Insurance: Mathematics and Economics 48 (2011) 355–367

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Mortality density forecasts: An analysis of six stochastic mortality models

Andrew J.G. Cairns [a,*], David Blake [b], Kevin Dowd [b], Guy D. Coughlan [c,1], David Epstein [c], Marwa Khalaf-Allah [c]

[a] Maxwell Institute for Mathematical Sciences, and Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom

[b] Pensions Institute, Cass Business School, City University, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom

[c] Pension Advisory Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom

ARTICLE INFO

Article history:
Received November 2010
Received in revised form
December 2010
Accepted 20 December 2010

Keywords:
Plausibility
Fan charts
Model risk
Forecasting
Model selection criteria

ABSTRACT

This paper develops a framework for developing forecasts of future mortality rates. We discuss the suitability of six stochastic mortality models for forecasting future mortality and estimating the density of mortality rates at different ages. In particular, the models are assessed individually with reference to the following qualitative criteria that focus on the plausibility of their forecasts: biological reasonableness; the plausibility of predicted levels of uncertainty in forecasts at different ages; and the robustness of the forecasts relative to the sample period used to fit the model. An important, though unsurprising, conclusion is that a good fit to historical data does not guarantee sensible forecasts. We also discuss the issue of model risk, common to many modelling situations in demography and elsewhere. We find that even for those models satisfying our qualitative criteria, there are significant differences among central forecasts of mortality rates at different ages and among the distributions surrounding those central forecasts.

1. Introduction

The last twenty years has seen a growing range of models for forecasting mortality. Early work on stochastic models by McNown and Rogers (1989) and Lee and Carter (1992) has been followed by:

- developments on the statistical foundations by, for example, Lee and Miller (2001), Brouhns et al. (2002), Booth et al. (2002a), Czado et al. (2005), Delwarde et al. (2007), and Li et al. (2009); and
- the development of new stochastic models by Booth et al. (2002a,b, 2005), Cairns et al. (2006b) (CBD), Renshaw and Haberman (2006), Hyndman and Ullah (2007), Cairns et al. (2009), Plat (2009) and Debonneuil (2010).

These stochastic models vary significantly according to a number of key elements: number of sources of randomness driving mortality improvements at different ages; assumptions of smoothness in the age and period dimensions; inclusion or not of cohort effects; estimation method.

A number of studies have sought to draw out more formal comparisons between a number of these models. Some of these limit themselves to comparison of some variants of the Lee–Carter model (Lee and Miller, 2001; Booth et al., 2002a,b, 2005). Hyndman and Ullah (2007) compare out-of-sample forecasting performance of the Lee–Carter model and its Lee–Miller and Booth–Maindonald–Smith variants with a new class of multifactor models. CMI (2005, 2006, 2007) compare the Lee–Carter, Renshaw and Haberman and *P*-splines models. These types of analysis have been extended to a wider range of models with substantially different characteristics by the present authors; this paper is one part of this endeavour.

Cairns et al. (2009) focused on quantitative and qualitative comparisons of eight stochastic mortality models (see Table 1 in Section 2), based on their general characteristics and ability to explain *historical* patterns of mortality. The criteria employed included: quality of fit, as measured by the Bayes information criterion (BIC); ease of implementation; parsimony; transparency; incorporation of cohort effects; ability to produce a non-trivial correlation structure between ages; robustness of parameter estimates relative to the period of data employed.

Complementing this, Dowd et al. (2010a,b) carry out a range of formal, out-of-sample backtesting and goodness-of-fit tests using mortality data for English and Welsh males. They find that some models fare better under some criteria than others,

* Corresponding author.
E-mail address: A.Cairns@ma.hw.ac.uk (A.J.G. Cairns).

[1] Disclaimer: This report has been partially prepared by the Pension Advisory group, and not by any research department, of JPMorgan Chase & Co. and its subsidiaries ("JPMorgan"). Information herein is obtained from sources believed to be reliable but JPMorgan does not guarantee its completeness or accuracy. Opinions and estimates constitute JPMorgan's judgment and are subject to change without notice. Past performance is not indicative of future results. This material is provided for informational purposes only and is not intended as a recommendation or an offer or solicitation for the purchase or sale of any security or financial instrument.

0167-6687/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2010.12.005

but that no single model can claim superiority under all the criteria considered. In any event, different patterns of mortality improvements in different countries means that models that are best for one country might not be as suitable for another. Finally, this paper focuses on the *ex ante* plausibility and robustness of forecasts produced by the different models. The present paper, therefore, focuses on the *ex ante* qualitative aspects of forecasts, while the previous works (Cairns et al., 2009; Dowd et al., 2010a,b) focus on the *ex post* quantitative aspects.

Building on the analyses of historical data of Cairns et al. (2009) and Dowd et al. (2010a,b), the present paper focuses on *ex ante* qualitative aspects of mortality forecasts and the distribution of results around central forecasts. Specifically, we introduce a number of qualitative criteria that focus on the plausibility of forecasts made using different models.

