64	1323000	17368
15	1496400	339	40	1755900	1905	65	1329000	19277
16	1576800	412	41	1844500	2207	66	1344200	20991
17	1670500	535	42	1837500	2517	67	1370100	23665
18	1744500	561	43	1812200	2565	68	1408200	26365
19	1822800	592	44	1777100	2918	69	1337400	27664
20	1883200	591	45	1699800	3077	70	1249500	28397
21	1930400	640	46	1563200	3119	71	1174200	29178
22	1964400	623	47	1499200	3369	72	1098800	30437
23	2015600	653	48	1453200	3677	73	1029500	32146
24	2051700	668	49	1408400	3740	74	1052400	35728

obtained the 1% and 99% percentile values by discussion with a colleague, Dr A. Kostaki). The 1% values are set at

$$(10^{-4}, 10^{-4}, 10^{-2}, 5 \times 10^{-5}, 0, 15, 10^{-7}, 1)$$

and the 99% values at

$$(2 \times 10^{-2}, 15 \times 10^{-2}, 3 \times 10^{-1}, 10^{-2}, 20, 110, 10^{-3}, 12 \times 10^{-1}).$$

For illustration, these represent vague, but still informative, prior beliefs. Given experience with a series of such problems, more informative forms of prior beliefs would, of course, evolve, as would some knowledge of parameter correlations.

We can now perform the Markov chain Monte Carlo analysis algorithm; see Smith and Roberts (1993) for details. For our example, the sampling strategy chosen was to simulate a single chain. After a number of 'burn-in' iterations required for convergence we stored the outcome of every kth iteration until a sample of 2500 realizations was formed. The length of burn-in was assessed with the method of Raftery and Lewis (1992). Moreover, with the aid of visual inspection of trace plots of the realizations of the chain and of the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983) we increased the length of the burn-in period so that we were fairly sure that it was sufficiently long. k was such that there was no serial correlation in the stored sample. For the logistic non-linear model for example, the length of burn-in was 100000 and k was 50. Fig. 2 shows the results for all three models. The dots represent the observed probabilities of death. For each model, curve (1) was evaluated at the posterior mean of $\boldsymbol{\theta}$ which was estimated by the corresponding sample mean.

Fig. 2 also shows 95% predictive intervals for the logarithm of the observed probabilities

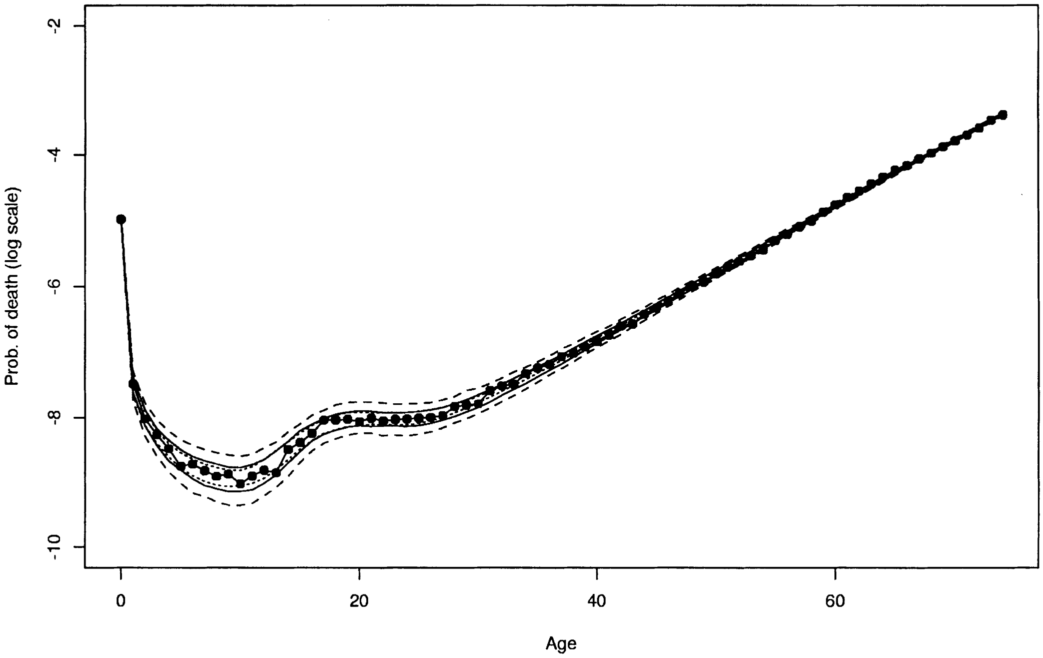

Fig. 2. Empirical data (●) and fitted and 95% predictive curves for a full table provided by the logistic non-linear (———), log-normal (··········) and extra-binomial variation (- - - - -) models

of death across age. They were calculated as follows. For the logistic non-linear model and for each integer age we have, via the stored output sample of $\boldsymbol{\theta}$, 2500 estimates of the true probability of dying at that age, $q_x(i)$ say, $i = 1, \ldots, 2500$. For each $q_x(i)$ we sample one value $d_x(i)$ from the corresponding binomial distribution $\text{Bin}\{b_x, q_x(i)\}$ and thus we obtain a posterior predictive sample $d_x(i)/n_x(i)$ for the observed probability of dying. The empirical 2.5% and 97.5% quantiles of that sample form the above-mentioned intervals.

For the log-normal model, for each point in the posterior sample of $\boldsymbol{\theta}$, σ^2, we generate one value from the corresponding log-normal distribution (5) and thus, after a straightforward transformation, we obtain a posterior predictive sample $q_x(i)$ for the observed probability of dying. The construction of the predictive intervals proceeds as before. For the model dealing with extra-binomial variation we form the predictive intervals with the same procedure as before with the inclusion of an extra step where $q_x(i)$ is obtained by sampling from a beta distribution.

We see that the estimated curves for all three models provide a good fit for the observed frequencies. However, the allowance for a changing variance across age in the binomial model gives better predictive intervals than the log-normal model does. The intervals provided by the latter are narrower and they miss some of the observed probabilities of dying. Note that the intervals resulting from the model accounting for possible extra variation are the widest intervals. They reflect a deterioration in the precision of our estimates due to the inclusion of ζ in the logistic model. In the actual example that we present here the posterior sample mean of ζ was 537292.6 with the Markov chain Monte Carlo output sample ranging between 293605.1 and 1098843.

The pairwise bivariate marginals, in the case of the logistic non-linear model, are shown in

Table 2. Approximate posterior correlations for the parameters of the logistic non-linear model

A	B	C	D	E	F	G	H
1.00							
0.89	1.00						
0.82	0.98	1.00					
0.16	0.23	0.24	1.00				
−0.20	−0.33	−0.36	0.39	1.00			
0.01	−0.04	−0.06	−0.16	−0.05	1.00		
0.15	0.23	0.25	0.05	0.19	−0.22	1.00	
−0.13	−0.21	−0.23	−0.04	−0.20	0.22	−0.99	1.00

Table 2 in the form of a table of correlations. We can see the strong posterior correlations between some pairs of parameters (e.g. between G and H, and B and C).

The resulting marginal distributions, along with those from the grouped data case, are given in Fig. 3 in the form of box plots. We note, in particular, that most of the marginals are markedly skewed to the right, underlining the danger of posterior or likelihood normal approximations in the original parameterization.

3. Inference for incomplete life-tables

Mortality data are often collected by using 5-year age groups rather than individual years of life, except for the first 5 years, which are presented in the two intervals [0, 1) and [1, 5). Thus, the available data—so-called incomplete or abridged life-tables—are of the form d_0, D_1, D_x and $x = 5, 10, 15, \ldots, n'$, where D_x denotes the number of people who died at an age in the interval starting at x, $[x, x+1, \ldots, x+k]$ say, and n' is the starting age of the last group. Such data are collected, for example, from the World Health Organization. This is a common method of describing mortality patterns for reasons of convenience, especially in countries with incomplete and unstable documentation of vital statistics.

The problem of extending an incomplete to a full life-table has been studied by Mode and Busby (1982), Pollard (1989) and Kostaki (1991). From a Bayesian perspective, the incomplete life-table problem can be seen as an incomplete data problem, or, as we shall show below, as a constrained parameter problem. We can then use a general approach to such situations using a Markov chain Monte Carlo strategy; see Gelfand *et al.* (1992).

We assume, as in the case of complete data, that the true probability π_x of a person dying at age x is given by formula (1). We again denote by $\boldsymbol{\theta}$ the parameter vector, and by d_x the number of people dying at age x which is, as before, binomially distributed with parameters n_x and π_x. The only known non-random quantities are the grouped population counts N_x and the population with age 0 at last birthday and the only observed data are the grouped death counts D_x and the number of dead children under 1 year old. All the other quantities, i.e. $\boldsymbol{\theta}$, d_x and n_x, will be considered, in a Bayesian perspective, as unknowns. Let $\mathbf{N}$, $\mathbf{D}$, $\mathbf{n}$ and $\mathbf{d}$ be the vectors of the corresponding quantities. Note that $\mathbf{n}$ and $\mathbf{d}$ do not contain the quantities for $x = 0$.