Often in this paper, we will refer to the concept of *biological reasonableness* (which was first proposed in Cairns et al., 2006a). The concept is not intended to refer to criteria based on hard scientific (biological or medical) facts. Instead, it is intended to cover a wide range of subjective criteria, related to biology, medicine and the environment. What the modeller needs to do is look at the results and ask the question: *what mixture of biological factors, medical advances and environmental changes would have to happen to cause this particular set of forecasts?* As one example, the upper set of projections in Fig. 4 at age 85 looks rather more unusual than the two lower sets of projections under a particular model. Under the upper scenario, we would have to think of a convincing biological, medical or environmental reason why, *with certainty*, age 85 mortality rates are going to deteriorate to 1960's levels. If the modeller cannot think of any good reason why this might happen, then she must rule out the model (at least with its current method of calibration) on grounds of biological unreasonableness.

Besides biological reasonableness, we also consider the issue of *the plausibility of forecast levels of uncertainty in projections at different ages*. The objective here is to judge whether or not the pattern of uncertainty at different ages is consistent with historical levels of variability at different ages: we can sometimes conclude that a particular model is less plausible on the basis of forecast levels of uncertainty.

An important additional issue concerns the *robustness of forecasts* relative to the choice of sample period and age range. If we make a small change either to the sample period (for example, when we add in the latest mortality data) or to the age range, we would normally expect to see, with a robust model, only modest changes in the forecasts at all ages. Where a model is found to lack robustness with one sample population, there is a danger that it will lack robustness if applied to another sample population and should, therefore, either be used with great care or not used at all.

Although application of such a wide ranging set of model selection criteria will eliminate some models, we will demonstrate that mortality forecasting is no different from many other modelling problems where model risk is significant: mortality forecasters should acknowledge this fact and make use of multiple models rather than pretend that it is sufficient to make forecasts based on any single model.

1.1. Plan for this paper

We will consider qualitative assessment criteria that allow us to examine the *ex ante* plausibility of the forecasts generated by six stochastic mortality models, illustrating with national population data for England & Wales (EW) for an age group consisting of males 60–89 years old and estimated over the years 1961–2004. This is supplemented by a briefer discussion of forecasts for the equivalent US dataset. We focus on higher ages because our current principal research interest is the longevity risk facing pension plans and annuity providers.

Table 1
Formulae for six out of the original eight mortality models investigated by Cairns et al. (2009). The functions $\beta_x^{(i)}$, $\kappa_t^{(i)}$, and $\gamma_{t-x}^{(i)}$ are age, period and cohort effects, respectively. $\bar{x}$ is the mean age over the range of ages being used in the analysis. $\hat{\sigma}_x^2$ is the mean value of $(x-\bar{x})^2$. n_a is the number of ages.

Model	Formula
M1	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$
M2	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}$
M3	$\log m(t,x) = \beta_x^{(1)} + n_a^{-1}\kappa_t^{(2)} + n_a^{-1}\gamma_{t-x}^{(3)}$
M5	$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x})$
M7	$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}\left((x-\bar{x})^2 - \hat{\sigma}_x^2\right) + \gamma_{t-x}^{(4)}$
M8	$\text{logit}\, q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_{t-x}^{(3)}(x_c - x)$

We will concentrate on six of the models discussed by Cairns et al. (2009): these are labelled in Table 1 as M1, M2, M3, M5, M7 and M8. Models M2, M3, M7 and M8 include a cohort effect and these emerged in Cairns et al. (2009) as the best fitting, in terms of BIC, of the eight models considered on the basis of male mortality data from EW and the US for the age group under consideration. M2 is the Renshaw and Haberman (2006) extension of the original Lee–Carter model (M1), M3 is a special case of M2, and M7 and M8 are extensions of the original CBD model (M5). The original Lee–Carter and CBD models had no cohort effect, and provide useful benchmarks for comparison with the four models involving cohort effects. M4 is not considered any further in this study because of its low BIC and qualitative rankings for these datasets in Cairns et al. (2009, Table 3). (M4 focuses on identifying the smooth underlying trend. However, this means that it is not as good as the other models at capturing short-term deviations from this trend.) Although M3 is a special case of M2, we include it here because it had a relatively high BIC ranking for the US data, and because it avoids a problem with the robustness of parameter estimates for M2 identified by CMI (2007), Cairns et al. (2009), and Dowd et al. (2010a,b). M6 was also dropped from the original set of eight models: M6 is a special case of M7, and M7 was found to be stable and to deliver consistently better and more plausible results than M6.

The structure of the paper is as follows. In Section 2, we specify the stochastic processes needed for forecasting the term structure of mortality rates for each of the models. Results for the different models obtained using EW male mortality data are compared and contrasted in Section 3. Section 5 examines two applications of the forecast models, namely applications to survivor indices and annuity prices, and makes additional comments on model risk and plausibility of the forecasts. Each model is then tested for the robustness of its forecasts in Section 4. Finally, in Section 6, we summarise an analysis for US male mortality data: our aim is to draw out features of the US data that are distinct from those of the EW data. Section 7 concludes.

2. Forecasting with stochastic mortality models

We take six stochastic mortality models which, on the basis of fitting to historical data, appear to be suitable candidates for forecasting future mortality at higher ages, and prepare them for forecasting. To do this, we need to specify the stochastic processes that drive the age, period and (if present) cohort effects in each model.

We define $m(t,x)$ to be the death rate in year t at age x, and $q(t,x)$ to be the corresponding mortality rate, with the relationship between them given by $q(t,x) = 1 - \exp[-m(t,x)]$. The models considered are outlined in Table 1.

All but M5 require the use of one or more identifiability constraints (see Appendix A.1), and parameter values for the age,