In this case the full model for the data and the unknowns given everything that is known will be

$$p(\mathbf{D}, \mathbf{d}, d_0, \mathbf{n}, \boldsymbol{\theta}|\mathbf{N}, n_0) = p(\mathbf{D}|\mathbf{d}, \mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, d_0, n_0)\, p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, d_0, n_0)\, p(d_0|\mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, n_0)\, p(\mathbf{n}, \boldsymbol{\theta}|\mathbf{N}, n_0).$$

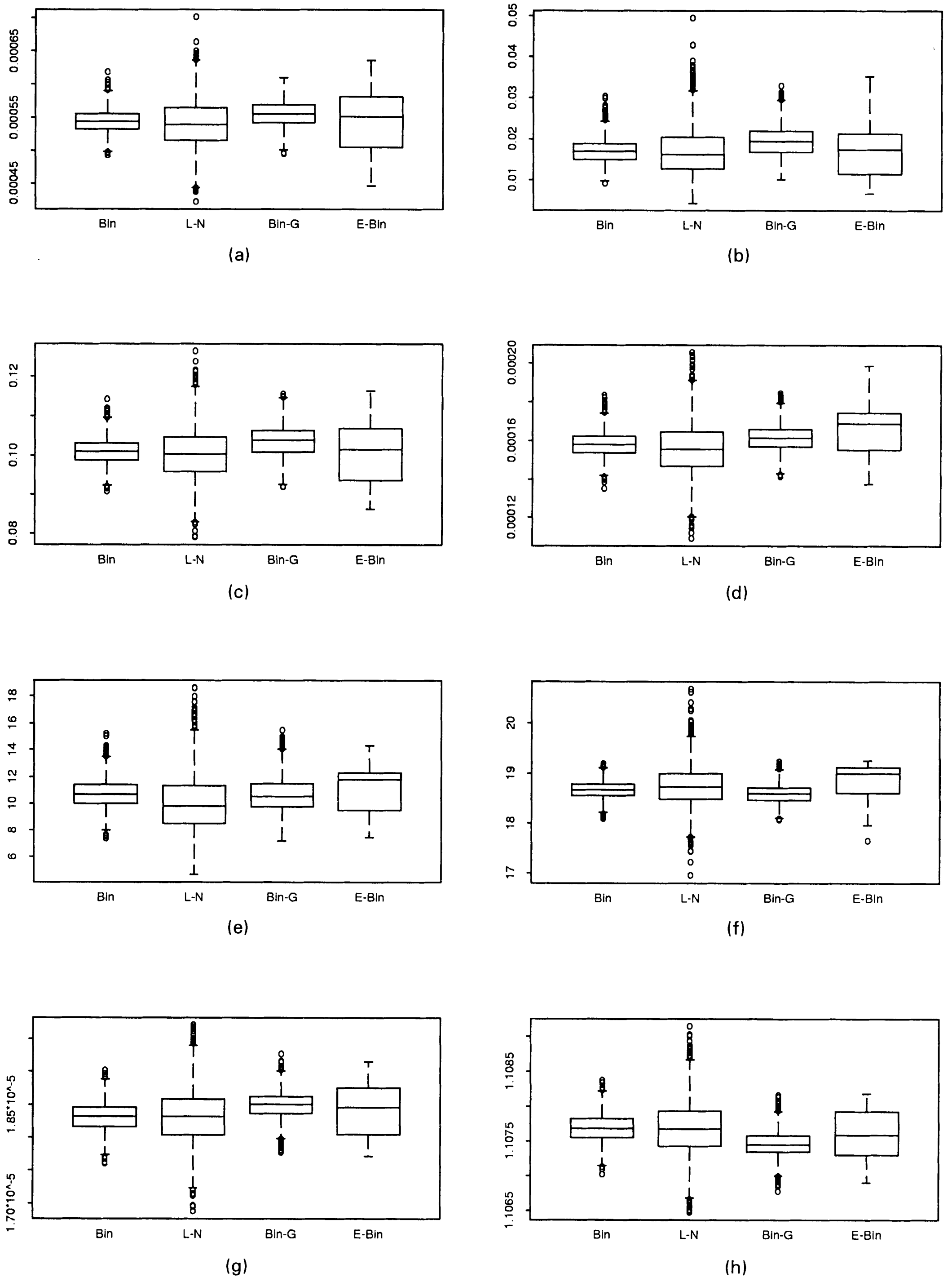

Fig. 3. Box plots of posterior marginal distributions (Bin, logistic non-linear model—full table; L-N, log-normal model—full table; Bin-G, logistic non-linear model—abridged table; E-Bin, model accounting for extra-binomial variation—full table): (a) *A*; (b) *B*; (c) *C*; (d) *D*; (e) *E*; (f) *F*; (g) *G*; (h) *H*

We now examine each of the above conditional densities in detail. First, we note that $\mathbf{D}$ depends only on $\mathbf{d}$ since

$$D_x = \sum_{i=x}^{x+k} d_i$$

and therefore

$$p(\mathbf{D}|\mathbf{d}, \mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, d_0, n_0) = \prod p(D_x|d_x, \ldots, d_{x+k}),$$

a product of indicator functions, the product ranging over $x = 1, 5, 10, \ldots, n'$.

When $\mathbf{n}$ and $\boldsymbol{\theta}$ are known, $\mathbf{d}$ depends only on these quantities, and, as in the previous section, its components are independent random quantities binomially distributed:

$$p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, d_0, n_0) = p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta}) = \prod_{x=1}^{n} p\{d_x|n_x, \pi_x(\boldsymbol{\theta})\},$$

where by $\pi_x(\boldsymbol{\theta})$ we denote the probability of dying as a function of the parameters of the Heligman–Pollard formula (1).

Similarly,

$$p(d_0|\mathbf{n}, \boldsymbol{\theta}, \mathbf{N}, n_0) = p\{d_0|n_0, \pi_0(\boldsymbol{\theta})\}.$$

Finally, we can consider that, *a priori*, $\boldsymbol{\theta}$ is independent of anything else and that $\mathbf{n}$ given $\mathbf{N}$ is independent of $\boldsymbol{\theta}$ so that

$$p(\mathbf{n}, \boldsymbol{\theta}|\mathbf{N}, n_0) = p(\boldsymbol{\theta})\, p(\mathbf{n}|\mathbf{N}, n_0)$$

where $p(\boldsymbol{\theta})$ is the same prior of $\boldsymbol{\theta}$ as in the complete-data case, whereas $p(\mathbf{n}|\mathbf{N}, n_0)$ is, likewise, the expression of our prior beliefs about population counts per year of age in the light of the observed counts alone. The prior that we choose is of the form

$$p(\mathbf{n}|\mathbf{N}, n_0) = \prod p(n_x, \ldots, n_{x+k}|N_x).$$

The product is, again, over $x = 1, 5, 10, \ldots, n'$. We assume that counts from different age groups are independent given the grouped counts and that within each group the distribution of the individual counts is uniform over the set of all combinations of positive values that sum to the grouped count. We do this because *a priori* we have no information about $\mathbf{n}$.

3.1. Simulation details

The quantity of interest is $\boldsymbol{\theta}$ only but since we have missing data we must also carry these along during the simulation stage. Therefore, we seek to sample from $p(\boldsymbol{\theta}, \mathbf{d}, \mathbf{n}|\mathbf{D}, \mathbf{N}, d_0, n_0)$. This is proportional to the full model mentioned in the previous section. A convenient simulation method is the Gibbs sampler.

We update each vector separately because of their different characteristics. The vectors have complicated conditional posterior distributions, which we shall mention below, and therefore for each vector we use the Metropolis–Hastings algorithm, which can handle distributions of arbitrary complexity.

On the basis of the results of the previous section it is seen that the conditional posterior distribution of $\boldsymbol{\theta}$, given everything else, is

$$p(\boldsymbol{\theta}|\cdot) \propto p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta})\, p\{d_0|n_0, \pi_0(\boldsymbol{\theta})\}\, p(\boldsymbol{\theta}),$$

the same as in the full data case. The conditional posterior distribution of $\mathbf{d}$ is

$$p(\mathbf{d}|\cdot) \propto p(\mathbf{D}|\mathbf{d})\, p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta})$$

and that of $\mathbf{n}$ is easily seen to be

$$p(\mathbf{n}|\cdot) \propto p(\mathbf{d}|\mathbf{n}, \boldsymbol{\theta})\, p(\mathbf{n}|\mathbf{N}, n_0).$$

The sampling from the conditional of $\boldsymbol{\theta}$ is achieved by using the Metropolis–Hastings sampler of the full data case.

The death counts vector $\mathbf{d}$ is updated in blocks, each block referring to an age group. This is done because, as we said before, when $\mathbf{D}$ is given, components of $\mathbf{d}$ from different age groups are independent, but components from the same group are not. The conditional posterior distribution of each block of death counts is the same as that of the full vector but without the terms referring to the other age groups. Therefore it is of the form

$$p(d_x, \ldots, d_{x+k}|\cdot) \propto p(D_x|d_x, \ldots, d_{x+k}) \prod_{i=x}^{x+k} p\{d_i|n_i, \pi_i(\boldsymbol{\theta})\}.$$

This is the product of several binomial distributions and of an indicator function. Again, at each Gibbs step we perform one Metropolis step for each age group as follows.

Assume that we are working in group $[x, x+k]$ and that the current death counts are $d_x, \ldots, d_{x+k}$. One way to propose a new set of counts $d'_x, \ldots, d'_{x+k}$ is to sample one value from each of the first k binomial distributions of the group, independently from the others. This will give proposed values for all except the last age of the group. A value for that can then be obtained by subtracting the sampled counts from the observed total. Of course, since the sampling is done independently, nothing guarantees us that the result of the subtraction will be positive. But then, the posterior distribution of the values proposed will be 0 and therefore they will be rejected. In other words, we use a sampler with proposal distribution

$$p(d_x, \ldots, d_{x+k}) = \prod_{i=x}^{x+k-1} p\{d_i|n_i, \pi_i(\boldsymbol{\theta})\}.$$

If we keep in mind the forms of the posterior and the proposal we shall see that the probability of accepting the proposed values is

$$\alpha = \min\left(1, \frac{\left[\prod_{i=x}^{x+k} p\{d'_i|n_i, \pi_i(\boldsymbol{\theta})\}\right]\left[\prod_{i=x}^{x+k-1} p\{d_i|n_i, \pi_i(\boldsymbol{\theta})\}\right] p(D_x|d'_x, \ldots, d'_{x+k})}{\left[\prod_{i=x}^{x+k} p\{d_i|n_i, \pi_i(\boldsymbol{\theta})\}\right]\left[\prod_{i=x}^{x+k-1} p\{d'_i|n_i, \pi_i(\boldsymbol{\theta})\}\right] p(D_x|d_x, \ldots, d_{x+k})}\right).$$

This simplifies considerably because we only form the ratio if the last count is positive; otherwise we retain the old values. When the last count is positive the acceptance ratio becomes

$$\alpha = \min\left[1, \frac{p\{d'_{x+k}|n_{x+k}, \pi_{x+k}(\boldsymbol{\theta})\}}{p\{d_{x+k}|n_{x+k}, \pi_{x+k}(\boldsymbol{\theta})\}}\right].$$

The case of the population counts per year of age is very similar. Again we update block by block and we use the Metropolis algorithm within the Gibbs structure. Assume again that we are working with group $[x, x+k]$. The proposal distribution for the counts is a multinomial distribution with parameters the total count for the group and the same proportion for each

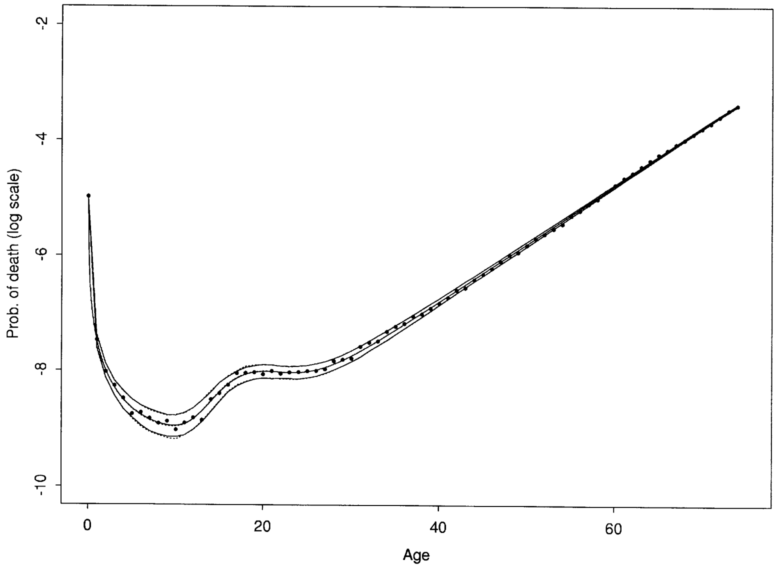

Fig. 4. Empirical data (●) and fitted and 95% predictive curves for both the full (———) and the abridged (··········) tables provided by the logistic non-linear model

age. Therefore, all the values proposed give $p(N_x|n_x, \ldots, n_{x+k}) = 1$ and so the acceptance ratio, after some cancellations, is

$$\alpha = \min\left[1, \frac{\prod_{i=x}^{x+k} p\{d_i|n_i', \pi_i(\boldsymbol{\theta})\} \prod_{i=x}^{x+k} n_i'!}{\prod_{i=x}^{x+k} p\{d_i|n_i, \pi_i(\boldsymbol{\theta})\} \prod_{i=x}^{x+k} n_i!}\right].$$

3.2. An example

For illustration—and to aid a direct comparison with our earlier analysis of the English and Welsh data—we have grouped the data of Table 1 into the above-mentioned age intervals. We have used exactly the same prior distributions for $\boldsymbol{\theta}$ as in the complete-data case and a similar simulation strategy. Fig. 4 shows the fitted curve and the 95% predictive intervals and also repeats the results of the logistic non-linear model for the full data case. Fig. 3 depicts, in the form of box plots, the marginal distributions derived.

We see in Fig. 4 that the predictive intervals arising from the analysis of the grouped data are slightly wider than those of the full data case. Furthermore, we observe from Fig. 3 that, in the case of the logistic non-linear model, the posterior distributions derived from the abridged table have greater variance than those referring to the full table. This is expected because the full table provides more information than the abridged table does.

4. Other posterior summaries

Many other quantities derived from the underlying model parameters are of interest to actuaries and demographers. Least squares approaches are not well suited to deriving inferences for non-linear transformations of basic parameters. However, the simulation-based Bayesian framework provides an easy way to calculate such quantities. For non-Bayesian approaches, see Pollard (1991) and Congdon (1994). In the following subsections, we shall outline the Bayesian approach to inference for four such quantities (based on either complete or incomplete data). An illustrative analysis will be presented for the first of these only.

4.1. *Survival probability*

The survival probability, denoted ${}_tp_x$, is defined as the probability of surviving from age x to age $x+t$, so

$$ {}_tp_x = 1 - {}_t\pi_x = \prod_{i=0}^{t-1} (1 - \pi_{x+i}). \tag{8}$$

It is simple to see that a sample from ${}_tp_x$ is readily available if we utilize the sample from π_{x+i} obtained as described in Section 2. The posterior survivor function (or lifetime) of a person can then be plotted in a way similar to that presented in Dellaportas and Smith (1993): for a given age, a posterior sample of the survivor probability can be summarized in a box plot form as shown for the English and Welsh females data and for $t = 5$ in Fig. 5.

4.2. *First to die*

Assume that there are two independent people Z and Y subject to the same life-table probabilities, and that we are asked to estimate the probability that one of them, Z say, will die first. Let $\pi_{1,x}$ and $\pi_{2,x}$, and ${}_tp_{1,x}$ and ${}_tp_{2,x}$, for $x = 0, 1, \ldots, n$ and $t = 1, 2, \ldots, n-x$, be the probabilities of dying and the survival probabilities of the two people respectively. Let the current ages of the two people be z and y respectively, with $z \geqslant y$, and denote by $d_{z,y}$ the probability that the first person dies first when their ages are z and y. It is readily shown that the probability required is

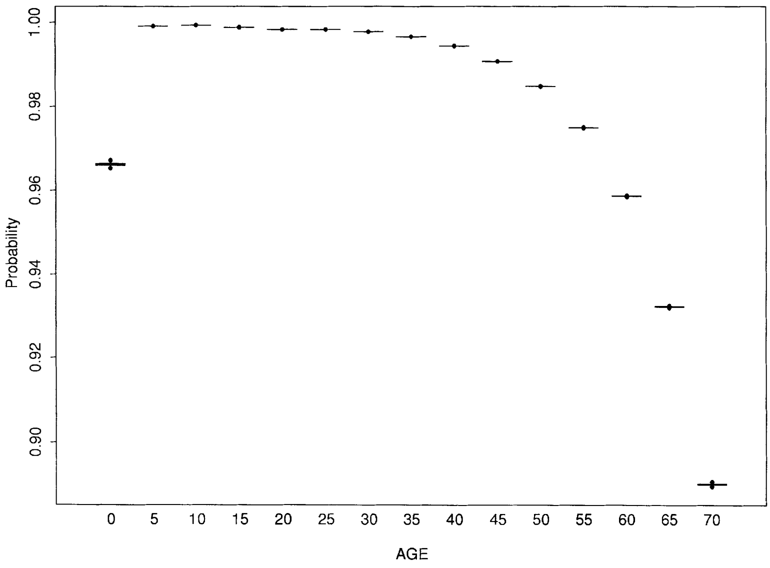

Fig. 5. Posterior survivor function ${}_tp_x$ for a full table provided by the logistic non-linear model, $t = 5$

$$
\begin{aligned}
P(Z \text{ dies first}) = {} & d_{z,y} \\
& + d_{z+1,y+1} \times {}_1p_{1,z} \times {}_1p_{2,y} \\
& + d_{z+2,y+2} \times {}_2p_{1,z} \times {}_2p_{2,y} \\
& \vdots \\
& + d_{n,y+n-z} \times {}_{n-z}p_{1,z} \times {}_{n-z}p_{2,y}.
\end{aligned} \tag{9}
$$

Again, for a set of sampled values of $\pi_{1,z}$ and $\pi_{2,y}$, and ${}_tp_{1,z}$ and ${}_tp_{2,y}$, we can easily estimate all the probabilities on the right-hand side of equation (9). Probabilities such as $d_{z,y}$ can be estimated by simply checking when, in the posterior samples of $\pi_{1,z}$ and $\pi_{2,y}$, the value of the first variate is larger than the value of the second.

4.3. Joint lifetime

The time to death of two people, considered as an entity, is sometimes defined as the time to the death of the first of them. Assume that the two people, aged y and z, have probabilities of dying $\pi_{1,y}$ and $\pi_{2,z}$ and survival probabilities ${}_tp_{1,y}$ and ${}_tp_{2,z}$. The probability that at least one death occurs in the next t years, say J_t, is just

$$J_t = {}_tp_{1,y} \times {}_t\pi_{2,z} + {}_t\pi_{1,y} \times {}_tp_{2,z} + {}_t\pi_{1,y} \times {}_t\pi_{2,z} \tag{10}$$

and illustrative graphs can be produced in the same way as for the lifetime of one person as described earlier.

4.4. Median time to death

The median lifetime can be determined by equating ${}_tp_x$ in equation (8) to 0.5. For m sampled values of $(\boldsymbol{\theta}, \sigma^2)$ we can produce m values of $(\pi_x, \pi_{x+1}, \ldots, \pi_{x+n})$, and therefore m values of ${}_tp_x$, $t = 1, \ldots, n - x$. We can then solve equation (8) m times and produce a posterior sample of the median life time to death. The equation can be solved via an efficient search numerical algorithm; see for example Ripley (1987), section 3.3.

5. Discussion

We have taken up the theme in Congdon (1993) that parametric modelling of mortality data offers many advantages. However, the implementation of a statistical analysis for highly parameterized non-linear functions is non-trivial, even for the log-normal case, let alone for generalized non-linear model cases. We have shown that Markov chain Monte Carlo techniques overcome this problem and enable rich and flexible inference summaries to be provided, not only for the mortality curves and their parameters, but also for a variety of related predictive problems such as calculations of survival probabilities, first-to-die probabilities, joint lifetimes and median times to death.

Future work will include attempts at modelling the time evolution of parameters (a form of non-linear filtering).

Acknowledgements

We are grateful to the Joint Editor and the reviewers for valuable suggestions on presentation and modelling. We are also grateful to Dr A. Kostaki of the Athens University of Economics and Business for many helpful discussions.

The first author was financially supported by the National Hellenic Research Foundation and the Royal Society of London to visit the second author at Imperial College of Science, Technology and Medicine, London. The third author was financially supported by the Engineering and Physical Sciences Research Council, UK, and the Alexander S. Onassis Public Benefit Foundation, Greece.

Appendix A

Assume that, for a given parameter vector $\boldsymbol{\theta} \in \Re^k$ and data $\mathbf{x}$, we require a sample from the posterior density $p(\boldsymbol{\theta}|\mathbf{x})$. Markov chain Monte Carlo approaches construct an irreducible and aperiodic Markov chain with realizations $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^t, \ldots$ in the parameter space, equilibrium distribution $p(\boldsymbol{\theta}|\mathbf{x})$ and a transition probability

$$K(\boldsymbol{\theta}', \boldsymbol{\theta}) = P(\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}'|\boldsymbol{\theta}^t = \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are the realized states at times t and $t+1$ respectively. Under appropriate regularity conditions, asymptotic results guarantee that, as $t \rightarrow \infty$, $\boldsymbol{\theta}^t$ tends in distribution to a random variable with distribution $p(\boldsymbol{\theta}|\mathbf{x})$ and that the ergodic average of an integrable function of $\boldsymbol{\theta}$ is a consistent estimator of the posterior mean of the function.

Different choices of the transition kernel K lead to different Markov chain Monte Carlo sampling schemes. A range of hybrid algorithms can also be derived by combining different schemes. We briefly describe here the two schemes used in the paper.

A.1. The Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm simulates a Markov chain by using the transition kernel

$$K(\boldsymbol{\theta}', \boldsymbol{\theta}) = \begin{cases} q(\boldsymbol{\theta}'|\boldsymbol{\theta})\,\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta}' \neq \boldsymbol{\theta}, \\ 1 - \sum\limits_{\boldsymbol{\theta}''} q(\boldsymbol{\theta}''|\boldsymbol{\theta})\,\alpha(\boldsymbol{\theta}''|\boldsymbol{\theta}) & \text{otherwise,} \end{cases}$$

where $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ is a *proposal* conditional distribution and $\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta})$ is an *acceptance* conditional distribution, defined (here) by

$$\alpha(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \min\left\{\frac{p(\boldsymbol{\theta}'|\mathbf{x})\,q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}|\mathbf{x})\,q(\boldsymbol{\theta}'|\boldsymbol{\theta})}, 1\right\}.$$

A.2. The Gibbs sampler

The Gibbs sampler can be regarded as a special case of the Metropolis–Hastings algorithm. The transition from state t to state $t+1$ is achieved via a sequence of k steps, each updating every coordinate of $\boldsymbol{\theta}$ in turn. The algorithm requires the ability to sample from all full conditional densities of the form $p(\theta_i|\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$. The transition kernel is given by

$$K(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \prod_{i=1}^{k} p(\theta_i^{t+1}|\theta_1^{t+1}, \ldots, \theta_{i-1}^{t+1}, \theta_{i+1}^{t}, \ldots, \theta_k^{t}).$$

References

Benjamin, B. and Pollard, J. (1980) *The Analysis of Mortality and Other Actuarial Statistics*, 2nd edn. London: Heinemann.

Carlin, B. P. (1992) A simple Monte-Carlo approach to Bayesian Graduation. *Trans. Soc. Act.*, **44**, 55–76.

Congdon, P. (1993) Statistical graduation in local demographic analysis and projection. *J. R. Statist. Soc.* A, **156**, 237–270.

——(1994) Analysing mortality in London: life-tables with frailty. *Statistician*, **43**, 277–308.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.*, **42**, 443–459.

Forfar, D. O., McCutcheon, J. J. and Wilkie, A. D. (1988) On graduation by mathematical formula. *J. Inst. Act.*, **115**, 1–149.

Forfar, D. O. and Smith, D. M. (1987) The changing shape of English life tables. *Trans. Fac. Act.*, **40**, 98–133.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gelfand, A. E., Smith, A. F. M. and Lee, T.-M. (1992) Bayesian analysis of constrained parameter and truncated data problems. *J. Am. Statist. Ass.*, **87**, 523–532.

Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 169–193. Oxford: Oxford University Press.

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. *Phil. Trans. R. Soc.*, **115**, 513–585.

Haberman, S. and Renshaw, A. E. (1996) Generalized linear models and actuarial science. *Statistician*, **45**, 407–436.

Hartmann, M. (1983) Past and recent attempts to model mortality at all ages. *J. Off. Statist.*, **3**, 19–36.

Heidelberger, P. and Welch, P. (1983) Simulation run length control in the presence of an initial transient. *Ops Res.*, **31**, 1109–1144.

Heligman, L. and Pollard, J. H. (1980) The age pattern of mortality. *J. Inst. Act.*, **107**, 49–80.

Hickman, J. C. and Miller, R. B. (1977) Notes on Bayesian graduation. *Trans. Soc. Act.*, **29**, 1–21.

Hills, S. E. and Smith, A. F. M. (1992) Parameterization issues in Bayesian inference. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 227–246. Oxford: Oxford University Press.

Kimeldorf, G. S. and Jones, D. A. (1967) Bayesian graduation. *Trans. Soc. Act.*, **19**, 66–112.

Kostaki, A. (1991) The Heligman-Pollard formula as a tool for expanding an abridged life table. *J. Off. Statist.*, **7**, 311–323.

——— (1992a) Methodology and applications of the Heligman-Pollard formula. *PhD Thesis*. Department of Statistics, University of Lund, Lund.

——— (1992b) A nine-parameter version of the Heligman-Pollard formula. *Math. Popln Stud.*, **3**, 277–288.

Mode, C. and Busby, R. (1982) An eight parameter model of human mortality — the single decrement case. *Bull. Math. Biol.*, **44**, 647–659.

Pollard, J. H. (1989) On the derivation of a full life table from mortality data recorded in five-year age groups. *Math. Popln Stud.*, **2**, 1–14.

——— (1991) Fun with Gompertz. *Genus*, **57**, 1–19.

Raftery, A. L. and Lewis, S. M. (1992) How many iterations in the Gibbs sampler? In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 763–773. Oxford: Oxford University Press.

Renshaw, A. E. (1991) Actuarial graduation practice and generalised linear and non-linear models. *J. Inst. Act.*, **118**, 295–312.

Ripley, B. D. (1987) *Stochastic Simulation*. New York: Wiley.

Rogers, A. (1986) Parametrized multistate population dynamics and projections. *J. Am. Statist. Ass.*, **81**, 48–61.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.